COMMON FORMULAS AND CALIBRATION FRONTIERS IN PAC–BAYES BOUNDS

Anonymous authors
Paper under double-blind review

ABSTRACT

PAC–Bayes bounds certify predictors selected from data, but their tightness can be limited by the need to use one formula across sample outcomes or recover classical confidence bounds when posterior equals prior. We characterize when one normalized convex comparator attains every statewise optimum and derive matching excess bounds in two regimes for fixed finite interior targets with positive weights. A fixed task using empirical risk minimization has low risk and bounded Kullback–Leibler (KL) complexity, yet its optimal expected excess is of the same order as the benchmark width above empirical risk. For finite binary-loss libraries, our construction combines Clopper–Pearson (CP) exactness at a shared point mass with simultaneous posterior coverage, continuity and KL learning, whereas exact recovery at a mixed prior can obstruct learning. Under the specified randomized evaluation, every rule with CP recovery at the original confidence level and simultaneous coverage throughout the full model gives a perfect predictor a bound tending to one in probability on a common fixed task. Allowing pathwise one-sided calibration tolerance yields matching asymptotic limits for width quantiles under each distribution and in the worst case, with the finite model, positive prior and KL budget fixed. The worst-case width can vanish when the tolerance decays subexponentially, whereas exponential decay can leave a positive limit. One explicit rule uses joint losses to attain each distribution's optimal limit without knowing the distribution. This attainment is uniform over growing finite libraries and all positive priors and data laws at specified confidence levels with a fixed positive precision exponent and KL budget.

1 INTRODUCTION

Generalization bounds connect a predictor's performance on its training sample to its population risk on future observations. Their usefulness depends on both coverage at the prescribed confidence level and how tightly they constrain the predictor's risk. How tight can a bound be under given statistical requirements? PAC–Bayes bounds combine empirical risk with posterior-to-prior Kullback–Leibler (KL) complexity. The prior is fixed before sampling, while one high-probability event covers all posteriors, permitting data-dependent selection (McAllester, 1999; Seeger, 2002; Catoni, 2007). Valid formulas can still give different bounds for the same predictor and sample. We study two requirements: using one formula across sample states, and recovering a classical confidence bound when posterior equals prior. The first concerns the limits of standard convex-comparator derivations. The second asks whether classical recovery is compatible with posterior learning.

Convex comparators connect empirical and population risk (Germain et al., 2009; Bégin et al., 2016). For each empirical risk and complexity penalty, Foong et al. (2021) identify a Catoni formula attaining the optimal endpoint, but this formula can vary across sample states. Can one comparator chosen before sampling attain the expected statewise optimum, and if not, how close can it get? This is their Open Problem 1. We characterize when states share an optimal formula and give matching excess bounds in two finite interior regimes. This cost also arises in a fixed low-risk empirical risk minimization (ERM) task: two classifiers have equal population risks but unequal prior weights, so ERM's sample-dependent choices incur different KL penalties. The optimal expected excess is of order $N^{-1/2}$, as is the benchmark width above empirical risk: a common formula can lose leading-order precision even when the selected predictor always has low risk.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

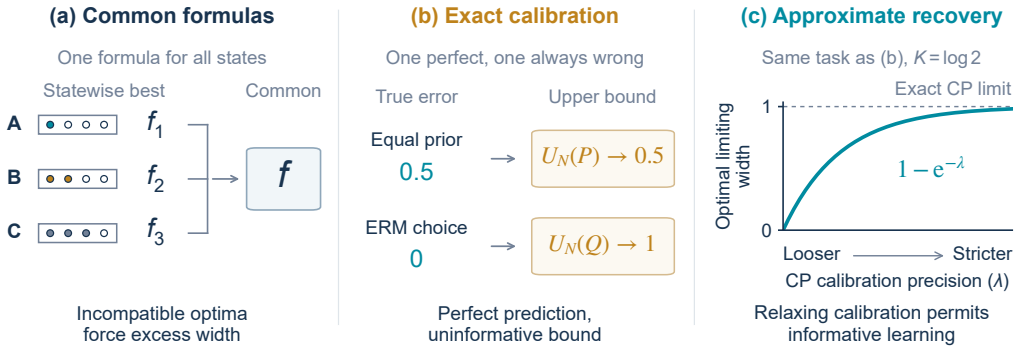


Figure 1: Two requirements and their costs. (a) A common normalized comparator incurs excess when statewise optima conflict; sample strips are schematic. (b) On constant loss vector $(0, 1)$, $P = (1/2, 1/2)$, ERM selects $Q = (1, 0)$. Randomized binomial evaluation, original-confidence prior CP recovery, and full-model common coverage force $U_N(P) \to 1/2, U_N(Q) \to 1$ in probability (Theorem 5). (c) On this task with KL budget $\log 2$, pathwise one-sided tolerance $\epsilon_N \to 0$ gives limiting optimal width $1 - e^{-\lambda}$, $\lambda = \lim_N N^{-1} \log(1/\epsilon_N)$. Panels (b,c) show asymptotic limits.

Even the optimal comparator endpoint remains strictly above the classical Clopper–Pearson (CP) bound for a fixed binary-loss predictor at error counts between zero and the sample size. Foong et al.'s Open Problem 2 therefore asks whether another PAC–Bayes construction can recover the binomial tail bound gracefully when posterior equals prior. For finite binary-loss libraries, we construct a common confidence set for risks that allows exact CP recovery at shared point masses while retaining simultaneous posterior coverage, continuity, and a KL learning guarantee (Theorem 4).

For mixed priors, averaged losses need not be binary. We use the randomized binomial evaluation of Foong et al. (2021, Appendix I.2). Our transport construction retains exact CP recovery and common coverage, but exact recovery can obstruct posterior learning. Under a uniform prior on one always-correct and one always-wrong classifier, ERM selects the perfect predictor. Nevertheless, original-confidence CP recovery and common coverage throughout the full binary-loss model force its upper bound toward one (Theorem 5). Near CP thresholds, failures at the prior can use almost the entire failure budget, even when the rule uses all individual losses.

How much relaxation restores informative guarantees? At fixed coverage, we allow the prior bound to exceed CP by at most ϵ_N on every sample, including evaluation randomness. In fixed finite models, subexponentially vanishing tolerance permits vanishing worst-case width. Exponential tolerance can leave a positive limit (Figure 1). The matching bounds below cover all jointly valid rules under this randomized calibration requirement, not just comparator-derived bounds.

Our main contributions are:

- **Compatibility and optimal cost of a common formula.** For fixed finite positive-weight interior targets, we characterize exact compatibility and derive sharp excess laws: sensitivity-weighted entropy for separated states sharing an endpoint, and a Gaussian-envelope limit near an interior diagonal (Section 3). A fixed low-risk ERM task incurs a cost of the same order as the benchmark width despite bounded posterior KL (Theorem 1).
- **Feasibility and learning limits of exact CP recovery.** We construct common posterior guarantees that recover CP, with continuity and KL learning at shared point masses (Theorem 4). Under mixed-prior randomized evaluation, every rule with original-confidence CP recovery and full-model common coverage must give the perfect ERM predictor a bound tending to one in probability on the same fixed zero-risk task, even for rules that use the full loss matrix (Theorem 5).
- **Matching precision–width bounds and uniform attainment.** For fixed finite binary-loss models, positive priors, and fixed KL budgets, we determine matching first-order width limits for each data law and the worst law (Theorem 6). *One explicit rule uses joint loss information to attain each law's local frontier without knowing the law.* Attainment is uniform over growing finite libraries, positive priors, and data laws, at a fixed positive precision exponent and fixed KL budget under the confidence levels of Theorem 7.

These results give a negative answer to Open Problem 1 and establish sharp selection costs in two specified regimes. For finite binary-loss libraries we answer Open Problem 2 under randomized Gibbs evaluation: the bound recovers CP exactly when posterior equals prior and remains valid for all posteriors. Total variation controls how far it moves from the prior's CP endpoint as the posterior changes. We also quantify how this recovery requirement limits posterior learning.

1.1 Related Work

Nagarajan & Kolter (2019) construct low-risk learners for which even algorithm-dependent two-sided uniform convergence gives vacuous bounds. Gastpar et al. (2024) establish limits on uniformly tight generalization bounds; Gastpar et al. (2025) connect risk estimability for algorithms and distribution families to stability and conditional variance. We quantify optimal width under common posterior coverage and specified calibration constraints, including an attaining rule.

Within convex comparators (Germain et al., 2009; Bégin et al., 2016), Foong et al. (2021) identify the statewise Catoni optimum. Hellström & Guedj (2024) optimize an average-risk objective and give a Cramér-comparator benchmark with near-optimal high-probability bounds. We average inverse endpoints across prescribed sample states. Rodríguez-Gálvez et al. (2024) permit sample-dependent temperature choice at the Seeger–Langford threshold $t_{\mathrm{SL}} = t_0 + \log \xi(N)/N$; we study attainment at $t_0 = [\mathrm{KL}(Q\|P) + \log(1/\delta)]/N$.

CP inversion covers one binomial parameter (Clopper & Pearson, 1934); we additionally require common posterior coverage. Our risk sets use p-to-e calibration (Vovk & Wang, 2021) and vector-risk ideas (Adams et al., 2024), with coordinates indexing hypotheses. Coin-betting (Jang et al., 2023) and non-KL complexity (Kuzborskij et al., 2024) sharpen certificates without imposing our original-confidence CP recovery condition. Relative-entropy ambiguity sets (Van Parys et al., 2021; Sutter et al., 2024) and their geometric-mean dual (Van Parys et al., 2021) are established; here pathwise CP tolerance determines the limiting radius at fixed coverage, with a prior-constrained frontier and uniform attainability. Appendix Q gives detailed comparisons.

2 Preliminaries and Problem Setup

Let $S \sim \mathcal{D}^N$ and measurable losses $\ell(h, z) \in [0, 1]$ be defined on standard Borel spaces. The prior P is independent of S; the posterior Q may depend on it. Write $L(Q)$ for expected loss under $h \sim Q$ and an independent observation, and $\widehat{L}_S(Q)$ for the corresponding sample average. A valid upper endpoint $U(S, Q)$ satisfies $L(Q) \le U(S, Q)$ for all admissible Q on one event of probability at least $1-\delta$, permitting posterior selection on the same data. Throughout, $\delta \in (0, 1)$ is the failure probability, logarithms are natural, and kl denotes binary relative entropy with its lower-semicontinuous boundary extension.

The generic PAC–Bayes theorem uses convex comparators of empirical and population risk (Germain et al., 2009; Bégin et al., 2016). Let $\mathcal{F}_N$ consist of proper jointly convex lower-semicontinuous functions $f : [0, 1]^2 \to \mathbb{R} \cup \{+\infty\}$, never $-\infty$, satisfying

$$I_{f,N} := \sup_{r \in [0,1]} \mathbb{E}_{K \sim \mathrm{Bin}(N,r)} e^{N f(K/N, r)} \le 1. \tag{1}$$

This condition must hold for every population risk r; zero-probability summands are omitted at $r = 0, 1$. For any comparator Δ with $0 < I_{\Delta,N} < \infty$, the shift $f = \Delta - N^{-1} \log I_{\Delta,N}$ normalizes this moment without changing the bound. The Catoni comparator $f_\beta(q, r) = -\beta q - \log(1 - r + re^{-\beta})$ already has unit moment for every $\beta > 0$; its parameter selects a risk bound.

Write $c_\delta(Q) = [\mathrm{KL}(Q\|P) + \log(1/\delta)]/N$ for the complexity level. For any $f \in \mathcal{F}_N$ fixed before sampling, the generic theorem gives $f(\widehat{L}_S(Q), L(Q)) \le c_\delta(Q)$ on one event of probability at least $1 - \delta$, simultaneously for all finite-KL posteriors. We compare its rightmost feasible endpoint with the statewise benchmark (the latter defined for $c \ge 0$):

$$\begin{aligned} U_f(q, c) &:= \sup\big(\{r \in [0, 1] : f(q, r) \le c\} \cup \{0\}\big), \\ F(q, c) &:= \sup\{r \in [q, 1] : \mathrm{kl}(q, r) \le c\}. \end{aligned} \tag{2}$$

U_f is the largest allowed population risk. Optimizing the comparator separately at each interior state gives F (Foong et al., 2021); the optimizing comparator may change with the state.

For a fixed learner Q_S, write $p_f(S)$ and $p_{\tan}(S)$ for the endpoints evaluated at $(\widehat{L}_S(Q_S), c_\delta(Q_S))$. Both equal one at infinite complexity. The optimal expected excess for one comparator is

$$\inf_{f \in \mathcal{F}_N} \mathbb{E}_S[p_f(S) - p_{\tan}(S)]. \tag{3}$$

The infimum lies outside the expectation because one comparator must serve every sample outcome. The learner, prior, and data law stay fixed; changing f changes the certificate without changing $L(Q_S)$.

The CP benchmark in Section 4 concerns a broader class of simultaneously valid bounds. Scopes and boundary conventions are in Table 1 and Appendix B.1.

3 The Cost of Using One Formula

3.1 A Fixed Learning Task

In the task below, both classifiers have risk ρ, while prior weights determine the cost of certifying ERM's data-dependent choice.

Theorem 1 (Selection cost on a fixed low-risk task)**.** *Fix $0 < \rho < 1/2$ and $\pi, \delta \in (0,1)$. Let $\Pr(X = a) = \Pr(X = b) = \rho$ and $\Pr(X = c) = 1 - 2\rho$, with label zero. Classifier h_0 errs only on a, classifier h_1 only on b, and $P(h_0) = \pi$. Let Q_S be ordinary ERM with a tie rule fixed before sampling. There is a finite constant $V_\rho = V(\rho, \pi, \delta) \geq 0$ such that*

$$\sqrt{N} \inf_{f \in \mathcal{F}_N} \mathbb{E}_S\Big[U_f(\widehat{L}_S(Q_S), c_\delta(Q_S)) - F(\widehat{L}_S(Q_S), c_\delta(Q_S))\Big] \longrightarrow V_\rho.$$

Moreover, $V_\rho > 0$ exactly when $\pi \neq 1/2$. For every sample, $L(Q_S) = \rho$, and the selected point posterior has KL either $\log(1/\pi)$ or $\log(1/(1-\pi))$.

For unequal prior weights, the positive excess disproves universal attainability in Foong et al.'s Open Problem 1. It changes the leading width constant while true risk stays ρ; Figure 5(a) compares the two prior regimes.

3.2 When Can One Formula Attain Every Optimum?

To separate compatibility from sampling randomness, prescribe a finite target family $\mathsf{S} = \{(w_j, q_j, p_j)\}_{j=1}^m$, with $w_j > 0$, $\sum_j w_j = 1$, and $0 < q_j < p_j < 1$. Here q_j is an empirical risk, p_j its desired optimal endpoint, and w_j its weight in the objective. Set $t_j = \mathrm{kl}(q_j, p_j)$, so that $F(q_j, t_j) = p_j$, and define

$$\mathcal{G}_N^\star(\mathsf{S}) = \inf_{f \in \mathcal{F}_N} \sum_j w_j \{U_f(q_j, t_j) - p_j\}. \tag{4}$$

Targets are fixed before observation; realizing them as a learning task additionally requires a common data law, prior, and learner.

Even two targets with distinct empirical risks and a common optimal endpoint can conflict. Let ℓ_j support binary KL at (q_j, p_j) and $H = \max_j \ell_j$. Exact contact at every target forces a comparator above this envelope, which must itself pass the moment test.

Proposition 2 (Exact compatibility)**.** *For every finite positive-weight interior target family,*

$$\mathcal{G}_N^\star(\mathsf{S}) = 0 \iff I_{H,N} = 1 \iff \textit{one } f \in \mathcal{F}_N \textit{ attains every } p_j.$$

This equivalence does not assume that the infimum is attained. If all targets share p, compatibility holds exactly when all q_j coincide.

A normalizer above one rules out simultaneous contacts. Even approaching them arbitrarily closely forces $I_{H,N} = 1$, making H valid and sharp (Appendix F). At most one affine support per target preserves scores and positive-threshold endpoints. This represents the full-class optimum with finitely many planes, rather than restricting it to familiar formulas (Appendix A, Propositions 9–10).

The fixed task's targets approach the diagonal as N grows; their cost therefore requires a local analysis.

3.3 Why the Fixed Task Pays at the Leading Order

At fixed posterior KL and confidence, both sampling fluctuations and interior benchmark widths have order $N^{-1/2}$. Fix $q_0 \in (0,1)$, a finite number m of targets, and positive weights w_j summing to one. Suppose

$$\sqrt{N}(q_{j,N} - q_0) \to a_j, \qquad \sqrt{N}(p_{j,N} - q_0) \to b_j, \qquad d_j = b_j - a_j > 0.$$

The scaled benchmark gap d_j gives bounded $Nt_{j,N}$, where $t_{j,N} = \mathrm{kl}(q_{j,N}, p_{j,N})$; a realization at fixed confidence therefore has bounded posterior KL. Binomial fluctuations yield a Gaussian optimization over at most m planes, with a moment constraint for every population parameter. Let V be its optimal weighted endpoint value; (35) in Appendix C gives the exact variational definition.

Theorem 3 (Near-diagonal endpoint limit)**.** *For the fixed finite target sequences above,*

$$\sqrt{N}\,\mathcal{G}_N^\star \longrightarrow V, \qquad V = 0 \iff d_1 = \cdots = d_m. \tag{5}$$

Moreover, $0 \le V \le \sqrt{\sum_j w_j d_j^2} - \sum_j w_j d_j$.

Unequal gaps force a leading cost; equal gaps remove only that term, not necessarily the finite-sample excess. The upper bound uses one plane; several planes can do better.

Proof idea. Put $v_0 = q_0(1-q_0)$. Zero leading cost at gap d_j forces slope d_j/v_0. Under Gaussian fluctuations, each plane gives an exponential with mean one. Different gaps give distinct exponentials whose maximum has expectation above one. Restoring normalization moves at least one endpoint outward; equal gaps allow one plane to serve all targets. The full-class converse extracts supporting planes at any near-optimal comparator's actual endpoints and controls their rescaled parameters. Conversely, limiting planes yield finite-N comparators after a vanishing normalization correction, controlled over the entire population-parameter interval, including its boundary (Appendix C).

For the two-classifier task, put $v = \rho(1-\rho)$, $d_0 = \sqrt{2v\log(1/(\pi\delta))}$, and $d_1 = \sqrt{2v\log(1/((1-\pi)\delta))}$. Swapping the counts of a, b preserves the sample law and selected risk but exchanges the winner. The two KL penalties therefore have equal weight away from ties. Thus $V_\rho = V$ at $q_0 = \rho$, weights $(1/2, 1/2)$, and gaps (d_0, d_1); Appendix D controls ties and fluctuations. Unequal prior weights give different slopes and a positive leading cost.

3.4 How the Cost Changes for Separated Targets

For distinct fixed targets $(W_j, q_j, p)_{j=1}^r$, positive weights summing to one, and $0 < q_j < p < 1$, the sensitivity $s_j = 1/\partial_p \,\mathrm{kl}(q_j, p)$ converts score loss into endpoint displacement. Set $A = \sum_j W_j s_j$ and $\alpha_j^\star = W_j s_j / A$. The full-class endpoint cost is

$$\mathcal{G}_N^\star(\mathsf{S}) = \frac{AH(\alpha^\star)}{N} + O_{\mathsf{S}}(N^{-3/2}), \qquad H(\alpha) = -\sum_j \alpha_j \log \alpha_j. \tag{6}$$

Allocating normalization mass α_j costs state j a score loss $-\log \alpha_j / N$. Weighting by $W_j s_j$ and optimizing gives this entropy law, with a matching converse (Appendix G). Here, at fixed confidence, $\mathrm{KL}(Q\|P) = N\,\mathrm{kl}(q_j, p) - \log(1/\delta)$, which grows linearly with N.

These costs concern the comparator class. Its statewise optimum still exceeds CP at interior counts. We next leave that class and ask which forms of CP recovery preserve informative posterior bounds.

4 Exact Binomial Recovery with Posterior Learning

For a predictor fixed before sampling with binary error count $k < N$, the Clopper–Pearson endpoint $\mathrm{CP}_{N,\delta}(k)$ solves $\Pr\{\mathrm{Bin}(N, r) \le k\} = \delta$ for r; at $k = N$ it is one. At a shared point mass $P = Q = \delta_h$ and an interior count $0 < k < N$,

$$\mathrm{CP}_{N,\delta}(k) < F(k/N, \log(1/\delta)/N) \le U_f(k/N, \log(1/\delta)/N)\,, \qquad f \in \mathcal{F}_N. \tag{7}$$

CP and F agree at $k = 0, N$; the strict interior gap rules out scalar recovery there (Appendix H). A confidence set for candidate risks can impose CP directly while retaining one common event. Consider a finite library of M predictors with binary losses. Write $r^\star_h = L(h)$, $k_h = \sum_i \ell(h, Z_i)$, and $x_h = k_h/N$. The coordinates of $r^\star \in [0,1]^M$ need not sum to one, and losses may be dependent. Since $L(Q) = Q \cdot r^\star$, a confidence set $\mathcal{R}(P, S)$ gives

$$U(P, Q) = \max_{r \in \mathcal{R}(P,S)} Q \cdot r.$$

The event $r^\star \in \mathcal{R}(P, S)$ covers every data-selected Q. Unlike the error-type coordinates of Adams et al. (2024), ours index predictors.

The risk set splits failure probability between a global learning constraint, a CP cap on a dominant prior atom, and its complement. The allocation varies continuously with P, assigning the entire budget to CP at a point mass. Let U_{sq} be the support function of the squared-evidence risk set in (105). It has the following guarantees.

Theorem 4 (Shared-point-mass exactness with posterior learning)**.** *Fix $M, N \geq 1$, $\delta \in (0,1)$, a prior P before sampling, and i.i.d. binary-loss observations. With probability at least $1-\delta$, $L(Q) \leq U_{\mathrm{sq}}$ simultaneously for all posteriors Q, including those singular with respect to P. For every fixed loss matrix, U_{sq} is jointly continuous in (P, Q) on the entire product simplex and equals $\mathrm{CP}_{N,\delta}(k_h)$ at $P = Q = \delta_h$. On every sample, for $K = \mathrm{KL}(Q\|P) < \infty$,*

$$U_{\mathrm{sq}} \leq \min\left\{1, \widehat{L}_S(Q) + \sqrt{\frac{K + \log(5/\delta)}{N}}\right\}. \tag{8}$$

CP exactness concerns endpoint equality; discrete coverage need not equal $1-\delta$. The three failure budgets sum to δ; change of measure controls all posteriors through KL. Continuity follows by slightly shrinking feasible vectors to preserve slack under nearby priors, including constraints that disappear at the simplex boundary (Appendix H.2).

Calibrating binomial tail probabilities into evidence (Vovk & Wang, 2021) strengthens the same construction. The resulting endpoint U_{bin} retains coverage, continuity, shared-point-mass CP equality, and the KL guarantee, with

$$U_{\mathrm{bin}} \leq U_{\mathrm{sq}}, \qquad d_+(\widehat{L}_S(Q), U_{\mathrm{bin}}) \leq \frac{2[K + \log(5/\delta)]}{N}, \tag{9}$$

where $d_+(x, r) = \mathrm{kl}(x, r)$ for $r > x$ and zero otherwise. At zero empirical risk, this gives excess $O((K + \log(1/\delta))/N)$. Appendix H gives the proof and computation.

Known, globally loss-equivalent predictors can be aggregated before sampling to remove dependence on duplicate labels; observed agreement alone is insufficient (Appendix J, Proposition 47; see also Aladrah & Anselmi, 2026).

5 Calibration Precision and Posterior Width

Transport gives exact mixed-prior recovery with an unavoidable width cost. Under approximate calibration, logarithmic mixture tests attain the optimal width.

5.1 Exact Recovery at a Mixed Prior

Fix P before sampling. To obtain a binomial count from possibly nonbinary prior-averaged losses, augment each observation with an independent $V_i \sim \mathrm{Unif}(0,1)$ (Foong et al., 2021, Appendix I.2). Let K_P count observations for which $V_i \leq \sum_h P_h \ell(h, Z_i)$. Then $K_P \sim \mathrm{Bin}(N, L(P))$ on (S, V). Writing $c = \mathrm{CP}_{N,\delta}(K_P)$, define

$$T(P, Q, c) = \max\{Q \cdot r : r \in [0,1]^M,\ P \cdot r \leq c\}.$$

On the CP event $L(P) \leq c$, the true risk vector is feasible, so $L(Q) \leq U_{\mathrm{tr}}(P, Q) = T(P, Q, c)$ for all $Q = Q(S, V)$, under every binary-loss law. For every finite library, $N \geq 1$, and $\delta \in (0,1)$, it has $U_{\mathrm{tr}}(P, P) = c$ and excess above c between zero and $\mathrm{TV}(P, Q) \leq \sqrt{\mathrm{KL}(Q\|P)/2}$ (Appendix I, Proposition 42). This answers Open Problem 2 under randomized Gibbs evaluation for finite binary-loss libraries; total variation specifies graceful recovery.

A two-classifier test. For $P = (1/2, 1/2)$ and $Q = (1, 0)$, $T(P, Q, c) = \min\{1, 2c\}$. On constant loss vector $(0, 1)$, this tends to one although the first classifier is perfect. Could individual losses improve the certificate? The obstruction below applies even to rules using the entire loss matrix.

Theorem 5 (Exact recovery can conceal a perfect classifier)**.** *Fix $M = 2$, $P = (1/2, 1/2)$, and $\delta \in (0, 1)$. Let U_N be measurable $[0, 1]$-valued rules independent of the unknown data law. For every coordinate-mean vector $r \in [0, 1]^2$ in the full product-Bernoulli model, assume* (i) *the designated count $K_P \sim \mathrm{Bin}(N, P \cdot r)$;* (ii) *$U_N(P) = \mathrm{CP}_{N,\delta}(K_P)$ almost surely; and* (iii) *$\mathbb{P}_r\{Q \cdot r \le U_N(Q)$ for all posteriors $Q\} \ge 1 - \delta$. On the fixed law with loss vector $(0, 1)$ at every observation, ordinary ERM always selects $Q = (1, 0)$ and*

$$\widehat{L}_S(Q) = L(Q) = 0, \qquad \mathrm{KL}(Q\|P) = \log 2, \qquad U_N(Q) \xrightarrow{p} 1.$$

Why exactness blocks learning. Consider a population whose prior risk lies just above a CP threshold. Failure at the prior then consumes arbitrarily close to all of δ. If the rule also undercovers a fixed posterior outside that failure event on a set of positive probability, their union violates common coverage. Varying the population up to the threshold therefore forces, almost surely for each fixed Q,

$$U_N(Q) \ge T(P, Q, \mathrm{CP}_{N,\delta}(K_P - 1)) \quad \text{on } K_P \ge 1.$$

The preceding count separates prior failures from samples on which no further failure is affordable. The lower bound tends to one on this task and applies even to rules using all individual losses. Appendix I gives the converse and characterizes when deterministic binomialization is possible.

5.2 Releasing a Calibration Budget for Learning

We retain common coverage while allowing the prior endpoint to exceed CP by at most ϵ_N on each sample, including evaluation randomness. This pathwise tolerance controls inflation; average or high-probability calibration would allow larger samplewise deviations. Write $\lambda_N = N^{-1} \log(1/\epsilon_N)$ for exponential precision.

A change in the data law can raise posterior risk without changing prior risk. For $D = \delta_{(0,1)}$ with uniform prior and KL budget $\log 2$, the alternative $E_t = (1 - t)D + t\delta_{(1,0)}$, $0 \le t < 1$, leaves prior risk at $1/2$ but raises the risk of $Q = (1, 0)$ from zero to t. The likelihood cost is $\mathrm{KL}(D\|E_t) = -\log(1 - t)$, so a cost of λ allows a risk increase of $1 - e^{-\lambda}$.

To quantify this effect, fix $\mathcal{Z} = \{0, 1\}^M$, a positive prior P, a KL budget $K \ge 0$, and $\delta, \tau \in (0, 1)$, all fixed as N grows. For a law D on $\mathcal{Z}$, write $r(D) = \mathbb{E}_D Z$ and $\hat{r} = N^{-1} \sum_i Z_i$, where $Z_1, \ldots, Z_N$ are independent draws from D. Coordinate losses within each observation may be dependent. Using the randomized count K_P above, let $\mathcal{A}_N^+(\epsilon)$ consist of jointly measurable $[0, 1]$-valued rules that, under every D, cover all posteriors on one event of probability at least $1 - \delta$ and satisfy $U_N(P) \le \mathrm{CP}_{N,\delta}(K_P) + \epsilon$ almost surely over both the sample and evaluation randomness. For such a rule, $W_N(U)$ is the supremum of $[U_N(Q) - Q \cdot \hat{r}]_+$ over $\mathrm{KL}(Q\|P) \le K$. We measure its typical width by the $(1 - \tau)$-quantile under (D, V). Here δ controls failure of coverage, whereas τ controls how often the width may exceed its quantile. The infimum of this quantile over $\mathcal{A}_N^+(\epsilon)$ defines $\mathfrak{R}_{N,+}^{\mathrm{loc}}(D; \epsilon)$. Taking the worst quantile over D before the infimum over rules gives $\mathfrak{R}_{N,+}^{\mathrm{glob}}(\epsilon)$.

The local limit is determined by the largest posterior risk increase under an alternative law E that does not increase prior risk and has likelihood cost at most λ:

$$A_\lambda(D; P, K) = \sup_{\substack{\mathrm{KL}(Q\|P) \le K,\ \mathrm{KL}(D\|E) \le \lambda \\ P \cdot r(E) \le P \cdot r(D)}} [Q \cdot (r(E) - r(D))]_+. \tag{10}$$

Here E is a law on $\mathcal{Z}$, and $\mathrm{KL}(D\|E)$ measures the cost of observing D-typical data under E. For the limit over all data laws, define $\Gamma(P, K)$ as the maximum of $Q \cdot v$ over $v \in [-1, 1]^M$ and posteriors Q satisfying $P \cdot v = 0$ and $\mathrm{KL}(Q\|P) \le K$, and $\nu(\lambda)$ as the maximum of $b - a$ over $0 \le a \le b \le 1$ with $\mathrm{kl}(a, b) \le \lambda$. Thus $\Gamma(P, K)$ measures how much a posterior can respond to risk changes that cancel under the prior, while $\nu(\lambda)$ measures the separation allowed by the likelihood cost, with $\nu(0) = 0$ and $\nu(\infty) = 1$. With $K = 0$ there is no reweighting, so $\Gamma(P, 0) = 0$. The product $\Gamma(P, K)\nu(\lambda)$ gives the worst-law limit (Appendix K).

Theorem 6 (Calibration precision and optimal width)**.** *Fix M, P, K, δ, τ as above. If $\epsilon_N > 0$, $\epsilon_N \to 0$, and $\lambda_N \to \lambda \in [0, \infty]$, then*

$$\begin{aligned} \mathfrak{R}^{\mathrm{loc}}_{N,+}(D; \epsilon_N) &\longrightarrow A_\lambda(D; P, K) \quad \textit{for every fixed } D, \\ \mathfrak{R}^{\mathrm{glob}}_{N,+}(\epsilon_N) &\longrightarrow \Gamma(P, K)\nu(\lambda). \end{aligned} \tag{11}$$

At $\lambda = \infty$, the constraint $\mathrm{KL}(D\|E) \le \lambda$ is omitted. Zero-tolerance one-sided calibration has limits $A_\infty(D; P, K)$ and $\Gamma(P, K)$. At each finite limiting exponent, one explicit sequence independent of D, K, τ attains all local limits and the global limit. It also satisfies $|U_N(P) - \mathrm{CP}_{N,\delta}(K_P)| \le \epsilon_N$. The lower bounds apply to every one-sided admissible rule, without continuity assumptions in Q.

At zero tolerance, worst-law width tends to one exactly when the KL budget permits concentration on a nonempty subset with prior mass at most $1/2$ (Corollary 62). For priors $(1/2, 1/2)$ and $(0.9, 0.1)$, the thresholds are $\log 2$ and $\log 10$: concentrating on the rarer classifier requires a larger KL budget.

Equal limits, different rates. On the fixed task $D = \delta_{(0,1)}$ with $P = (1/2, 1/2)$, $K = \log 2$, $\delta + \tau < 1$, and $\epsilon_N = N^{-1}$, the optimal local width is $(\log N)/(2N) + O(N^{-1})$ under one-sided calibration, but $\Theta(N^{-1/2})$ under two-sided calibration (Appendix M). The latter retains a prior-width floor because the width supremum includes $Q = P$; one-sided calibration permits a lower prior endpoint.

Why the lower and upper bounds meet. A CP cutoff leaves at most $N\epsilon_N$ probability for posterior failures on its protected event. Under a risk-increasing alternative E, D-typical data have exponential rate $\mathrm{KL}(D\|E)$; conditional evaluation randomization cancels from the likelihood ratio. Below rate λ, these data exceed the failure budget, forcing coverage of the higher risk. Optimizing alternatives gives (10). The upper construction spends a test budget proportional to ϵ_N, yielding the same penalty λ while retaining joint loss information.

5.3 Joint Information and Uniform Attainment

Fix $P = (1/2, 1/2)$ and $K = \log 2$. Under $D_s = (\delta_{00} + \delta_{11})/2$, both classifiers err together; under $D_o = (\delta_{10} + \delta_{01})/2$, exactly one errs. Every posterior has risk $1/2$ under both laws, yet

$$A_\lambda(D_s; P, K) = \frac{1 - e^{-\lambda}}{2}, \qquad A_\lambda(D_o; P, K) = \frac{\sqrt{1 - e^{-2\lambda}}}{2}.$$

Both use the full four-vector model (Appendix L, (186)). Their different joint errors allow different risk shifts invisible to the prior, despite identical risks and KL budgets.

Theorem 7 (Uniform attainment of local frontiers)**.** *Fix $0 < \lambda < \infty$, $0 \le K < \infty$, $\delta = 1/20$, $\tau = 1/10$, and $\epsilon_N = e^{-\lambda N}$. The optimal local quantiles $\mathfrak{R}^{\mathrm{loc}}_{N,+}(D; \epsilon_N)$ converge to $A_\lambda(D; P, K)$ uniformly over finite M, positive P, and binary-loss laws D. One explicit rule U_N, independent of D, K, τ, has finite-sample simultaneous coverage and pathwise two-sided calibration error at most $\epsilon_N/4$. For every $a > 0$,*

$$\sup_{M,P,D} \Pr_{D,V} \{|W_N(U_N) - A_\lambda(D; P, K)| > a\} \longrightarrow 0.$$

The rule adapts to each law's joint errors without knowing the law, attaining its local frontier. Uniformity allows M, P, D to vary with N, with additive error at fixed λ, K.

Proof idea. The reverse-KL dual (Van Parys et al., 2021) tests contrasts $Q - \beta P$ with $\beta \ge 0$ through mixtures $X_R = \sum_j R_j X_j$ of the $2M$ features $z_h, 1 - z_h$. Empirical averages of $\log(t + X_R)$, $t > 0$, retain whether classifiers err together; coordinatewise log averages would discard this information. A common mixture test controls these averages through KL, with errors independent of library size. Finite dual grids and a CP budget adjustment give a valid rule approaching each law's frontier. The matching converse smooths alternative laws to control likelihood-ratio tails without a minimum atom mass. Prior-risk boundary estimates complete the uniform control. Appendix L proves both bounds and gives the construction and its evaluation cost for a fixed posterior. Appendix N treats bounded means.

6 Numerical Illustrations

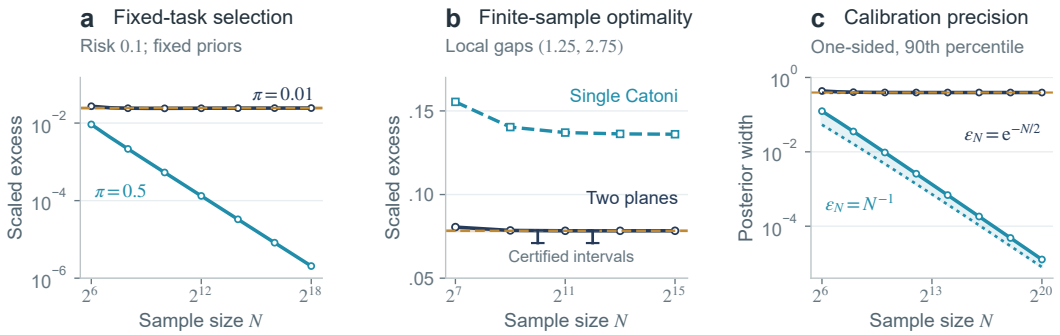


Figure 2: Finite-sample evidence ($\delta = 0.05$). (a) Catoni excess for fixed-task ERM. (b) Construction upper bounds and full-class intervals ($N = 1024, 4096$). Excess in (a,b) is scaled by $\sqrt{N}$. (c) One-sided width bounds on $D = \delta_{(0,1)}$, $P = (1/2, 1/2)$, $K = \log 2$: solid upper, dotted lower. Gold: limiting optima. Protocols and full curves: Appendices O.6 and P.

We examine selection cost, finite-sample optimality, and calibration width at $\delta = 0.05$ (Figure 2). ERM expectations and calibration quantiles are evaluated from count laws without Monte Carlo sampling. In (a), risk stays 0.1: the unequal-prior Catoni excess, scaled by $\sqrt{N}$, has a limit within 1.1×10^{-5} of the certified optimum $V_\rho \approx 0.02441$; the balanced-prior curve decays. These are attainable construction values, not finite-N optima (Appendix O.6). The contrast isolates the effect of prior weights: the task and population risk stay fixed, while unequal weights make ERM's two possible choices favor different comparator parameters.

Panel (b) compares single-Catoni and two-plane constructions on designed local gaps $(1.25, 2.75)$. At $N = 32768$, their scaled excesses are 0.1361 and 0.078336. Certified full-class intervals at $N = 1024, 4096$ have relative gaps $(U - L)/U < 9.6\%$. The lower bounds rule out smaller average endpoint excess over the full normalized convex class. The upper bounds validate explicit constructions over the entire population-risk interval (Appendix P). The two-plane construction reduces excess above the statewise benchmark. These intervals distinguish improvement over a single Catoni formula from proximity to the optimum over the full comparator class.

On the fixed zero-risk task in (c), 90th-percentile width brackets shrink toward zero for $\epsilon_N = N^{-1}$, but approach $1 - e^{-1/2}$ for $e^{-N/2}$. The upper curve uses the preselected-atom rule; the lower curve applies to every one-sided admissible rule. Both tolerances tend to zero, but only the exponential choice retains a positive limiting width. Thus the precision demanded at the prior affects what can be certified about the perfect predictor, whose risk remains zero throughout. Appendix O adds conditional resampling and matched-posterior comparisons of evidence refinement.

7 Discussion and Conclusion

PAC–Bayes tightness depends on both the learned predictor and the requirements imposed on its risk bound. For normalized convex comparators, we characterize exact compatibility and sharp selection costs in two regimes of fixed finite interior targets with positive weights. These costs reflect conflicting formula preferences and the sensitivity of risk endpoints. A fixed ERM task with low risk and bounded posterior KL incurs excess of the same order as the benchmark width, disproving universal attainability in Foong et al.'s Open Problem 1. CP recovery when prior and posterior share a point mass remains compatible with learning. For finite libraries with binary losses, randomized Gibbs evaluation answers Open Problem 2 by recovering CP exactly at the prior while retaining simultaneous posterior coverage, with total variation controlling the departure from CP. Yet requiring recovery at the original confidence level throughout the full model can force a perfect predictor's bound toward one. Allowing pathwise calibration tolerance yields matching limits for width quantiles under each distribution and in the worst case in fixed finite models. One explicit rule uses joint losses to attain each distribution's frontier without knowing the distribution. At a fixed positive precision exponent and KL budget, attainment is uniform over growing finite libraries, positive priors and data laws under the stated confidence levels.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, manuscript organization, presentation refinement, figure preparation, and LaTeX preparation. The research questions, mathematical claims, constructions, theorem statements, proof strategies, and proofs originated with and were developed by the authors. Generative AI tools were also used as auxiliary verification tools to stress-test proofs already developed by the authors and to help identify potential gaps, hidden assumptions, edge cases, or inconsistencies. All mathematical claims and proofs were independently verified, revised where necessary, and finalized by the authors. The authors take full responsibility for the correctness and content of the final manuscript.

ETHICS STATEMENT

This work studies theoretical properties of risk guarantees. Its numerical illustrations use synthetic data and the publicly available Digits and Breast Cancer benchmark datasets. We do not collect new human-subject data or make claims about clinical decision-making or deployment. The experiments evaluate risk certificates under the protocols and limitations stated in Appendix O.

REPRODUCIBILITY STATEMENT

The statistical setting and assumptions are stated in Section 2 and alongside the corresponding results. Complete proofs and explicit constructions are provided in the appendices, whose opening roadmap links the main results to their proofs. Appendix O describes the data splits, classifier libraries, posterior selection, resampling protocols, and numerical illustrations, including the count-law evaluations in Appendix O.6. Appendix P details the endpoint optimization, normalization checks, and interval-arithmetic certification methods.

REFERENCES

Reuben Adams, John Shawe-Taylor, and Benjamin Guedj. Controlling multiple errors simultaneously with a pac-bayes bound. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.nips.cc/paper_files/paper/2024/file/09887fac6cb071922e870090ce32aeff-Paper-Conference.pdf.

Nicola Aladrah and Fabio Anselmi. PAC–Bayes bounds on quotient parameter spaces: Geometry-induced implicit-bias priors, 2026. URL https://arxiv.org/abs/2607.18422. arXiv preprint arXiv:2607.18422.

Luc Bégin, Pascal Germain, François Laviolette, and Jean-Francis Roy. PAC-Bayesian Bounds Based on the Rényi Divergence. In *Proceedings of the 19th International Conference on AISTATS*, volume 51 of *Proceedings of Machine Learning Research*, pp. 435–444. PMLR, 2016. URL https://proceedings.mlr.press/v51/begin16.html.

Ioar Casado, Luis A. Ortega, Aritz Pérez, and Andrés R. Masegosa. PAC-Bayes-Chernoff Bounds for Unbounded Losses. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0767. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2b567c9aa100a8edcd13607a91ed3022-Abstract-Conference.html.

Olivier Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *Institute of Mathematical Statistics Lecture Notes–Monograph Series*. Institute of Mathematical Statistics, 2007. ISBN 978-0-940600-72-0. URL https://arxiv.org/abs/0712.0248.

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.

Jinshuo Dong, Aaron Roth, and Weijie J. Su. Gaussian differential privacy. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):3–37, 2022. doi: 10.1111/rssb.12454.

A. A. Fedotov, Peter Harremo"es, and Flemming Topsoe. Refinements of Pinsker's inequality. *IEEE Transactions on Information Theory*, 49(6):1491–1498, 2003. doi: 10.1109/TIT.2003.811927.

Andrew Y. K. Foong, Wessel P. Bruinsma, David R. Burt, and Richard E. Turner. How Tight Can PAC-Bayes Be in the Small Data Regime? In *Advances in Neural Information Processing Systems*, volume 34, pp. 4093–4105, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/214cfbe603b7f9f9bc005d5f53f7a1d3-Abstract.html.

Michael Gastpar, Ido Nachum, Jonathan Shafer, and Thomas Weinberger. Fantastic generalization measures are nowhere to be found. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/a9e8e05221b60d4161a26a00a8fd6c78-Paper-Conference.pdf.

Michael Gastpar, Ido Nachum, Jonathan Shafer, and Thomas Weinberger. Which algorithms have tight generalization bounds? In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1227. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/33ffc1a4a06e19865ac6cba04c753121-Abstract-Conference.html.

Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. PAC-Bayesian Learning of Linear Classifiers. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 353–360. ACM, 2009. doi: 10.1145/1553374.1553419.

Fredrik Hellström and Benjamin Guedj. Comparing Comparators in Generalization Bounds. In *Proceedings of the 27th International Conference on AISTATS*, volume 238 of *Proceedings of Machine Learning Research*, pp. 73–81. PMLR, 2024. URL https://proceedings.mlr.press/v238/hellstrom24a.html.

Kyoungseok Jang, Kwang-Sung Jun, Ilja Kuzborskij, and Francesco Orabona. Tighter PAC-Bayes Bounds Through Coin-Betting. In *Proceedings of the 36th Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 2240–2264. PMLR, 2023. URL https://proceedings.mlr.press/v195/jang23a.html.

Ilja Kuzborskij, Kwang-Sung Jun, Yulian Wu, Kyoungseok Jang, and Francesco Orabona. Better-than-KL PAC-Bayes Bounds. In *Proceedings of the 37th Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 3325–3352. PMLR, 2024. URL https://proceedings.mlr.press/v247/kuzborskij24a.html.

David A. McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, 1999. doi: 10.1023/A:1007618624809.

Jiahao Ming, Aaditya Ramdas, Yi Shen, Ruodu Wang, and Ian Waudby-Smith. Gaffke's confidence interval for the mean of bounded data is inadmissible but asymptotically efficient, 2026. URL https://arxiv.org/abs/2607.18661.

Vaishnavh Nagarajan and J. Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/file/05e97c207235d63ceb1db43c60db7bbb-Paper.pdf.

Luca Oneto, Sandro Ridella, and Davide Anguita. Generalization performances of randomized classifiers and algorithms built on data dependent distributions. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pp. 93–98, 2017. URL https://www.esann.org/sites/default/files/proceedings/legacy/es2017-53.pdf.

Borja Rodríguez-Gálvez, Ragnar Thobaben, and Mikael Skoglund. More PAC-Bayes bounds: From bounded losses, to losses with general tail behaviors, to anytime validity. *Journal of Machine Learning Research*, 25(110):1–43, 2024. URL https://www.jmlr.org/papers/v25/23-1360.html.

Matthias Seeger. PAC-Bayesian generalisation error bounds for Gaussian process classification. *Journal of Machine Learning Research*, 3:233–269, 2002. URL https://www.jmlr.org/papers/v3/seeger02a.html.

Tobias Sutter, Bart P. G. Van Parys, and Daniel Kuhn. A pareto dominance principle for data-driven optimization. *Operations Research*, 72(5):1976–1999, 2024. doi: 10.1287/opre.2021.0609.

Bart P. G. Van Parys, Peyman Mohajerin Esfahani, and Daniel Kuhn. From data to decisions: Distributionally robust optimization is optimal. *Management Science*, 67(6):3387–3402, 2021. doi: 10.1287/mnsc.2020.3678.

Nikos Vlassis and Philip S. Thomas. An exact distribution-free test for means of nonnegative random variables, 2026. URL https://arxiv.org/abs/2607.08415.

Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination, and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021. doi: 10.1214/20-AOS2020.

PROOF GUIDE AND NOTATION

The appendices follow two proof paths. Equation (7) connects them: no normalized scalar comparator reaches interior CP, which motivates the risk-vector construction. The calibration theorems then use confidence sets and testing arguments of their own.

Arc	Formal dependency	Proof location
Scalar comparators	Shared ceiling and information tools branch into: affine and KL representations, compatibility, and information lower bounds; a separate local-limit argument gives the Gaussian endpoint characterization	Appendices A–G
Learning-compatible CP	Strict CP interface ⇒ risk-vector set ⇒ Theorem 4 ⇒ binomial refinement, paired use, and fixed-program computation	Appendix H
Every-prior CP	Randomized binomial count ⇒ transport upper bound; exactness + common coverage ⇒ previous-count lower bound ⇒ fixed-task saturation	Appendix I
Approximate CP	Budget relaxation; fixed-model width limits; logarithmic mixture tests give a uniform local frontier for growing libraries at positive exponent; a sharp local rate on one fixed two-classifier task	Appendices J.1–M
Bounded-mean calibration	Deterministic prior recovery by transport; a root-N lower obstruction for a fixed posterior under fiber-invisible alternatives	Appendix N
Computations	Conditional resampling, transport enumeration, and endpoint certificates; a compact comparison with prior guarantees	Appendices O–Q

Scope and probability spaces. Propositions 8 and 2, and Theorem 39 concern finite interior target families. Theorem 3 fixes the number and positive weights of targets converging locally to one interior diagonal point. Theorem 1 uses a fixed low-risk task, fixed prior, and ordinary ERM; its sharp limit is for fixed $0 < \rho < 1/2$. Theorem 22 uses fair labels, a fixed prior, and odd sample sizes; its asymmetric rate is not uniform as the prior approaches symmetry. Theorem 28 permits biased labels but is only a finite-N positivity result. The all-N construction allows the task and prior to depend on N. Theorem 4 fixes finite M and binary losses, takes probability over S, and allows boundary priors and posteriors; its KL rate requires finite $\mathrm{KL}(Q\|P)$. Theorem 5 takes probability over (S, V); its converse holds almost surely for each fixed Q over the full product-Bernoulli risk-vector model. Theorem 6 fixes finite M, positive P, K, δ, and the width quantile level; it requires validity under every binary-vector law and pathwise one-sided upper calibration over (S, V). The matching construction also satisfies the stronger two-sided condition. Its limits are first-order as $N \to \infty$. The uniform extension in Theorem 7 (full statement: Theorem 65) fixes $\lambda > 0$, K, $\delta = 1/20$, and $\tau = 1/10$, but permits arbitrary finite M and varying positive P and laws D; it controls additive error. Theorem 76 gives a sharper logarithmic rate only for the fixed law $D_0 = \delta_{(0,1)}$, $P = (1/2, 1/2)$, and $K = \log 2$. Corollary 81 instead uses the deterministic Gaffke endpoint and a fixed posterior on a three-vector task. Its root-N lower obstruction is not a matching width frontier or a claim about the ERM-selected posterior.

Object	Mathematical guarantee	Computational status
Comparator score optimum	Exact affine representation; uniform KL localization for fixed interior targets	Convex semi-infinite score program (Appendix A)
Comparator endpoint optimum	Exact finite affine representation; uniform interior transfer and a separate near-diagonal Gaussian limit	Full-class finite-N brackets and limiting brackets are distinct (Appendix P)
$U_{\mathrm{sq}}, U_{\mathrm{bin}}$	Exact support functions of the stated finite convex risk sets	Convex dual computation; Proposition 40 gives exact fixed-program compression
U_{tr}	Exact every-prior randomized-evaluation endpoint	Fractional knapsack, solved exactly after sorting in $O(M \log M)$

We use the main-text notation throughout. In the proofs, $\widehat{R}_S = \widehat{L}_S$ and $R_D = L$ are equivalent risk symbols, and $B[f(q,\cdot), c] = U_f(q,c)$ is the rightmost feasible endpoint, with value zero for an empty sublevel set and one at infinite complexity. $\mathcal{F}_N$ denotes the normalized convex class; $\mathcal{C}$ denotes source comparators before normalization. The symbol H denotes entropy when its argument is a probability vector and the tangent envelope in the compatibility proof. Auxiliary binomial laws in the comparator proofs are proof devices, not the observed loss counts of their learning tasks. The symbols U_{sq} and U_{bin} denote support functions of the squared-deviation and binomial-evidence risk sets. In randomized general-prior evaluation, K_P is an observed augmented-sample count and U_{tr} is the distinct scalar-transport endpoint.

Table 1: Requirements and scope of the main comparisons. The growing-library extension is stated in Theorem 7 (Section 5.3).

Requirement	Scope
One comparator for several states	Normalization (1); finitely many positively weighted targets whose empirical risks and larger benchmark bounds lie in $(0,1)$ (Section 3.2).
CP at a shared point mass	Finite binary-loss library; $P = Q = \delta_h$; simultaneous coverage for all Q (Section 4).
CP at a mixed prior	Binomial evaluation count and original-confidence CP equality; simultaneous coverage throughout the model of Theorem 5.
Approximate prior CP	Prior endpoint at most CP $+ \epsilon_N$ on every augmented sample; validity under every law on a fixed finite binary-loss alphabet, positive P (Section 5.2).

A FINITE TANGENT REPRESENTATION

The following compact-interior representation supports the separated-target analysis. The exact affine reductions and the proof follow below.

Proposition 8 (Finite tangent representation)**.** *Fix m, weights $w_j \geq w_{\min} > 0$ summing to one, and targets $x_{j,N} = (q_{j,N}, p_{j,N})$ in a compact $\mathcal{K} \Subset \{0 < q < p < 1\}$. For all sufficiently large N, the full-class score infimum equals that over normalized maxima of at most m shifted KL tangents with contacts in one compact convex interior neighborhood $\mathcal{K}'$. Writing $g(q,p) = \mathrm{kl}(q,p)$ and $s(q,p) = p(1-p)/(p-q)$,*

$$\mathcal{G}_N^\star(\mathsf{S}_N) = \inf_{f \in \mathcal{F}_N} \sum_j w_j s(x_{j,N})\{g(x_{j,N}) - f(x_{j,N})\} + O(N^{-3/2}). \tag{12}$$

The neighborhood, threshold, and remainder depend only on $(m, w_{\min}, \mathcal{K})$.

A.1 AN EXACT AFFINE SCORE PROGRAM AT EVERY SAMPLE SIZE

The finite affine score representation holds at every sample size. It is the convex starting point; uniform KL localization and conversion to endpoint costs require the additional arguments in Proposition 8.

Proposition 9 (Finite affine score representation)**.** *Fix $N \geq 1$, $m \geq 1$, targets $x_i \in (0,1)^2$, and nonnegative weights w_i summing to one. Every $f \in \mathcal{F}_N$ admits affine functions $\ell_1, \dots, \ell_m$ such that $H = \max_j \ell_j \in \mathcal{F}_N$ and $H(x_i) = f(x_i)$ for all i. Consequently, $\sup_{f \in \mathcal{F}_N} \sum_i w_i f(x_i)$ equals the value of the following convex semi-infinite program, with finite-dimensional variables $a_j \in \mathbb{R}^2$, $b_j, v_j \in \mathbb{R}$:*

$$\begin{aligned} &\textit{maximize} \quad \sum_i w_i v_i, \\ &a_i^\top x_i + b_i = v_i \quad \textit{for all } i, \\ &a_j^\top x_i + b_j \leq v_i \quad \textit{for all } i, j, \\ &\sum_{k=0}^{N} \binom{N}{k} r^k (1-r)^{N-k} \exp\Big\{ N \max_j [a_j^\top (k/N, r) + b_j] \Big\} \leq 1 \quad \textit{for every } r \in [0,1]. \end{aligned} \tag{13}$$

At $r = 0, 1$, zero-probability terms are omitted. Values are suprema; attainment is not asserted.

Proof. For each $r \in (0,1)$, all binomial grid probabilities are positive. Normalization makes $f(k/N, r)$ finite, and convexity in the first coordinate then makes $f(q,r)$ finite throughout the open square. An interior target therefore has a supporting affine function $\ell_i \leq f$ with $\ell_i(x_i) = f(x_i)$. Their maximum is finite, continuous, and convex, satisfies $H \leq f$, and agrees with f at every target. Monotonicity of the moment functional gives $I_{H,N} \leq I_{f,N} \leq 1$. The contact constraints in (13) say exactly that $\max_j \ell_j(x_i) = v_i$. Thus every comparator supplies a feasible tuple with the same objective, and every feasible tuple supplies a comparator H with that objective. For fixed r, each summand in the moment constraint is a nonnegative multiple of the exponential of a maximum of affine functions of the coefficients, hence convex. Intersecting these constraints over r preserves convexity; all remaining constraints are affine. □

The same support argument also gives an exact finite-sample representation of any finite vector of rightmost endpoints. This does not make their joint optimization convex: the contact locations at which the planes must be taken depend on the comparator being represented.

Proposition 10 (Finite affine endpoint representation)**.** *Fix $N \geq 1$ and finitely many pairs (q_i, t_i) with $q_i \in (0,1)$ and $t_i > 0$. For every $f \in \mathcal{F}_N$, there is a maximum H of at most one affine function per target such that $H \in \mathcal{F}_N$ and*

$$U_H(q_i, t_i) = U_f(q_i, t_i) \qquad \textit{for every } i.$$

If all the displayed endpoints equal one, one may take $H \equiv 0$. Consequently, every finite-target endpoint infimum over $\mathcal{F}_N$ is unchanged if the comparator class is restricted to maxima of at most one affine component per target. Attainment of that infimum is not asserted.

Proof. Put $u_i = U_f(q_i, t_i)$. The KL ceiling gives $f(q_i, q_i) \le 0 < t_i$, so q_i is feasible and $u_i \ge q_i$. Suppose first that $u_i < 1$. The vertical sublevel set

$$\{r \in [0,1] : f(q_i, r) \le t_i\}$$

is closed by lower semicontinuity. Hence its supremum is attained at u_i. The comparator is finite and continuous on the open square, so strict inequality at (q_i, u_i) would make a point immediately to its right feasible. Therefore $u_i > q_i$ and $f(q_i, u_i) = t_i$.

Take a supporting affine function $L_i(q, r) = A_i q + D_i r + C_i \le f(q, r)$ at (q_i, u_i). Since

$$L_i(q_i, q_i) \le f(q_i, q_i) \le 0 < t_i = L_i(q_i, u_i),$$

its vertical slope D_i is strictly positive. Let H be the maximum of these L_i over the unsaturated targets. Then $H \le f$, so H is convex, lower semicontinuous, proper, and satisfies $I_{H,N} \le I_{f,N} \le 1$. At every unsaturated target, $H(q_i, u_i) = t_i$, while for every $r > u_i$,

$$H(q_i, r) \ge L_i(q_i, r) > t_i.$$

Because q_i remains feasible under $H \le f$, this proves $U_H(q_i, t_i) = u_i$.

If $u_i = 1$, closedness of the same nonempty sublevel set gives $f(q_i, 1) \le t_i$. Thus $H(q_i, 1) \le f(q_i, 1) \le t_i$, so its endpoint is also one. When every target is saturated, $H \equiv 0$ is normalized and has endpoint one at every positive threshold. Applying this construction to each comparator, while the reverse restriction is immediate, proves the infimum statement. □

Neither proposition identifies the support slopes with KL contact coordinates or bounds those coordinates uniformly. Those are additional steps in Proposition 8, together with the uniform score-to-endpoint expansion. Nonconvexity of the contact-coordinate parameterization is therefore not intrinsic to score optimization, but can remain in finite-sample endpoint optimization. Computationally, finitely many r constraints give a relaxation, while a global upper bound $B \ge I_{H,N}$ makes $H - N^{-1} \log B$ feasible. A valid upper bound on the relaxed score optimum and the score of this normalized candidate bracket the full-class score optimum. A sampled maximum or an uncertified solver objective alone does not establish such a bracket.

Proposition 11 (A barycentric upper certificate for the optimal score)**.** *Choose $r_l \in (0,1)$ and nonnegative numbers γ_{ilk} satisfying*

$$\sum_{l,k} \gamma_{ilk} = w_i, \qquad \sum_{l,k} \gamma_{ilk}(k/N, r_l) = w_i x_i.$$

Put $\mu_{lk} = \sum_i \gamma_{ilk}$, $\beta_l = \sum_k \mu_{lk}$, and $\pi_{lk} = \binom{N}{k} r_l^k (1 - r_l)^{N-k}$. Then

$$\sup_{f \in \mathcal{F}_N} \sum_i w_i f(x_i) \le \frac{1}{N} \sum_{l,k} \mu_{lk} \log \frac{\mu_{lk}}{\beta_l \pi_{lk}},$$

with zero-mass terms omitted.

Proof. For $w_i > 0$, convexity bounds $w_i f(x_i)$ by $\sum_{l,k} \gamma_{ilk} f(k/N, r_l)$. For each l with $\beta_l > 0$, the log-sum inequality and normalization give

$$N \sum_k \frac{\mu_{lk}}{\beta_l} f(k/N, r_l) \le \mathrm{KL}(\mu_{l\cdot}/\beta_l \| \pi_{l\cdot}) + \log \sum_k \pi_{lk} e^{N f(k/N, r_l)} \le \mathrm{KL}(\mu_{l\cdot}/\beta_l \| \pi_{l\cdot}).$$

Multiply by β_l, sum, and take the supremum over f. □

This upper certificate only needs feasible barycentric masses, not an optimal solver output or exact floating-point stationarity. With rational targets, weights, grid points, and masses, the barycenter equations can be checked exactly; outward interval evaluation of the logarithms then bounds the score. It pairs with global normalization of an affine candidate to give a full-class optimality bracket.

A.2 Finite tangent families and their normalization

Write $g(q,p) = \mathrm{kl}(q,p)$, and let $T_y g$ denote its affine tangent at y. A nonempty family $\mathcal{C} = \{(y_i, c_i)\}_{i=1}^{L}$, $L \le m$, records contacts and positive relative heights. Scaling all c_i together changes neither its comparator nor its deficits. Define

$$\Lambda_N(\mathcal{C}) = \sup_{r\in[0,1]} \mathbb{E}_{K\sim\mathrm{Bin}(N,r)} \max_{i\le L} c_i e^{N T_{y_i} g(K/N,r)}, \tag{14}$$

$$\Delta_N(x;\mathcal{C}) = \min_{i\le L} \left\{ B_g(x,y_i) + \frac{1}{N}\log\frac{\Lambda_N(\mathcal{C})}{c_i} \right\}, \qquad B_g(x,y) = g(x) - T_y g(x). \tag{15}$$

The Bregman term measures contact mismatch; the logarithmic term pays for joint normalization. Each family supplies the legal comparator

$$f_{\mathcal{C}}(x) = \max_i \{T_{y_i} g(x) + N^{-1}\log c_i\} - N^{-1}\log\Lambda_N(\mathcal{C}). \tag{16}$$

Since each unshifted tangent has unit binomial moment, $\max_i c_i \le \Lambda_N(\mathcal{C}) \le \sum_i c_i$. For the targets and compact contact region of Proposition 8, the full score and endpoint formulations are

$$\inf_{f\in\mathcal{F}_N} \sum_j w_j\{g(x_{j,N}) - f(x_{j,N})\} = \inf_{\substack{|\mathcal{C}|\le m \\ y_i\in\mathcal{K}'}} \sum_j w_j \Delta_N(x_{j,N};\mathcal{C}), \tag{17}$$

$$\mathcal{G}_N^\star(\mathsf{S}_N) = \inf_{\substack{|\mathcal{C}|\le m \\ y_i\in\mathcal{K}'}} \sum_j w_j s(x_{j,N}) \Delta_N(x_{j,N};\mathcal{C}) + O(N^{-3/2}). \tag{18}$$

The remainder is uniform over the fixed finite interior target class.

A.3 Uniform KL localization and endpoint conversion

We give the two ingredients behind Proposition 8. All constants below are uniform over a fixed compact $\mathcal{K} \Subset \{0 < q < p < 1\}$.

Lemma 12 (Global KL ceiling)**.** *Every normalized convex comparator satisfies $f(q,p) \le g(q,p) = \mathrm{kl}(q,p)$ on $(0,1)^2$.*

Proof. Change measure from $\mathrm{Bin}(N,p)$ to $\mathrm{Bin}(N,q)$. The log likelihood ratio is affine in K, so Jensen's inequality and convexity of $f(\cdot,p)$ give

$$1 \ge \mathbb{E}_{\mathrm{Bin}(N,p)} e^{N f(K/N,p)} \ge \exp\{N f(q,p) - N\,\mathrm{kl}(q,p)\}.$$

Taking logarithms proves the claim. □

Lemma 13 (Uniform tangent representation)**.** *There are a compact convex neighborhood $\mathcal{K}'$, $\varepsilon_0 > 0$, and $C < \infty$ such that the following holds. If an affine $L \le g$ satisfies $g(x) - L(x) \le \varepsilon_0$ at $x \in \mathcal{K}$, then unique $y \in \mathcal{K}'$ and $\eta \ge 0$ satisfy*

$$L = T_y g - \eta, \qquad g(x) - L(x) = B_g(x,y) + \eta, \qquad \|x - y\| \le C\sqrt{g(x) - L(x)}. \tag{19}$$

Proof. Choose $\rho > 0$ so every $\overline{B}(x,\rho)$, $x \in \mathcal{K}$, lies in a compact convex interior set $\mathcal{K}'$. Let $0 < \mu \le M$ bound the eigenvalues of $\nabla^2 g$ there. Put $d = g(x) - L(x)$ and let h be the slope of L. Evaluating $L \le g$ at $x \pm tu$ and using Taylor's theorem gives, for every unit u,

$$|\langle h - \nabla g(x), u\rangle| \le d/t + Mt/2.$$

Taking $t = \sqrt{2d/M}$ yields $\|h - \nabla g(x)\| \le \sqrt{2Md}$; the case $d = 0$ follows by a limit. For sufficiently small d, the minimizer of $y \mapsto g(y) - \langle h, y\rangle$ over $\overline{B}(x,\rho)$ cannot lie on the boundary, by strong convexity. Thus $h = \nabla g(y)$ for a unique interior y, and strong monotonicity gives $\|y - x\| \le \mu^{-1}\sqrt{2Md}$. The difference between L and $T_y g$ is a constant; evaluating at y shows it is $-\eta$ with $\eta \ge 0$, proving the identities. □

Lemma 14 (Uniform deficits for near minimizers). *For targets $x_{j,N} \in \mathcal{K}$, put*

$$J_N(f) = \sum_j w_j\{g(x_{j,N}) - f(x_{j,N})\}, \qquad V_N = \inf_{f \in \mathcal{F}_N} J_N(f).$$

Then $0 \le V_N \le (\log m)/N$. Moreover, every $f \in \mathcal{F}_N$ with $J_N(f) \le V_N + 1/N$ satisfies

$$0 \le g(x_{j,N}) - f(x_{j,N}) \le \frac{\log m + 1}{w_{\min} N} \quad \text{for every } j. \tag{20}$$

Consequently, for all sufficiently large N, every supporting plane used at these target contacts satisfies the hypothesis of Lemma 13, uniformly over the stated target class.

Proof. The lower bound follows from Lemma 12. For the upper bound, take the maximum of the m KL tangents at the requested targets and subtract $(\log m)/N$. The maximum has binomial moment at most the sum of its m unit tangent moments, so the shifted maximum lies in $\mathcal{F}_N$. At target j, its own tangent gives deficit at most $(\log m)/N$. Thus $V_N \le (\log m)/N$. Every summand of $J_N(f)$ is nonnegative and $w_j \ge w_{\min}$, which gives (20). The right side converges uniformly to zero, so the final claim follows from the fixed ε_0 in Lemma 13. □

Lemma 15 (Uniform active-tangent endpoint inversion). *Fix compact interior sets $\mathcal{K} \subset \operatorname{int} \mathcal{K}' \Subset \{q < p\}$. There are $\varepsilon, C > 0$ such that, for every normalized component family with contacts in $\mathcal{K}'$, every $x = (q,p) \in \mathcal{K}$, and $\Delta_N(x;\mathcal{C}) \le \varepsilon$,*

$$0 \le U_{f_\mathcal{C}}(q, g(x)) - p \le s(x)\Delta_N(x;\mathcal{C}) + C\Delta_N(x;\mathcal{C})^{3/2}.$$

The constants are uniform in the number and switching pattern of the affine pieces, and the displayed endpoint lies below one.

Proof. Choose a component i attaining the minimum in Equation (15). Its normalized affine piece A_i satisfies $g(x) - A_i(x) = \Delta := \Delta_N(x;\mathcal{C})$. Moreover $\Lambda_N(\mathcal{C}) \ge c_i$, because the unshifted tangent has moment one at the population coordinate of its contact. Hence $B_g(x, y_i) \le \Delta$, and uniform strong convexity gives $\|y_i - x\| \le C_0\sqrt{\Delta}$. On $\mathcal{K}'$, the vertical slope $b_i = \partial_p g(y_i)$ is uniformly positive and $b_i^{-1} = s(x) + O(\sqrt{\Delta})$. For every $t > \Delta/b_i$, $A_i(q, p+t) > g(x)$, so the maximum defining $f_\mathcal{C}$ is infeasible there regardless of switches among the other pieces. Thus the rightmost feasible point is at most $p + \Delta/b_i$, which gives the upper bound. The KL ceiling makes p feasible, giving the lower bound. Finally compact separation of p from one and a smaller fixed ε make clipping inactive uniformly. □

Proof of Proposition 8. For any component family, the comparator in Equation (16) is convex and has normalized moment one by definition of Λ_N. Since the maximum selects the best component at each target, direct subtraction from g proves the right-to-left inequality in Equation (17).

For the converse, choose near minimizers within $\min\{1/N, k^{-1}\}$ of the infimum and let $k \to \infty$. At each target, select a supporting affine plane $L_j \le f \le g$. Lemma 14 makes every contact deficit uniformly $O(N^{-1})$, so Lemma 13 gives $L_j = T_{y_j} g - \eta_j$. Set $c_j = e^{-N\eta_j}$. Since $f \ge \max_j L_j$ and f is normalized, $\Lambda_N(\{(y_j, c_j)\}) \le 1$. At target x_j, assigning its own support plane in Equation (15) gives

$$\Delta_N(x_j;\mathcal{C}) \le B_g(x_j, y_j) + \eta_j \le g(x_j) - f(x_j).$$

Delete components serving no target, average, and let the near-optimality error vanish. This proves equality.

For endpoint transfer, denote the right side of Equation (18) without its remainder by $\mathcal{C}_N^\star$, and use two separate N^{-2}-near minimizers: a component family for $\mathcal{C}_N^\star$ and a normalized comparator for the endpoint problem. The one-tangent-per-target construction bounds both optimal objectives by C/N.

We first record the comparison from endpoint excess to score deficit. Write $d_f(q,p) = g(q,p) - f(q,p)$ and $e_f(q,p) = U_f(q, g(q,p)) - p$. Fix a small $h > 0$, uniformly allowed on the compact target region, and put $M_{q,p}(h) = g(q, p+h) - g(q,p)$. Convexity of the vertical slice of f, together with $f(q, p+h) \le g(q, p+h)$, gives

$$e_f(q,p) \ge h\,\frac{d_f(q,p)}{d_f(q,p) + M_{q,p}(h)}. \tag{21}$$

Thus $d_f \le Ce_f$ whenever e_f is small. Since m is fixed and $w_j \ge w_{\min} > 0$, the endpoint-near-minimizing comparator consequently has $e_j, d_j = O(N^{-1})$ at every target. To obtain the sharp local direction, use compactness to choose $\bar{h} > 0$ and $L < \infty$ such that for $0 < h \le \bar{h}$, uniformly over the target region,

$$M_{q,p}(h) \le h/s(q,p) + Lh^2.$$

For $d_f(q,p) > 0$, substitute $h = \sqrt{d_f(q,p)}$ in Equation (21). Uniformly expanding the denominator gives

$$e_j \ge s(x_j)d_j - Cd_j^{3/2}. \tag{22}$$

The case $d_j = 0$ follows from endpoint nonnegativity. This argument uses only the KL ceiling and vertical convexity of f; the supporting plane is needed later to recover the component family.

For the independent component-family near minimizer, positive weights give $\Delta_j = O(N^{-1})$. Lemma 15 gives $e_j \le s(x_j)\Delta_j + C\Delta_j^{3/2}$, including arbitrary switches among active affine pieces. Summing proves

$$\mathcal{G}_N^\star \le \mathcal{C}_N^\star + O(N^{-3/2}).$$

For the reverse direction, put

$$A_N = \sum_j w_j s(x_j), \qquad \bar{w}_{j,N} = \frac{w_j s(x_j)}{A_N}.$$

Compactness gives constants $0 < s_- \le s(x) \le s_+ < \infty$ on $\mathcal{K}$. Hence $s_- \le A_N \le s_+$ and $\bar{w}_{j,N} \ge w_{\min} s_-/s_+$, so the exact score reduction uses one threshold depending only on $(m, w_{\min}, \mathcal{K})$. The same compact set bounds the vertical slope away from zero and supplies positive distance to $p = 1$; enlarge that threshold once to make localization, both Taylor remainders, and absence of clipping simultaneous. Apply the exact score reduction with normalized weights $\bar{w}_{j,N}$ and multiply the result by A_N. Equation (22) then gives

$$\mathcal{G}_N^\star + N^{-2} \ge \sum_j w_j s(x_j) d_j - O(N^{-3/2}) \ge \mathcal{C}_N^\star - O(N^{-3/2}).$$

This proves both sides of Equation (18), with one uniform remainder constant and the asserted common N_0. □

Equivalent weighted-score form. The formula (12) is the same endpoint transfer written without contact coordinates. Put $A_N = \sum_j w_j s(x_{j,N})$ and $\bar{w}_{j,N} = w_j s(x_{j,N})/A_N$. On the fixed compact target class these weights have a uniform positive lower bound. Applying (17) with $\bar{w}_{j,N}$ and multiplying by A_N identifies the infimum in (18) with $\inf_{f \in \mathcal{F}_N} \sum_j w_j s(x_{j,N})[g(x_{j,N}) - f(x_{j,N})]$. This is precisely the reweighting used in the preceding proof.

B Shared Comparator Tools

We collect the normalization shift, the common-binomial information inequality, and the endpoint interpolation used by the selection proofs. These statements hold at each finite sample size. They do not require an optimal comparator to exist.

B.1 The finite-sample normalizer

Let $\mathcal{C}$ be the proper, convex, lower-semicontinuous functions $\Delta : [0,1]^2 \to \mathbb{R} \cup \{+\infty\}$, with no value $-\infty$. For a fixed integer $N \ge 1$, define

$$I_{\Delta,N} := \sup_{r \in [0,1]} \sum_{k=0}^{N} \binom{N}{k} r^k (1-r)^{N-k} \exp\{N\Delta(k/N, r)\}. \tag{23}$$

The sum is the extended nonnegative expectation under $\mathrm{Bin}(N, r)$: at $r \in \{0, 1\}$, zero-mass terms are omitted. This avoids both 0^0 and a formal product $0e^{+\infty}$. We call f *normalized at sample size* N

when $I_{f,N} \le 1$, exactly as in Equation (1); it is *unit-normalized* when equality holds. We use the endpoint conventions in Section 2. If $0 < I_{\Delta,N} < \infty$, put

$$f = \Delta - \frac{1}{N} \log I_{\Delta,N}. \tag{24}$$

Then f remains proper, convex, and lower semicontinuous, $I_{f,N} = 1$, and

$$G_{f,N}(r) := \sum_{k=0}^{N} \binom{N}{k} r^k (1-r)^{N-k} e^{N f(k/N,r)} \le 1 \quad (0 \le r \le 1). \tag{25}$$

A finite normalizer is automatically strictly positive: at any $r \in (0,1)$ all binomial weights are positive, so a finite moment has finite grid values and is itself strictly positive. Moreover, for every finite threshold t,

$$B\left[\Delta(q,\cdot), t + \frac{1}{N} \log I_{\Delta,N}\right] = B[f(q,\cdot), t]. \tag{26}$$

Thus the shift is lossless for every endpoint used below. Conversely, if f is normalized with $I_{f,N} < 1$, shifting it upward by $-N^{-1} \log I_{f,N}$ makes it unit-normalized and can only decrease its endpoint at any unchanged threshold. Hence allowing the budget $I_{f,N} \le 1$ or restricting to its unit boundary gives the same infimum in our fixed-threshold optimization.

The remaining source-comparator branch cannot improve that infimum. Under the endpoint convention of Appendix B.1, $B[g, +\infty] = 1$, a comparator with $I_{\Delta,N} = +\infty$ reports endpoint one at every finite source threshold after the normalizer correction. The constant comparator $f \equiv 0$ is unit-normalized and reports the same endpoint at the nonnegative thresholds used in this paper. Thus the source formulation and the normalized formulation are exactly equivalent for our objectives, for every finite N.

B.2 A GENERALIZED JENSEN–SHANNON OBSTRUCTION

Fix $p \in (0,1)$, points $q_1, \dots, q_m \in (0,1)$, and positive weights w_j summing to one. Write

$$\lambda_j(k) = \binom{N}{k} q_j^k (1-q_j)^{N-k}, \qquad \bar{\lambda} = \sum_{j=1}^{m} w_j \lambda_j, \tag{27}$$

and define the per-sample generalized Jensen–Shannon radius

$$J_N := \frac{1}{N} \sum_{j=1}^{m} w_j \, \mathrm{KL}(\lambda_j \| \bar{\lambda}). \tag{28}$$

Lemma 16 (Binomial generalized-JS deficit)**.** *If $f \in \mathcal{C}$ obeys (25), then*

$$\sum_{j=1}^{m} w_j \{\mathrm{kl}(q_j, p) - f(q_j, p)\} \ge J_N. \tag{29}$$

If two positive-weight q_j's differ, then $J_N > 0$.

Proof. Let

$$\nu(k) = \binom{N}{k} p^k (1-p)^{N-k}, \qquad a_k = f(k/N, p).$$

Every $\nu(k)$ is positive. Equation (25) therefore forces every a_k to be finite and gives

$$Z = \sum_{k=0}^{N} \nu(k) e^{N a_k} \in (0, 1].$$

Consequently

$$\pi_f(k) = \frac{\nu(k) e^{N a_k}}{Z} \tag{30}$$

is a strictly positive probability mass function.

The mean of k/N under λ_j is q_j. Horizontal convexity and the binomial KL identity therefore give

$$\begin{aligned} N\{\mathrm{kl}(q_j,p) - f(q_j,p)\} &\geq \mathrm{KL}(\lambda_j\|\nu) - N\sum_k \lambda_j(k)a_k \\ &= \mathrm{KL}(\lambda_j\|\pi_f) - \log Z \\ &\geq \mathrm{KL}(\lambda_j\|\pi_f). \end{aligned}$$

For any positive probability mass function π on the common support,

$$\sum_j w_j \,\mathrm{KL}(\lambda_j\|\pi) = \sum_j w_j \,\mathrm{KL}(\lambda_j\|\bar{\lambda}) + \mathrm{KL}(\bar{\lambda}\|\pi). \tag{31}$$

Average the preceding inequality, use (31) with $\pi = \pi_f$, and discard the last nonnegative KL term. This proves (29).

If $J_N = 0$, every positive-weight λ_j equals $\bar{\lambda}$. Equal binomial laws have equal means, so all corresponding q_j's are equal. The contrapositive proves strict positivity. □

Corollary 17 (The normalized binary-KL ceiling)**.** *Every f obeying (25) satisfies $f(q,r) \leq \mathrm{kl}(q,r)$ for $q, r \in (0,1)$.*

Proof. Repeat the first part of Lemma 16 with a single state and with p replaced by r. The resulting right-hand side is nonnegative. □

For $N > 1$, the quantity in (28) is generally smaller than the Bernoulli Jensen–Shannon radius. Indeed, if $\bar{q} = \sum_j w_j q_j$, then

$$J_N = \sum_j w_j \,\mathrm{kl}(q_j,\bar{q}) - \frac{1}{N}\,\mathrm{KL}\big(\bar{\lambda}\big\|\mathrm{Bin}(N,\bar{q})\big)\,. \tag{32}$$

A nontrivial mixture of binomial laws is generally not binomial. We therefore use J_N only as the certified lower bound just proved, never as an asserted exact optimum.

B.3 Turning vertical deficit into endpoint displacement

Lemma 18 (Endpoint interpolation)**.** *Assume $0 < q_j < p < s < 1$ for every j, and that at least two positive-weight q_j's differ. Set*

$$t_j = \mathrm{kl}(q_j,p), \qquad d_j = t_j - f(q_j,p), \qquad M_j = \mathrm{kl}(q_j,s) - \mathrm{kl}(q_j,p),$$

and $b_j = B[f(q_j,\cdot), t_j]$. Then

$$b_j - p \geq (s-p)\frac{d_j}{d_j + M_j}. \tag{33}$$

Consequently, with $w_{\min} = \min_j w_j$ and $M = \max_j M_j$,

$$\sum_j w_j(b_j - p) \geq w_{\min}(s-p)\frac{J_N}{J_N + M} > 0. \tag{34}$$

Proof. Corollary 17 gives $d_j \geq 0$ and $f(q_j,s) \leq t_j + M_j$. Since

$$\partial_r \,\mathrm{kl}(q_j,r) = \frac{r - q_j}{r(1-r)} > 0 \quad \text{on } [p,s],$$

we have $M_j > 0$. Put $\alpha_j = d_j/(d_j + M_j)$. Vertical convexity gives

$$\begin{aligned} f(q_j,(1-\alpha_j)p + \alpha_j s) &\leq (1-\alpha_j)(t_j - d_j) + \alpha_j(t_j + M_j) \\ &= t_j. \end{aligned}$$

The displayed point belongs to the sublevel set defining b_j, which proves (33); no monotonicity of the slice is needed.

Lemma 16 implies that some j_* has $d_{j_*} \geq J_N$. Now $w_{j_*} \geq w_{\min}$, $M_{j_*} \leq M$, and $d/(d+M)$ increases in d and decreases in M. Keeping only this state's contribution proves (34). □

C NEAR-DIAGONAL ENDPOINT COSTS

Comparator class and endpoints. Write $\mathrm{kl}(q,r) = q\log(q/r) + (1-q)\log((1-q)/(1-r))$ with the usual extended boundary values. Let $\mathcal{F}_N$ consist of proper jointly convex lower-semicontinuous functions $f : [0,1]^2 \to \mathbb{R} \cup \{+\infty\}$, never taking $-\infty$, such that

$$I_{f,N} := \sup_{r\in[0,1]} \mathbb{E}_{K\sim\mathrm{Bin}(N,r)} e^{Nf(K/N,r)} \le 1.$$

Zero-probability terms are omitted at $r = 0, 1$. Define

$$U_f(q,t) = \sup\big(\{r \in [0,1] : f(q,r) \le t\} \cup \{0\}\big).$$

Normalization forces all binomial grid values to be finite for interior r; convexity then makes f finite and continuous on the open square. The KL ceiling $f(q,r) \le \mathrm{kl}(q,r)$ follows from Jensen in the empirical coordinate and the entropy inequality, applied to $\mathrm{Bin}(N,q)$ and $\mathrm{Bin}(N,r)$: $Nf(q,r) \le N\mathbb{E}_q f(K/N,r) \le \mathrm{KL}(\mathrm{Bin}(N,q)\|\,\mathrm{Bin}(N,r)) + \log \mathbb{E}_r e^{Nf(K/N,r)} \le N\,\mathrm{kl}(q,r)$. An affine tangent to kl at an interior point has normalizer one: writing it as $Aq + Dr + C$, the concave function $Dr + C + \log(1 - r + re^A)$ has maximum zero at the contact population coordinate.

Setup. Fix $q_0 \in (0,1)$ and put $v_0 = q_0(1-q_0)$. Let m be fixed, let $w_j > 0$ with $\sum_{j=1}^m w_j = 1$, and suppose

$$\sqrt{N}\,(q_{j,N} - q_0) \to a_j, \qquad \sqrt{N}\,(p_{j,N} - q_0) \to b_j, \qquad d_j := b_j - a_j > 0.$$

Set $t_{j,N} = \mathrm{kl}(q_{j,N}, p_{j,N})$ and

$$\mathcal{G}_N^\star = \inf_{f\in\mathcal{F}_N} \sum_{j=1}^m w_j\{U_f(q_{j,N}, t_{j,N}) - p_{j,N}\}.$$

For $\beta > 0$ and $\theta \in \mathbb{R}$, define

$$M(\beta,\theta) := \max_{0\le r\le 1} \left\{\theta r + \frac{\beta^2}{2} r(1-r)\right\}.$$

For a plane family $\mathcal{C} = \{(\beta_i,\theta_i,\gamma_i)\}_{i=1}^L$, $1 \le L \le m$, with $\beta_i > 0$ and $\gamma_i \ge 0$, let

$$Y_i(r,Z) := \exp\Big(-\beta_i\sqrt{r(1-r)}\,Z + \theta_i r - M(\beta_i,\theta_i) - \gamma_i\Big), \qquad Z \sim N(0,1),$$

and

$$\Lambda(\mathcal{C}) := \sup_{0\le r\le 1} \mathbb{E} \max_{1\le i\le L} Y_i(r,Z).$$

Finally define, for $d > 0$,

$$\Phi_d(\beta,\theta,\gamma) := \frac{d^2}{2v_0\beta} - d + \frac{M(\beta,\theta) + \gamma - \theta q_0}{\beta},$$

and

$$V := \inf_{\substack{1\le L\le m \\ \beta_i>0,\ \theta_i\in\mathbb{R},\ \gamma_i\ge 0 \\ \Lambda(\mathcal{C})=1}} \sum_{j=1}^m w_j \min_{1\le i\le L} \Phi_{d_j}(\beta_i,\theta_i,\gamma_i). \tag{35}$$

Theorem 19 (Near-diagonal Gaussian-envelope limit)**.** *Under the setup above,*

$$\boxed{\sqrt{N}\,\mathcal{G}_N^\star \longrightarrow V.}$$

Moreover,

$$0 \le V \le \sqrt{\sum_{j=1}^m w_j d_j^2} - \sum_{j=1}^m w_j d_j,$$

and

$$V = 0 \quad\Longleftrightarrow\quad d_1 = \cdots = d_m.$$

Hence, for a common local endpoint $b_j \equiv b$, the limit is strictly positive exactly when the local empirical coordinates a_j are not all equal.

Proof. We divide the proof into six steps.

Step 1: a uniform $O(N^{-1/2})$ benchmark. At target $(q_{j,N}, p_{j,N})$, let $\ell_{j,N}$ be the affine tangent of $g(q,p) = \mathrm{kl}(q,p)$. Its vertical slope equals

$$D_{j,N} = \frac{p_{j,N} - q_{j,N}}{p_{j,N}(1-p_{j,N})} = \frac{d_j}{v_0\sqrt{N}} + o(N^{-1/2}).$$

Each tangent has binomial normalizer one. Therefore $H_N = \max_j \ell_{j,N}$ has normalizer at most m and

$$\widehat{H}_N := H_N - \frac{1}{N}\log I_{H_N,N} \in \mathcal{F}_N.$$

At target j, the jth tangent alone excludes every

$$r > p_{j,N} + \frac{\log I_{H_N,N}}{N D_{j,N}}.$$

The KL ceiling makes $p_{j,N}$ feasible, so

$$0 \le U_{\widehat{H}_N}(q_{j,N}, t_{j,N}) - p_{j,N} \le \frac{\log m}{N D_{j,N}} = O(N^{-1/2}).$$

Consequently $\mathcal{G}_N^\star = O(N^{-1/2})$.

Step 2: support the comparator at its actual right endpoint. Choose $f_N \in \mathcal{F}_N$ within N^{-1} of the infimum. Put

$$e_{j,N} = U_{f_N}(q_{j,N}, t_{j,N}) - p_{j,N}, \qquad u_{j,N} = p_{j,N} + e_{j,N}.$$

Step 1 and the positive fixed weights imply $e_{j,N} = O(N^{-1/2})$ for every j. The KL ceiling gives $u_{j,N} \ge p_{j,N}$, and for large N all these points are interior.

Because f_N is finite and continuous on the open square, its vertical sublevel set at level $t_{j,N}$ is closed and convex. Since $u_{j,N}$ is its rightmost point and $u_{j,N} < 1$,

$$f_N(q_{j,N}, u_{j,N}) = t_{j,N}.$$

Take a supporting affine plane there,

$$L_{j,N}(q,r) = A_{j,N}q + D_{j,N}r + C_{j,N} \le f_N(q,r).$$

The envelope $\max_j L_{j,N}$ also lies below f_N, hence its binomial normalizer is at most one.

We now analyze one such plane and suppress the index j. Write

$$\ell_A(r) := \log(1 - r + re^A), \qquad h_N(r) := Dr + C + \ell_A(r).$$

Since $L \le f_N$ and f_N is normalized,

$$h_N(r) \le 0 \qquad (0 \le r \le 1).$$

At the contact (q_N, u_N),

$$t_N = Aq_N + Du_N + C = \underbrace{Aq_N - \ell_A(u_N)}_{=:C_A(q_N,u_N)} + h_N(u_N).$$

Thus $C_A(q_N, u_N) \ge t_N > 0$.

For $q_N < u_N$, positivity forces $A < 0$. Indeed $A \mapsto C_A(q_N, u_N)$ is strictly concave, vanishes at $A = 0$, and has derivative $q_N - u_N < 0$ there, while it is decreasing on $[0, \infty)$.

Also,

$$C_A(q_N, u_N) \le \mathrm{kl}(q_N, u_N)$$

by the Bernoulli log-moment variational formula. Since $u_N - q_N = \Theta(N^{-1/2})$ and Nt_N stays bounded above and away from zero, the following bounds hold:

$$0 < c \le -\sqrt{N}\,A \le C < \infty.$$

For completeness, Jensen gives $\ell_A(u_N) \geq u_N A$, hence $C_A(q_N, u_N) \leq (-A)(u_N - q_N)$ and therefore the lower bound on $-\sqrt{N}A$. Positivity of C_A first forces $A \to 0$. Indeed, since q_N and u_N remain uniformly interior, $Aq_N - \ell_A(u_N) \to -\infty$ if $A \to -\infty$. A bounded negative subsequence separated from zero has a further limit $A_* < 0$, where strict Jensen gives $A_* q_0 - \log(1 - q_0 + q_0 e^{A_*}) < 0$, another contradiction. Uniformly on an interior neighborhood,

$$\ell_A(u) \geq uA + c_0 A^2$$

for small A, so

$$0 < C_A(q_N, u_N) \leq (-A)(u_N - q_N) - c_0 A^2,$$

which gives the upper bound.

Moreover,

$$0 \leq C_A(q_N, u_N) - t_N = -h_N(u_N) \leq \mathrm{kl}(q_N, u_N) - t_N = O(N^{-1}).$$

Using $h_N(0) \leq 0$, $h_N(1) \leq 0$, and $h_N(u_N) \geq -C/N$ gives

$$\frac{-\ell_A(u_N) - C/N}{u_N} \leq D \leq \frac{\ell_A(u_N) - A + C/N}{1 - u_N}.$$

Since $\ell_A(u) = uA + O(A^2)$ uniformly,

$$D = -A + O(N^{-1}).$$

Define the scaled parameters

$$\beta_N := -\sqrt{N}\, A, \qquad \theta_N := N(A + D),$$

and

$$M_N := \sup_{0 \leq r \leq 1} \{Dr + \ell_A(r)\}, \qquad \gamma_N := -N(C + M_N).$$

Then β_N stays in a compact subset of $(0, \infty)$, θ_N is bounded, and $\gamma_N \geq 0$. Also

$$\gamma_N \leq -N h_N(u_N) = O(1),$$

because $Du_N + \ell_A(u_N) - M_N \leq 0$.

The Taylor expansion

$$\ell_A(r) = rA + \frac{1}{2} r(1 - r)A^2 + O(|A|^3)$$

is uniform in $r \in [0, 1]$. Therefore, along every convergent subsequence $\beta_N \to \beta > 0$, $\theta_N \to \theta$, $\gamma_N \to \gamma \geq 0$,

$$N M_N \to M(\beta, \theta) = \max_{0 \leq r \leq 1} \left\{ \theta r + \frac{\beta^2}{2} r(1 - r) \right\}.$$

If

$$q_N = q_0 + \frac{a + o(1)}{\sqrt{N}}, \qquad u_N = q_0 + \frac{c + o(1)}{\sqrt{N}}, \qquad N t_N \to \tau,$$

then the contact identity $L(q_N, u_N) = t_N$ yields

$$\tau = \beta(c - a) + \theta q_0 - M(\beta, \theta) - \gamma.$$

In our application $c = b_j + E_j$, where $E_j = \lim \sqrt{N}\, e_{j,N}$, and $\tau = d_j^2/(2v_0)$. Hence

$$E_j = \Phi_{d_j}(\beta, \theta, \gamma).$$

Lemma 20 (Convergence of finite-plane normalizers)**.** *Fix a finite number L of planes. Suppose $\beta_{i,N} \to \beta_i > 0$, $\theta_{i,N} \to \theta_i$, and $\gamma_{i,N} \to \gamma_i \geq 0$ for $1 \leq i \leq L$. Set*

$$A_{i,N} = -\beta_{i,N}/\sqrt{N}, \quad D_{i,N} = -A_{i,N} + \theta_{i,N}/N, \quad M_{i,N} = \sup_{0 \leq r \leq 1} \{D_{i,N} r + \log(1 - r + r e^{A_{i,N}})\},$$

and $L_{i,N}(q, r) = A_{i,N} q + D_{i,N} r - M_{i,N} - \gamma_{i,N}/N$. Then

$$\sup_{0 \leq r \leq 1} \mathbb{E}_{K \sim \mathrm{Bin}(N,r)} \max_{1 \leq i \leq L} e^{N L_{i,N}(K/N, r)} \longrightarrow \Lambda(\mathcal{C}),$$

where $\mathcal{C} = \{(\beta_i, \theta_i, \gamma_i)\}_{i=1}^L$.

Proof. The uniform Taylor expansion above gives $NM_{i,N} \to M(\beta_i, \theta_i)$ for every i. For a population parameter $r_N \to r$, let $K_N \sim \mathrm{Bin}(N, r_N)$ and

$$X_N := \frac{K_N - Nr_N}{\sqrt{N}}.$$

For the ith plane,

$$NL_{i,N}(K_N/N, r_N) = -\beta_{i,N} X_N + \theta_{i,N} r_N - NM_{i,N} - \gamma_{i,N}.$$

If $r \in (0,1)$, the triangular-array CLT gives

$$X_N \Rightarrow N(0, r(1-r)).$$

If $r \in \{0,1\}$, then $\mathbb{E}X_N^2 = r_N(1-r_N) \to 0$, so $X_N \to 0$ in L^2.

The exponential moments needed to pass expectations are uniform. Hoeffding's lemma gives, for every fixed s,

$$\sup_{N,r} \mathbb{E}e^{sX_N} \le e^{s^2/8}.$$

Hence $e^{B|X_N|}$ is uniformly integrable for every fixed B, and therefore

$$\mathbb{E}\max_i e^{NL_{i,N}(K_N/N, r_N)} \longrightarrow \mathbb{E}\max_i \exp\Big(-\beta_i \sqrt{r(1-r)}\, Z + \theta_i r - M(\beta_i, \theta_i) - \gamma_i\Big).$$

Taking a maximizing sequence r_N and then a convergent subsequence gives the limsup inequality for the supremum over r; evaluating at a fixed near-maximizer of the continuous limit gives the reverse inequality. Thus the finite-N envelope normalizer converges to $\Lambda(\mathcal{C})$. □

Step 4: the full-class liminf. Take a subsequence along which $\sqrt{N}\,\mathcal{G}_N^\star$ attains its liminf and use the near minimizers from Step 2. Passing to a further subsequence, all $E_{j,N} := \sqrt{N}\, e_{j,N}$ and all scaled support-plane parameters converge.

Let $\mathcal{C}$ be the resulting collection of at most m limit planes. Because $\max_j L_{j,N} \le f_N$, its finite-N normalizer is at most one. Lemma 20 therefore gives

$$\Lambda(\mathcal{C}) \le 1.$$

For the plane extracted from endpoint j, Step 2 gives

$$E_j = \Phi_{d_j}(\beta_j, \theta_j, \gamma_j).$$

If $\lambda := \Lambda(\mathcal{C}) < 1$, replace every γ_i by

$$\gamma_i' = \gamma_i + \log \lambda.$$

Since $\Lambda(\mathcal{C}) \ge \sup_r \mathbb{E}Y_i(r, Z) = e^{-\gamma_i}$, we have $\gamma_i' \ge 0$. The new plane family has normalizer exactly one, and every cost weakly decreases because $\log \lambda \le 0$. Consequently

$$V \le \sum_j w_j \min_i \Phi_{d_j}(\beta_i, \theta_i, \gamma_i') \le \sum_j w_j E_j.$$

This proves

$$\liminf_{N\to\infty} \sqrt{N}\, \mathcal{G}_N^\star \ge V.$$

Step 5: an exactly legal finite-N recovery sequence. Fix a plane family $\mathcal{C} = \{(\beta_i, \theta_i, \gamma_i)\}_{i=1}^L$ with $\Lambda(\mathcal{C}) = 1$. Define

$$A_{i,N} = -\frac{\beta_i}{\sqrt{N}}, \qquad D_{i,N} = -A_{i,N} + \frac{\theta_i}{N},$$

$$M_{i,N} = \sup_{0\le r\le 1} \left\{ D_{i,N} r + \log(1 - r + re^{A_{i,N}}) \right\}, \qquad C_{i,N} = -M_{i,N} - \frac{\gamma_i}{N},$$

and

$$H_N(q, r) = \max_i \{A_{i,N} q + D_{i,N} r + C_{i,N}\}.$$

By Lemma 20,

$$I_{H_N,N} \to \Lambda(\mathcal{C}) = 1.$$

Therefore

$$f_N := H_N - \frac{1}{N} \log I_{H_N,N}$$

is an exactly normalized comparator for every N, and $\log I_{H_N,N} = o(1)$.

For large N, every $D_{i,N} > 0$. Hence, at a fixed empirical coordinate $q_{j,N}$, the vertical slice $r \mapsto f_N(q_{j,N}, r)$ is strictly increasing. Let $r_{ji,N}$ be the unique root at which the normalized ith affine piece equals $t_{j,N}$. Writing

$$r_{ji,N} = p_{j,N} + \frac{E_{ji,N}}{\sqrt{N}}$$

and using $Nt_{j,N} \to d_j^2/(2v_0)$, $NM_{i,N} \to M(\beta_i, \theta_i)$, and $\log I_{H_N,N} \to 0$, gives

$$E_{ji,N} \longrightarrow \Phi_{d_j}(\beta_i, \theta_i, \gamma_i).$$

Because a maximum of increasing affine functions is below the threshold exactly when every component is below it,

$$U_{f_N}(q_{j,N}, t_{j,N}) = \min_i r_{ji,N}$$

for all sufficiently large N. Thus

$$\limsup_{N\to\infty} \sqrt{N}\, \mathcal{G}_N^\star \le \sum_j w_j \min_i \Phi_{d_j}(\beta_i, \theta_i, \gamma_i).$$

Taking the infimum over normalized plane families proves the limsup $\le V$. Together with Step 4 this proves the limit theorem.

Step 6: nonnegativity, a closed one-plane upper bound, and the zero criterion. Since q_0 is an admissible point in the definition of M,

$$M(\beta, \theta) \ge \theta q_0 + \frac{\beta^2 v_0}{2}.$$

Therefore

$$\Phi_d(\beta, \theta, \gamma) \ge \frac{d^2}{2v_0\beta} - d + \frac{\beta v_0}{2} + \frac{\gamma}{\beta} = \frac{(d - v_0\beta)^2}{2v_0\beta} + \frac{\gamma}{\beta} \ge 0.$$

For a single component choose

$$\theta = \beta^2(q_0 - \tfrac{1}{2}), \qquad \gamma = 0.$$

Then q_0 is the maximizer defining M, so the one-component Gaussian normalizer is exactly one and

$$\Phi_{d_j} = \frac{(d_j - v_0\beta)^2}{2v_0\beta}.$$

Averaging and minimizing over $\beta > 0$ gives

$$V \le \inf_{\beta>0} \left\{ \frac{\sum_j w_j d_j^2}{2v_0\beta} - \sum_j w_j d_j + \frac{v_0\beta}{2} \right\} = \sqrt{\sum_j w_j d_j^2} - \sum_j w_j d_j.$$

If all d_j equal d, choose $\beta = d/v_0$ in this one-component construction; then every cost is zero, so $V = 0$.

Conversely, suppose $V = 0$. Choose normalized plane families with objective tending to zero. Since all weights are positive and all costs are nonnegative, for each target j some serving component has cost tending to zero. The preceding lower bound then forces, for that component,

$$\beta \to d_j/v_0, \qquad \gamma \to 0,$$

and also

$$M(\beta, \theta) - \theta q_0 - \frac{\beta^2 v_0}{2} \to 0.$$

The maximized function $r \mapsto \theta r + \frac{1}{2}\beta^2 r(1-r)$ is a strictly concave quadratic. Because q_0 is interior, the last display forces

$$\theta \to \beta^2(q_0 - \tfrac{1}{2}).$$

If two targets have distinct d's, after passing to a subsequence they must therefore be served by two distinct components converging to two distinct zero-cost triples. At $r = q_0$, their Gaussian variables converge in L^1 to

$$X_d(Z) = \exp\left(-\frac{d}{\sqrt{v_0}}Z - \frac{d^2}{2v_0}\right),$$

each of mean one. For $d \neq d'$ these variables are not almost surely equal, so

$$\mathbb{E}\max\{X_d, X_{d'}\} = 1 + \frac{1}{2}\mathbb{E}|X_d - X_{d'}| > 1.$$

This contradicts the plane family constraint $\Lambda = 1$, which in particular requires the expectation of the full maximum at $r = q_0$ to be at most one. Hence all d_j must be equal. □

Corollary 21 (Explicit two-state/common-endpoint sandwich)**.** *Assume additionally that $b_j \equiv b$, and put $d_j = b - a_j$, $d_{\max} = \max_j d_j$, and*

$$P_j = N(a_j, v_0), \qquad \bar{P} = \sum_j w_j P_j, \qquad J_G = \sum_j w_j \,\mathrm{KL}(P_j \| \bar{P}).$$

Then

$$\boxed{\frac{w_{\min} v_0 J_G}{d_{\max} + \sqrt{2v_0 J_G}} \leq V \leq \sqrt{\sum_j w_j d_j^2} - \sum_j w_j d_j.}$$

In particular, for $m = 2$ and $a_1 \neq a_2$, both bounds are strictly positive.

Proof. Let $\lambda_{j,N} = \mathrm{Bin}(N, q_{j,N})$ and $\bar{\lambda}_N = \sum_j w_j \lambda_{j,N}$, and set

$$\mathfrak{J}_N = \frac{1}{N}\sum_j w_j \,\mathrm{KL}(\lambda_{j,N} \| \bar{\lambda}_N).$$

The hypothesis $b_j \equiv b$ does not require the finite-N endpoints $p_{j,N}$ to coincide. Introduce the auxiliary target family with the same $q_{j,N}, w_j$ and the exactly common endpoint $p_N^\circ = q_0 + b/\sqrt{N}$, and denote its optimum by $\mathcal{G}_N^{\circ\star}$. The already proved limit theorem applies to both families with identical local parameters, so $\sqrt{N}\mathcal{G}_N^{\circ\star} \to V$ as well. In the remainder of this proof write $p_N = p_N^\circ$; all finite-sample inequalities concern this auxiliary family.

The common-binomial generalized Jensen–Shannon inequality together with vertical endpoint interpolation gives, for every fixed $z > 0$ and $s_N = p_N + z/\sqrt{N}$,

$$\mathcal{G}_N^{\circ\star} \geq w_{\min}(s_N - p_N)\frac{\mathfrak{J}_N}{\mathfrak{J}_N + \max_j\{\mathrm{kl}(q_{j,N}, s_N) - \mathrm{kl}(q_{j,N}, p_N)\}}.$$

For completeness, put $d_j = \mathrm{kl}(q_{j,N}, p_N) - f(q_{j,N}, p_N) \geq 0$. With $\nu = \mathrm{Bin}(N, p_N)$, $a_k = f(k/N, p_N)$, and $\pi(k) = \nu(k)e^{Na_k}/\sum_l \nu(l)e^{Na_l}$, Jensen and normalization give $Nd_j \geq \mathrm{KL}(\lambda_{j,N}\|\pi)$. The KL mixture decomposition therefore yields $\sum_j w_j d_j \geq \mathfrak{J}_N$. Set

$$M_j = \mathrm{kl}(q_{j,N}, s_N) - \mathrm{kl}(q_{j,N}, p_N) > 0.$$

Convex interpolation with the KL ceiling at s_N makes $p_N + (s_N - p_N)d_j/(d_j + M_j)$ feasible at threshold $\mathrm{kl}(q_{j,N}, p_N)$. Some j has $d_j \geq \mathfrak{J}_N$; retaining this state proves the displayed inequality. If $\mathfrak{J}_N = 0$, the inequality follows directly from nonnegative endpoint excess.

We first identify the two limits in this expression. Under $\lambda_{j,N}$, let

$$X_N = \frac{K - Nq_0}{\sqrt{N}}.$$

Then $X_N \Rightarrow N(a_j, v_0)$. For every k, j the exact log-likelihood ratio is affine in X_N, and its coefficients satisfy

$$\log \frac{\lambda_{k,N}(K)}{\lambda_{j,N}(K)} = \alpha_{kj,N} X_N + \zeta_{kj,N}, \qquad \alpha_{kj,N} \longrightarrow \frac{a_k - a_j}{v_0}, \quad \zeta_{kj,N} \longrightarrow -\frac{a_k^2 - a_j^2}{2v_0}.$$

These are convergences of the two deterministic coefficients; an additive error uniform over all $K \in \{0, \dots, N\}$ is neither asserted nor needed. Hence

$$\mathrm{KL}(\lambda_{j,N} \| \bar{\lambda}_N) = -\mathbb{E}_{\lambda_{j,N}} \log \sum_k w_k \frac{\lambda_{k,N}(K)}{\lambda_{j,N}(K)} \longrightarrow \mathrm{KL}(P_j \| \bar{P}).$$

Uniform integrability follows because the log-sum-exp integrand grows at most linearly in $|X_N|$ and the local binomial variables are uniformly sub-Gaussian. Therefore

$$N \mathfrak{J}_N \to J_G.$$

The local KL expansion gives

$$N\{\mathrm{kl}(q_{j,N}, s_N) - \mathrm{kl}(q_{j,N}, p_N)\} \longrightarrow \frac{d_j z + z^2/2}{v_0}.$$

Thus

$$V \geq w_{\min} \frac{z J_G}{J_G + (d_{\max} z + z^2/2)/v_0}.$$

If $J_G = 0$, the claimed lower bound is zero and follows from $V \geq 0$. If $J_G > 0$, optimizing the right side over $z > 0$ gives $z = \sqrt{2 v_0 J_G}$ and therefore

$$V \geq \frac{w_{\min} v_0 J_G}{d_{\max} + \sqrt{2 v_0 J_G}}.$$

The upper bound is the one-plane bound from the theorem. □

Interpretation of the tangential parameter. Writing

$$\rho = \frac{1}{2} + \frac{\theta}{\beta^2},$$

one has

$$M(\beta, \theta) - \theta q_0 = \frac{\beta^2}{2} \left[v_0 + (q_0 - \rho)^2 - \mathrm{dist}(\rho, [0,1])^2 \right].$$

Thus ρ is the surviving location of the component's worst population parameter. It is precisely the degree of freedom that is lost by a purely local quadratic approximation at q_0.

D A Fixed-Task ERM Selection Rate

This section proves Theorem 1 and records a fair-label companion. For the latter, put

Theorem 22 (A fair-label fixed-task ERM rate)**.** *Let $Y \sim \mathrm{Ber}(1/2)$, let $h_0 \equiv 0$ and $h_1 \equiv 1$ be the two constant classifiers under zero–one loss, and fix $P(h_0) = \pi \in (0,1)$. For odd N, let Q_S be the unique ERM and let $\Gamma_N(\pi, \delta) = \inf_{f \in \mathcal{F}_N} \mathbb{E}_S[p_f(S) - p_{\tan}(S)]$. For every fixed $\delta \in (0,1)$,*

$$\pi \neq \tfrac{1}{2} \implies \Gamma_N(\pi, \delta) = \Theta_{\pi,\delta}(N^{-1/2}), \qquad \Gamma_N(\tfrac{1}{2}, \delta) = O_\delta(N^{-3/2}).$$

The first relation is along odd $N \to \infty$; the symmetric statement gives an achievable upper rate.

Put

$$A_0 = \log \frac{1}{\pi\delta}, \qquad A_1 = \log \frac{1}{(1-\pi)\delta}, \qquad d_j = \sqrt{A_j/2},$$

and let $\bar{A} = (A_0 + A_1)/2$, $\bar{d} = (d_0 + d_1)/2$, $D = \sqrt{\bar{A}/2}$. For a count $K = \sum_i Y_i$, write

$$J_N(K) = \mathbf{1}\{K > N/2\}, \qquad q_N(K) = \frac{\min(K, N-K)}{N},$$

$$p_N(K) = F\left(q_N(K), \frac{A_{J_N(K)}}{N}\right).$$

Thus the two reflected ERM branches have the same empirical risk and complexities $A_0/N, A_1/N$. The infimum defining $\Gamma_N(\pi, \delta)$ remains outside the expectation.

We repeatedly use two elementary consequences of the normalized-comparator contract. First, $f(q, r) \le \mathrm{kl}(q, r)$, including $q = 0, 1$, so every endpoint excess is nonnegative. Second, uniformly on compact $J \subset (0, 1)$,

$$\sqrt{N}\{F(q, A/N) - q\} = \sqrt{2Aq(1-q)} + O_{A,J}(N^{-1/2}), \tag{36}$$

while Pinsker's inequality gives the global bound

$$0 \le F(q, A/N) - q \le \sqrt{A/(2N)}. \tag{37}$$

Both statements also follow directly by integrating $\partial_r \mathrm{kl}(q, r) = (r - q)/(r(1 - r))$.

Lemma 23 (A globally legal affine construction)**.** *For every fixed $b > 0$,*

$$g_{N,b}(q, r) = \frac{b}{\sqrt{N}}(r - q) - \frac{b^2}{8N}$$

belongs to $\mathcal{F}_N$, and

$$U_{g_{N,b}}(q, A/N) = \min\left\{1, q + \frac{A/b + b/8}{\sqrt{N}}\right\}.$$

Its expected endpoint excess $G^{\mathrm{aff}}_{N,b}$ satisfies, for odd N,

$$0 \le \Gamma_N(\pi, \delta) \le G^{\mathrm{aff}}_{N,b} \le \frac{\bar{A}/b + b/8}{\sqrt{N}}, \tag{38}$$

$$\lim_{\substack{N\to\infty \\ N \text{ odd}}} \sqrt{N}\, G^{\mathrm{aff}}_{N,b} = \frac{\bar{A}}{b} + \frac{b}{8} - \bar{d}. \tag{39}$$

For $\pi = 1/2$, the choice $b = 4d$, $d = \sqrt{\log(2/\delta)/2}$, additionally gives

$$\Gamma_N(1/2, \delta) \le \frac{8d^3 + 16d}{3N^{3/2}} \quad \textit{for odd } N \ge 64d^2. \tag{40}$$

Proof. Hoeffding's centered Bernoulli moment bound gives, for every $r \in [0, 1]$,

$$\mathbb{E}_{K\sim\mathrm{Bin}(N,r)} e^{N g_{N,b}(K/N, r)} = e^{-b^2/8}\mathbb{E}e^{(b/\sqrt{N})(Nr-K)} \le 1.$$

Solving the affine sublevel inequality gives the stated endpoint. Reflection $K \leftrightarrow N - K$ makes each ERM branch have probability $1/2$, which proves (38).

Set $Q_N = q_N(K)$. Since $Q_N = 1/2 - |K/N - 1/2|$, we have $Q_N \to 1/2$ in probability. Equations (36) and (37) imply

$$\sqrt{N}\{F(Q_N, A_{J_N(K)}/N) - Q_N\} - d_{J_N(K)} \longrightarrow 0$$

in L^1. For all large N, the affine endpoint is unclipped at every ERM state, yielding (39).

It remains to show the sharper symmetric bound. Put $H = d/\sqrt{N}$, $h = F(q, 2d^2/N) - q$, and $x = 1/2 - q$. On $q \ge 1/4$, exact integration gives

$$2H^2 = 2h^2 + 16\int_0^h \frac{s(s - x)^2}{1 - 4(s - x)^2}\, ds.$$

When $H \le 1/8$, the denominator is at least $3/4$, hence

$$H - h \le \frac{8}{3}H^3 + \frac{16}{3}x^2 H.$$

For the random ERM state, $x = |K/N - 1/2|$, $\mathbb{E}x^2 = 1/(4N)$, and Chebyshev gives $\mathbb{P}(q < 1/4) \le 4/N$. On that complement use $0 \le H - h \le H$. Averaging proves (40). □

The next lemma supplies the rigidity argument for the full-class converse: it extracts from an arbitrary nearly sharp comparator two exponential tilts that must each spend essentially one unit of the same global moment budget.

Lemma 24 (Two sharp contacts have the same local gap)**.** *Fix $q_0 \in (0,1)$. Suppose $q_{j,N} \to q_0$, $0 < q_{j,N} < p_{j,N} < 1$, and*

$$\sqrt{N}(p_{j,N} - q_{j,N}) \to d_j > 0, \qquad t_{j,N} = \mathrm{kl}(q_{j,N}, p_{j,N}), \qquad j \in \{0,1\}.$$

If $f_N \in \mathcal{F}_N$ satisfies

$$\sqrt{N}\{U_{f_N}(q_{j,N}, t_{j,N}) - p_{j,N}\} \to 0 \quad (j = 0,1),$$

then $d_0 = d_1$.

Proof. Let $u_{j,N} = U_{f_N}(q_{j,N}, t_{j,N})$. These points are eventually interior, $f_N(q_{j,N}, u_{j,N}) = t_{j,N}$, and convexity supplies supports

$$L_{j,N}(x,r) = a_{j,N}x + b_{j,N}r + c_{j,N} \le f_N(x,r)$$

touching there. Their maximum remains below f_N.

Suppress j, N, and define

$$\ell_a(r) = \log(1 - r + re^a), \qquad h(r) = br + c + \ell_a(r).$$

The exact binomial moment of L equals $e^{Nh(r)}$, so global normalization gives $h(r) \le 0$ for every $r \in [0,1]$. At the contact, $t = aq - \ell_a(u) + h(u)$. With $z = ue^a/(1 - u + ue^a)$, direct expansion gives

$$\mathrm{kl}(q,z) = \mathrm{kl}(q,u) - \{aq - \ell_a(u)\}. \tag{41}$$

The local assumptions imply $N\{\mathrm{kl}(q,u) - t\} \to 0$. Since both terms into which (41) decomposes this deficit are nonnegative, Pinsker and the logit mean-value theorem give

$$\sqrt{N}\, a_{j,N} \longrightarrow -\frac{d_j}{q_0(1-q_0)}, \qquad -Nh_{j,N}(u_{j,N}) \longrightarrow 0. \tag{42}$$

The endpoint constraints $h(0), h(1) \le 0$ imply

$$\frac{-\ell_a(u) + h(u)}{u} \le b \le \frac{\ell_a(u) - a - h(u)}{1-u}.$$

Using $a = O(N^{-1/2})$, $\ell_a(r) = ra + O(a^2)$, and $\ell_a'(r) = a + O(a^2)$, uniformly in r, gives $b + a = O(N^{-1})$ and $\sup_r |h'(r)| = O(N^{-1})$. Consequently

$$Nh_{j,N}(q_0) \longrightarrow 0. \tag{43}$$

Let $\nu_N = \mathrm{Bin}(N, q_0)$, and set

$$m_{j,N} = e^{Nh_{j,N}(q_0)} \to 1, \qquad r_{j,N} = \frac{q_0 e^{a_{j,N}}}{1 - q_0 + q_0 e^{a_{j,N}}}, \qquad \lambda_{j,N} = \mathrm{Bin}(N, r_{j,N}).$$

Exact exponential tilting gives, for every count k,

$$\nu_N(k) e^{NL_{j,N}(k/N, q_0)} = m_{j,N}\lambda_{j,N}(k), \qquad \sqrt{N}(r_{j,N} - q_0) \to -d_j.$$

For binomial laws, their Hellinger affinity satisfies

$$\mathrm{Aff}(\lambda_{0,N}, \lambda_{1,N}) \le \exp\{-N(r_{0,N} - r_{1,N})^2/4\}.$$

Thus, if $d_0 \ne d_1$, their total variation distance has a positive liminf. On the other hand, the moment of the maximum of the two supports is at most one, whereas

$$1 \ge \sum_k \max_j \{m_{j,N}\lambda_{j,N}(k)\} \ge \min_j m_{j,N}\{1 + \mathrm{TV}(\lambda_{0,N}, \lambda_{1,N})\}.$$

Taking liminf is a contradiction. Hence $d_0 = d_1$. □

Proof of Theorem 22. Choosing $b = 4D = \sqrt{8\bar{A}}$ in Lemma 23 gives

$$\Gamma_N(\pi, \delta) \le D/\sqrt{N}, \qquad \limsup_{\substack{N\to\infty \\ N \text{ odd}}} \sqrt{N}\,\Gamma_N(\pi, \delta) \le D - \bar{d}.$$

The last constant is positive exactly when $d_0 \neq d_1$.

Suppose now that $\pi \neq 1/2$ but a subsequence of odd N satisfies $\sqrt{N}\Gamma_N(\pi, \delta) \to 0$. Choose deterministic $f_N \in \mathcal{F}_N$ whose full expected excess is at most $\Gamma_N + N^{-1} = o(N^{-1/2})$. Pair the reflected counts $l, N - l$ and put

$$\mu_N(l) = 2\binom{N}{l}2^{-N}, \qquad a_N(l) = \frac{1}{2}\sum_{j=0}^{1}\left\{U_{f_N}(l/N, A_j/N) - F(l/N, A_j/N)\right\}.$$

Then $\mathbb{E}e_{f_N,N} = \sum_{l=0}^{(N-1)/2} \mu_N(l)a_N(l)$. Chebyshev's inequality assigns mass at least $3/4$ to the paired central block $0 < 1/2 - l/N \le N^{-1/2}$. Hence some deterministic l_N in that block has $a_N(l_N) = o(N^{-1/2})$. Both nonnegative summands have the same order. Setting

$$q_{j,N} = l_N/N, \qquad p_{j,N} = F(l_N/N, A_j/N)$$

and using (36), Lemma 24 forces $d_0 = d_1$, equivalently $\pi = 1/2$, a contradiction. Therefore the nonnegative sequence $\sqrt{N}\Gamma_N$ has a positive liminf, proving the asymmetric $\Theta(N^{-1/2})$ claim. Finally, (40) proves the stated symmetric upper bound. □

The result concerns certificate width, not the learner's excess risk: both constant classifiers have population risk $1/2$. It does not identify the asymmetric leading constant, prove a matching symmetric lower bound, or give uniformity as $\pi \to 1/2$ or $\delta \to 0$.

For the low-risk theorem, put $\alpha_0 = \log(1/(\pi\delta))$ and $\alpha_1 = \log(1/((1-\pi)\delta))$; its d_j equal $\sqrt{2\rho(1-\rho)\alpha_j}$.

Lemma 25 (Uniformity under a compact local shift)**.** *Fix $q_0 \in (0, 1)$, positive weights $w_1, \dots, w_m$ summing to one, $c_j > 0$, and $A < \infty$. For $|a| \le A$ and sufficiently large N, let $q_N(a) = q_0 + a/\sqrt{N}$, $p_{j,N}(a) = F(q_N(a), c_j/N)$, and*

$$G_N(a) = \inf_{f\in\mathcal{F}_N} \sum_{j=1}^{m} w_j\{U_f(q_N(a), c_j/N) - p_{j,N}(a)\}.$$

Let V be the Gaussian-envelope value in (35) with $d_j = \sqrt{2q_0(1-q_0)c_j}$. Then $\sup_{|a|\le A}|\sqrt{N}G_N(a) - V| \to 0$. Moreover, for any fixed finite Gaussian plane family $\mathcal{C} = \{(\beta_i, \theta_i, \gamma_i)\}$ with $\Lambda(\mathcal{C}) = 1$, the exactly normalized recovery comparators from Step 5 of Theorem 19 satisfy, for every j,

$$\sup_{|a|\le A}\left|\sqrt{N}\{U_{f_N}(q_N(a), c_j/N) - p_{j,N}(a)\} - \min_i \Phi_{d_j}(\beta_i, \theta_i, \gamma_i)\right| \to 0.$$

Proof. The binary-KL inverse expansion is uniform for $q_N(a)$ in this compact interior neighborhood, so $\sqrt{N}\{p_{j,N}(a) - q_N(a)\} \to d_j$ uniformly in $|a| \le A$. If the first assertion failed, choose a violating sequence a_N and pass to a subsequence with $a_N \to a_*$. Theorem 19 applies to the resulting target sequence and gives the same value V, which does not depend on the common shift a_*. This is a contradiction. For the recovery assertion, Lemma 20 makes its normalizing correction $o(1/N)$. The affine-root expansion in Step 5 then gives the displayed limit along every convergent subsequence $a_N \to a_*$. The limit is independent of a_*; compactness gives uniform convergence. □

Proof of Theorem 1. Write N_a, N_b for the counts of the first two atoms, $R_N = \min(N_a, N_b)/N$, and J_N for the ERM branch. Off ties, swapping N_a and N_b preserves the sample probability and R_N and interchanges $J_N = 0$ and $J_N = 1$. Thus, for each grid point $q = l/N$, the two non-tie branch probabilities both equal $\omega_{l,N}/2$, where $\omega_{l,N} = \Pr(R_N = q, N_a \neq N_b)$. The tie probability

is $O(N^{-1/2})$: conditional on $N_a + N_b = m$, the first count is $\mathrm{Bin}(m, 1/2)$; its central atom is $O((m+1)^{-1/2})$, while $\Pr(N_a + N_b < \rho N) = O(N^{-1})$. Moreover,

$$\Pr\{\sqrt{N}|R_N - \rho| > M\} \le 2\rho(1-\rho)/M^2,$$

by the two binomial marginal variances and a union bound.

At grid point q put $p_{j,N}(q) = F(q, \alpha_j/N)$ and $e_{j,N}(q; f) = U_f(q, \alpha_j/N) - p_{j,N}(q) \ge 0$. If $\sqrt{N}(q_N - \rho)$ converges to a finite value, the local binary-KL inverse gives $\sqrt{N}(p_{j,N}(q_N) - q_N) \to d_j$. Theorem 19, applied to these two targets at the same q_N, therefore yields

$$\sqrt{N} \inf_{f \in \mathcal{F}_N} \tfrac{1}{2}\{e_{0,N}(q_N; f) + e_{1,N}(q_N; f)\} \to V_\rho.$$

Its value depends only on ρ, d_0, d_1, not on the limiting $\sqrt{N}(q_N - \rho)$. Lemma 25 makes this convergence uniform over grid points with $|\sqrt{N}(q - \rho)| \le M$ for each fixed M. For every single f the expected excess is at least

$$\sum_l \omega_{l,N} \inf_{g \in \mathcal{F}_N} \tfrac{1}{2}\{e_{0,N}(l/N; g) + e_{1,N}(l/N; g)\};$$

the tie contribution is nonnegative. Restricting the sum to the central band, taking the infimum over f, and then sending M to infinity proves the lower limit V_ρ.

For the reverse inequality fix a plane family in (35) with normalizer one and use the exactly normalized finite-N recovery comparator from Step 5 of Theorem 19. Its vertical affine slopes are positive for large N. The explicit affine roots, whose numerators are uniformly bounded in $q \in [0, 1]$ and denominators are of order $\sqrt{N}$, give $0 \le e_{j,N}(q; f_N) \le C/\sqrt{N}$ on the whole interval. On each fixed central band, Lemma 25 gives uniform convergence of $\sqrt{N} e_{j,N}(q; f_N)$ to $\min_i \Phi_{d_j}(\beta_i, \theta_i, \gamma_i)$. The tail bound above and the vanishing tie probability permit averaging the two equally weighted branches. The resulting upper limit is the plane family's objective; taking its infimum gives V_ρ. The positivity and one-plane bound follow from Step 6 of Theorem 19. □

E A POSITIVE GAP FOR EVERY FINITE SAMPLE SIZE

This appendix proves Theorem 26. For each finite N, it constructs a task and prior P_N, both fixed before sampling, on which every pre-data comparator has positive expected endpoint excess. A separate fixed binary classification task and fixed prior give ordinary ERM a positive cost for every odd $N \ge 3$. These positivity statements provide neither a sharp nor a uniform-in-N lower bound.

Theorem 26 (A positive gap at every finite sample size)**.** *For every integer $N \ge 1$, there is an explicit finite i.i.d. bounded-loss PAC–Bayes task, a pre-data prior P_N, and a permutation-invariant learner at $\delta = 1/2$ such that every pre-data $f \in \mathcal{F}_N$ satisfies*

$$\mathbb{E}[p_f(S)] - \mathbb{E}[p_{\tan}(S)] \ge \gamma_N > 0.$$

E.1 AN EXPLICIT SYMMETRIC I.I.D. TASK

Fix

$$q_1 = \frac{1}{4}, \qquad q_2 = \frac{1}{2}, \qquad p = \frac{99}{100}, \qquad \delta = \frac{1}{2}, \qquad s = \frac{199}{200}. \tag{44}$$

Let $\kappa_j = \mathrm{kl}(q_j, p)$ and $c_{j,N} = N\kappa_j - \log 2$. Draw $Z_1, \dots, Z_N \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(1/2)$. On $\mathcal{H} = \{h_0, h_1, h_2\}$, set before sampling

$$P_N(h_j) = 2e^{-N\kappa_j} \quad (j = 1, 2), \qquad P_N(h_0) = 1 - P_N(h_1) - P_N(h_2). \tag{45}$$

Use the bounded losses

$$\ell(h_j, z) = q_j \quad (j = 1, 2), \qquad \ell(h_0, z) = 0, \tag{46}$$

and let the deterministic posterior select h_1 when $\sum_i Z_i$ is even and h_2 when it is odd.

Lemma 27 (Legality and symmetry of the task)**.** *For every $N \geq 1$, (45) is a probability mass function with all masses positive, and $c_{j,N} > 0$. The posterior rule is measurable and permutation invariant, each coordinate is counterfactually essential, and the two selected states have probability $1/2$. In state j,*

$$\widehat{R}_S(Q_S) = q_j, \qquad \mathrm{KL}(Q_S \| P_N) = c_{j,N}. \tag{47}$$

Proof. Direct calculation gives

$$e^{-\kappa_1} = \left(\frac{99}{10\,546\,875}\right)^{1/4} < \frac{1}{16}, \qquad e^{-\kappa_2} = \frac{\sqrt{99}}{50} < \frac{1}{5}.$$

For the first inequality, $99 \cdot 16^4 = 6\,488\,064 < 10\,546\,875$; the second is immediate. Both bases lie in $(0, 1)$, so

$$P_N(h_1) + P_N(h_2) \leq 2e^{-\kappa_1} + 2e^{-\kappa_2} < \frac{1}{8} + \frac{2}{5} = \frac{21}{40} < 1.$$

Thus $P_N(h_0) > 19/40$, and $P_N(h_j) < 1$ implies $c_{j,N} > 0$.

All spaces are finite. The rule depends only on the number of ones, so it is measurable and permutation invariant; flipping any coordinate toggles the parity and hence the selected hypothesis. If $K \sim \mathrm{Bin}(N, 1/2)$, then

$$\mathbb{P}(K \text{ even}) - \mathbb{P}(K \text{ odd}) = \mathbb{E}[(-1)^K] = (1 - 2 \cdot \tfrac{1}{2})^N = 0.$$

The probabilities are therefore both $1/2$. Constant losses give the empirical risk in (47), and a point-mass posterior gives $\mathrm{KL}(\delta_{h_j} \| P_N) = -\log P_N(h_j) = c_{j,N}$. □

E.2 Proof of the persistence theorem

For this construction define

$$\lambda_{1,N} = \mathrm{Bin}(N, 1/4), \qquad \lambda_{2,N} = \mathrm{Bin}(N, 1/2), \qquad \bar{\lambda}_N = \frac{\lambda_{1,N} + \lambda_{2,N}}{2}, \tag{48}$$

$$J_N = \frac{1}{2N}\left\{\mathrm{KL}(\lambda_{1,N} \| \bar{\lambda}_N) + \mathrm{KL}(\lambda_{2,N} \| \bar{\lambda}_N)\right\} > 0, \tag{49}$$

and

$$M = \frac{1}{4}\log\frac{198}{199} + \frac{3}{4}\log 2 > 0. \tag{50}$$

Proof of the all-N positive-gap lower bound. Lemma 27 verifies the task. In state j,

$$\frac{c_{j,N} + \log(1/\delta)}{N} = \frac{N\kappa_j - \log 2 + \log 2}{N} = \mathrm{kl}(q_j, p). \tag{51}$$

Because $q_j < p$, the rightmost binary-KL root at this level is exactly p. Hence the separately optimized benchmark $p_{\tan}(S)$ equals p for every sample.

Now fix a pre-data comparator with $I_{\Delta,N} < \infty$, normalize it by (24), and use (26). Its endpoint in state j is

$$B[f(q_j, \cdot), \mathrm{kl}(q_j, p)].$$

Apply Lemmas 16 and 18 with equal weights and the constants in (44). Since

$$w_{\min}(s - p) = \frac{1}{2}\left(\frac{199}{200} - \frac{99}{100}\right) = \frac{1}{400},$$

we obtain

$$\mathbb{E}_S[p_\Delta(S)] - \mathbb{E}_S[p_{\tan}(S)] \geq \frac{J_N}{400(J_N + M)} > 0.$$

Finally,

$$\mathrm{kl}(q, s) - \mathrm{kl}(q, p) = q\log\frac{198}{199} + (1 - q)\log 2$$

is strictly decreasing in q. The maximum over $q_1 = 1/4$ and $q_2 = 1/2$ is therefore the value at q_1, which is exactly (50).

Under the stated endpoint convention $B[g, +\infty] = 1$, an infinite-normalizer comparator has excess endpoint $1 - p = 1/100$. This is larger than the displayed lower bound because $J_N/(J_N + M) < 1$ and $1/400 < 1/100$. Foong et al.'s source definition uses finite thresholds; our source-level theorem therefore quantifies over finite-normalizer comparators; the extended-value convention is stated only to make the appendix's endpoint semantics complete. □

E.3 A STRICTLY SUBUNIT WITNESS AT STANDARD CONFIDENCE

The preceding target $p = 99/100$ makes the prior legal down to one sample. Here we keep the target well inside the loss range and use $\delta = 1/20$. Fix

$$q_1 = \frac{1}{20}, \qquad q_2 = \frac{1}{10}, \qquad p = \frac{3}{10}, \qquad s = \frac{1}{2}, \tag{52}$$

and write $\kappa_j = \mathrm{kl}(q_j, p)$ and $a_j = e^{-\kappa_j}$. For every $N \geq 27$, use the same fair-bit parity rule and constant losses as above, but set

$$P_N(h_j) = 20e^{-N\kappa_j} \quad (j = 1, 2), \qquad P_N(h_0) = 1 - P_N(h_1) - P_N(h_2). \tag{53}$$

The prior is legal by exact rational power comparisons. Indeed,

$$a_1^{20} = 6\left(\frac{14}{19}\right)^{19}, \qquad a_2^{10} = 3\left(\frac{7}{9}\right)^{9}, \tag{54}$$

and direct integer cross-multiplication gives

$$a_1 < \frac{819}{1000}, \qquad a_2 < \frac{891}{1000}, \qquad 20\left\{\left(\frac{819}{1000}\right)^{27} + \left(\frac{891}{1000}\right)^{27}\right\} < 1. \tag{55}$$

Both bases lie in $(0, 1)$, so for every $N \geq 27$ the two displayed prior masses sum to less than one. In particular, each selected point-mass posterior has complexity

$$\mathrm{KL}(\delta_{h_j} \| P_N) = N\kappa_j - \log 20. \tag{56}$$

Adding $\log(1/\delta) = \log 20$ makes its threshold exactly κ_j. Thus the statewise binary-KL endpoint is $p = 3/10$ in both equally likely states, while the two predictors have population risks q_1 and q_2.

For the explicit gap, let

$$\lambda_{j,N}^{\mathrm{sub}} = \mathrm{Bin}(N, q_j), \qquad \bar{\lambda}_N^{\mathrm{sub}} = \frac{\lambda_{1,N}^{\mathrm{sub}} + \lambda_{2,N}^{\mathrm{sub}}}{2},$$

$$J_N^{\mathrm{sub}} = \frac{1}{2N}\sum_{j=1}^{2} \mathrm{KL}(\lambda_{j,N}^{\mathrm{sub}} \| \bar{\lambda}_N^{\mathrm{sub}}), \qquad M^{\mathrm{sub}} = \max_{j=1,2}\{\mathrm{kl}(q_j, 1/2) - \mathrm{kl}(q_j, 3/10)\}. \tag{57}$$

The binomial laws are distinct, hence $J_N^{\mathrm{sub}} > 0$, while $M^{\mathrm{sub}} > 0$. Lemma 18 with $w_{\min} = 1/2$ and $s - p = 1/5$ gives, for every normalized comparator,

$$\mathbb{E}[p_f(S)] - \mathbb{E}[p_{\tan}(S)] \geq \frac{1}{10}\frac{J_N^{\mathrm{sub}}}{J_N^{\mathrm{sub}} + M^{\mathrm{sub}}} := \gamma_N^{\mathrm{sub}} > 0. \tag{58}$$

This proves the stated nonvacuous witness. We claim only the sufficient range $N \geq 27$; no minimality of the threshold is needed.

E.4 ONE FIXED CLASSIFICATION TASK AND ORDINARY ERM

The preceding quantitative witnesses isolate the desired target pairs but use constant fractional losses and an N-dependent prior. The following result shows that strict incompatibility is also present in an ordinary fixed learning problem.

Theorem 28 (Finite-sample positivity for a fixed ERM task)**.** *Fix $\rho \in (0, 1)$ and $\delta \in (0, 1)$. Let $Y \sim \mathrm{Ber}(\rho)$, let $h_0 \equiv 0$ and $h_1 \equiv 1$ be the two constant binary classifiers under zero–one loss, and fix $P(h_0) = 1/3$, $P(h_1) = 2/3$. For every odd $N \geq 3$, let Q_S concentrate on the unique empirical-risk minimizer. Then there is $\gamma_{N,\rho,\delta} > 0$ such that every $f \in \mathcal{F}_N$ satisfies*

$$\mathbb{E}_S[p_f(S)] - \mathbb{E}_S[p_{\tan}(S)] \geq \gamma_{N,\rho,\delta} > 0. \tag{59}$$

Proof. Write $K = \sum_i Y_i$ and $q_N = (N-1)/(2N)$. The events

$$E_0 = \{K = (N-1)/2\}, \qquad E_1 = \{K = (N+1)/2\}$$

have strictly positive probabilities a_0, a_1. On E_j, ERM selects h_j, has empirical risk q_N, and has complexity

$$C_0 = \log 3, \qquad C_1 = \log(3/2).$$

Set $t_j = (C_j + \log(1/\delta))/N$ and $p_j = F(q_N, t_j)$. Since $0 < q_N < 1$, both targets are interior and

$$q_N < p_1 < p_0 < 1.$$

We verify that these two positive-weight targets are incompatible. Let ℓ_j be the KL tangent at (q_N, p_j), with vertical slope $D_j = (p_j - q_N)/[p_j(1-p_j)] > 0$, and set $H = \max\{\ell_0, \ell_1\}$. The tangent identity $\ell_j(1, p_j) = \log(q_N/p_j)$ gives

$$\ell_1(1, p_0) = \log(q_N/p_1) + D_1(p_0 - p_1) > \log(q_N/p_0) = \ell_0(1, p_0).$$

At ordinate p_0, every term in the binomial moment of H is at least the corresponding term for ℓ_0, and the $k = N$ term is strictly larger. Because the full moment of ℓ_0 at p_0 equals one, $I_{H,N} > 1$. Theorem 37, applied with weights $w_j = a_j/(a_0 + a_1)$, therefore yields a strictly positive two-state infimum $G_N^* > 0$.

For every other sample state the KL ceiling gives $p_f(S) \geq p_{\tan}(S)$. Consequently

$$\mathbb{E}_S[p_f(S) - p_{\tan}(S)] \geq (a_0 + a_1) G_N^*.$$

Taking $\gamma_{N,\rho,\delta} = (a_0 + a_1) G_N^* > 0$ proves the claim. □

The theorem is qualitative. The probability of the two near-tie events may decay with N when $\rho \neq 1/2$, and no uniform asymptotic lower rate is claimed. Its role is to show that the obstruction is not confined to parity selection, fractional constant losses, or sample-size-dependent priors.

The explicit bounds in the preceding constructions establish strict positivity rather than optimal finite-sample constants. Their parity constructions isolate the quantifier-level obstruction; Theorem 28 separately removes parity selection, fractional constant losses, and an N-dependent prior.

F Positive Recovery and Exact Comparator Compatibility

This appendix asks when the obstruction proved earlier becomes small, and when it disappears exactly. Throughout, the sample size N is fixed and finite, the state set is finite, all pre-data weights are strictly positive, and every target is interior: $0 < q_j < p_j < 1$. These restrictions are part of the results. In particular, the local estimates below are not uniform as $N \to \infty$, as a weight vanishes, or as a target approaches the boundary.

Use the normalized class, moment, and endpoint conventions from Appendix B. The KL ceiling makes each target feasible at level $t_j = \mathrm{kl}(q_j, p_j)$, so all endpoint excesses are nonnegative. The objective is $\mathcal{G}_N^\star(\mathsf{S})$ from (4), with one comparator chosen before the realized state is known.

F.1 A single mean-Catoni comparator gives global recovery

First suppose that all states have the same target endpoint p and that $0 < q_j < p < 1$. Set

$$\bar{q} = \sum_{j=1}^m w_j q_j, \qquad J_{\mathrm{Ber}} := J_1 = \sum_{j=1}^m w_j \,\mathrm{kl}(q_j, \bar{q}). \tag{60}$$

Because $\bar{q} < p$, define

$$A = \log \frac{\bar{q}(1-p)}{p(1-\bar{q})} < 0, \qquad C_A(q, r) = Aq - \log(1 - r + re^A), \qquad K_{\bar{q},p} = \frac{p(1-p)}{p - \bar{q}}. \tag{61}$$

Lemma 29 (Exact off-tangent Catoni identity). *For every finite N and every $r \in [0,1]$,*

$$G_{C_A,N}(r) = 1. \tag{62}$$

For every $q \in (0,1)$,

$$\mathrm{kl}(q,p) - C_A(q,p) = \mathrm{kl}(q,\bar{q}), \tag{63}$$

and the corresponding endpoint excess is exactly

$$B[C_A(q,\cdot), \mathrm{kl}(q,p)] - p = \min\Big\{1-p,\ K_{\bar{q},p}\big(1 - e^{-\mathrm{kl}(q,\bar{q})}\big)\Big\}. \tag{64}$$

Proof. For $K \sim \mathrm{Bin}(N,r)$, the binomial theorem gives

$$G_{C_A,N}(r) = e^{-N\log(1-r+re^A)}\mathbb{E}e^{AK} = 1.$$

Substituting the value of A and cancelling the two Bernoulli log-likelihood ratios proves (63). Since

$$\partial_r C_A(q,r) = \frac{1-e^A}{1-r+re^A} > 0,$$

the right endpoint is obtained by solving one strictly increasing scalar equation. If $a = \mathrm{kl}(q,\bar{q})$, then

$$C_A(q,r) = C_A(q,p) + a \quad\Longleftrightarrow\quad r - p = \frac{1-p+pe^A}{1-e^A}(1-e^{-a}) = K_{\bar{q},p}(1-e^{-a}).$$

Clipping this raw root at one yields (64). □

Theorem 30 (Mean-Catoni recovery). *For every fixed finite N,*

$$\boxed{\mathcal{G}_N^\star \le K_{\bar{q},p}\sum_{j=1}^m w_j\big(1-e^{-\mathrm{kl}(q_j,\bar{q})}\big) \le K_{\bar{q},p}J_1.} \tag{65}$$

Proof. Choose the single comparator C_A in (61) before the state is observed. It belongs to $\mathcal{F}_N$ by (62). Apply (64) state by state, remove clipping only in the upper-bound direction, average using the pre-data weights, and use $1 - e^{-x} \le x$ for $x \ge 0$. □

Thus exact universal sharpness can fail while a task-aware legal comparator still pays only for the dispersion of the possible empirical-risk states. The next result shows that this upper bound has the correct local order for every fixed N.

F.2 Matching local recovery for fixed finite N

Fix $0 < q_0 < p < 1$, fixed real directions $x_1, \dots, x_m$, and the same positive weights. For sufficiently small $|\varepsilon|$, put

$$q_j(\varepsilon) = q_0 + \varepsilon x_j, \qquad \bar{x} = \sum_j w_j x_j, \qquad \sigma_x^2 = \sum_j w_j(x_j - \bar{x})^2, \tag{66}$$

and assume $\sigma_x^2 > 0$. Then $\bar{q}(\varepsilon) = q_0 + \varepsilon\bar{x}$. Let

$$\lambda_{j,N,\varepsilon} = \mathrm{Bin}(N, q_j(\varepsilon)), \qquad \bar{\lambda}_{N,\varepsilon} = \sum_j w_j \lambda_{j,N,\varepsilon}, \tag{67}$$

and define

$$J_N(\varepsilon) = \frac{1}{N}\sum_j w_j \,\mathrm{KL}(\lambda_{j,N,\varepsilon} \| \bar{\lambda}_{N,\varepsilon}). \tag{68}$$

Lemma 31 (Local Bernoulli and binomial information). *As $\varepsilon \to 0$,*

$$J_1(\varepsilon) = \sum_j w_j \,\mathrm{kl}(q_j(\varepsilon), \bar{q}(\varepsilon)) = \frac{\sigma_x^2}{2q_0(1-q_0)}\varepsilon^2 + O(|\varepsilon|^3). \tag{69}$$

Moreover, for each fixed finite N,

$$J_N(\varepsilon) = J_1(\varepsilon) - \frac{1}{N}\,\mathrm{KL}\big(\bar{\lambda}_{N,\varepsilon} \big\| \mathrm{Bin}(N, \bar{q}(\varepsilon))\big)\,, \tag{70}$$

$$= J_1(\varepsilon) + O(\varepsilon^4), \qquad \frac{J_N(\varepsilon)}{J_1(\varepsilon)} \longrightarrow 1. \tag{71}$$

Proof. Taylor expansion of binary KL on a compact interior neighborhood of (q_0, q_0) gives

$$\mathrm{kl}(u, v) = \frac{(u-v)^2}{2q_0(1-q_0)} + O\big(\|(u-q_0, v-q_0)\|^3\big).$$

Weighted centering then proves (69). The exact KL chain-rule identity (32), applied to the present family, is (70).

It remains to control its correction. Regard the two laws in that correction as vectors on the fixed support $\{0, \dots, N\}$. They agree at $\varepsilon = 0$. Their first derivatives also agree, because differentiating a binomial mass with respect to its parameter makes the first-order mixture coefficient $\sum_j w_j x_j = \bar{x}$. Their coordinatewise difference is therefore $O(\varepsilon^2)$. For fixed N, every coordinate of the reference binomial law remains bounded away from zero when ε is small. Hence

$$\mathrm{KL}(P_\varepsilon \| Q_\varepsilon) \le \chi^2(P_\varepsilon \| Q_\varepsilon) = O(\varepsilon^4).$$

Combining this estimate with (69) proves (71). □

Theorem 32 (Quadratic recovery at fixed finite N)**.** *For every fixed finite N and every nondegenerate family (66),*

$$\boxed{\mathcal{G}_N^\star(\varepsilon) = \Theta(J_1(\varepsilon)) = \Theta(\varepsilon^2) \qquad (\varepsilon \to 0).} \tag{72}$$

The constants may depend on $N, q_0, p, w,$ and x.

Proof. Theorem 30 gives the upper bound. Its prefactor stays bounded because $\bar{q}(\varepsilon) \to q_0 < p$.

For the lower bound, fix any $s \in (p, 1)$ and set

$$M_\varepsilon = \max_j \{\mathrm{kl}(q_j(\varepsilon), s) - \mathrm{kl}(q_j(\varepsilon), p)\}.$$

The generalized-JS deficit and endpoint interpolation from Lemmas 16 and 18 give, for every $f \in \mathcal{F}_N$,

$$\sum_j w_j \{B[f(q_j(\varepsilon), \cdot), \mathrm{kl}(q_j(\varepsilon), p)] - p\} \ge w_{\min}(s-p)\frac{J_N(\varepsilon)}{J_N(\varepsilon) + M_\varepsilon}. \tag{73}$$

Here M_ε converges to the finite positive number $\mathrm{kl}(q_0, s) - \mathrm{kl}(q_0, p)$. Lemma 31 gives $J_N(\varepsilon) \sim J_1(\varepsilon) \asymp \varepsilon^2$. Thus the right-hand side of (73) is $\Omega(J_1(\varepsilon))$. Taking the infimum over f proves (72). □

If $\sigma_x^2 = 0$, all positive-weight directions coincide, $J_1 = 0$, and the positive quadratic conclusion is false. If N is allowed to vary with ε, the lower bound on the binomial coordinates used above can deteriorate. The theorem therefore makes no claim about a joint $N \to \infty$, $\varepsilon \to 0$ limit.

F.3 The tangent envelope decides exact compatibility

We now return to possibly distinct interior targets $0 < q_j < p_j < 1$. Let $t_j = \mathrm{kl}(q_j, p_j)$ and define the joint KL tangent at (q_j, p_j) by

$$\ell_j(q, r) = t_j + A_j(q - q_j) + D_j(r - p_j), \tag{74}$$

$$A_j = \log\frac{q_j(1-p_j)}{p_j(1-q_j)}, \qquad D_j = \frac{p_j - q_j}{p_j(1-p_j)} > 0. \tag{75}$$

Set

$$H = \max_{1 \le j \le m} \ell_j, \qquad I_{H,N} = \sup_{r \in [0,1]} G_{H,N}(r), \qquad \kappa_N = \frac{1}{N}\log I_{H,N}. \tag{76}$$

Lemma 33 (Moment bounds for the tangent envelope)**.** *Each tangent has normalizer one, and the envelope satisfies*

$$\sup_{r \in [0,1]} G_{\ell_j, N}(r) = 1, \qquad 1 \le I_{H,N} \le m, \qquad 0 \le \kappa_N \le \frac{\log m}{N}. \tag{77}$$

Proof. Write $\ell_j(q,r) = A_j q + D_j r + C_j$. The binomial theorem yields

$$G_{\ell_j,N}(r) = \left[e^{D_j r + C_j}(1 - r + r e^{A_j})\right]^N.$$

The logarithm of the bracket is strictly concave in r, has derivative zero at p_j, and equals zero there. Its supremum is therefore zero. Since $H \geq \ell_j$, its normalizer is at least one. Conversely, $e^{N \max_j \ell_j} \leq \sum_j e^{N\ell_j}$, so its normalizer is at most m. Taking logarithms gives the final bound. □

Theorem 34 (Exact tangent-envelope criterion)**.** *There exists one $f \in \mathcal{F}_N$ satisfying*

$$B[f(q_j,\cdot), t_j] = p_j \qquad \textit{for every } j \tag{78}$$

if and only if

$$\boxed{I_{H,N} = 1.} \tag{79}$$

Proof. Suppose first that an exact comparator f exists. The KL ceiling makes p_j feasible and gives $f(q_j,p_j) \leq t_j$. Since f is finite and continuous near the interior point (q_j,p_j), a strict inequality would make some $r > p_j$ feasible, contradicting (78). Thus $f(q_j,p_j) = \mathrm{kl}(q_j,p_j)$.

Take a supporting plane to f at this contact point. It lies below $f \leq \mathrm{kl}$ and touches the differentiable convex function kl at (q_j,p_j), so it must be the unique KL tangent ℓ_j. Hence $H \leq f$, and therefore

$$1 \leq I_{H,N} \leq \sup_r G_{f,N}(r) \leq 1.$$

This proves necessity.

Conversely, if $I_{H,N} = 1$, then $H \in \mathcal{F}_N$. At every target, $H(q_j,p_j) \geq \ell_j(q_j,p_j) = t_j$, whereas the KL ceiling gives the reverse inequality. Thus p_j is feasible. For every $r > p_j$,

$$H(q_j,r) \geq \ell_j(q_j,r) = t_j + D_j(r - p_j) > t_j.$$

No point to the right is feasible, so the endpoint is exactly p_j. □

Corollary 35 (A canonical approximate comparator)**.** *For arbitrary interior targets, the shifted envelope*

$$\widehat{H}_N = H - \kappa_N \tag{80}$$

belongs to $\mathcal{F}_N$ and satisfies, state by state,

$$0 \leq B[\widehat{H}_N(q_j,\cdot), t_j] - p_j \leq \min\left\{1 - p_j, \frac{\kappa_N}{D_j}\right\}. \tag{81}$$

In particular, $\kappa_N \leq (\log m)/N$ is an explicit coding-style upper penalty, not a claim of optimality.

Proof. The shift multiplies the normalizer by $e^{-N\kappa_N}$, so $\widehat{H}_N \in \mathcal{F}_N$. Its KL ceiling makes p_j feasible. On the other hand,

$$\widehat{H}_N(q_j,r) \geq t_j + D_j(r - p_j) - \kappa_N.$$

Every $r > p_j + \kappa_N/D_j$ is infeasible. Clipping at one proves (81). □

Corollary 36 (Common-endpoint compatibility)**.** *If all positive-weight states share one endpoint p, exact compatibility holds if and only if all corresponding q_j are equal.*

Proof. If all q_j are equal, all tangents coincide, and Theorem 34 applies. Conversely, choose $q_{\min} < q_{\max}$. Evaluate the envelope moment at $r = p$. The $k = N$ term of the q_j tangent satisfies

$$p^N \exp\{N\ell_j(1,p)\} = q_j^N.$$

Thus the maximum envelope contributes at least $q_{\max}^N$ in place of the $q_{\min}^N$ contribution. Every other term remains at least that of the $q_{\min}$ tangent. Since the complete $q_{\min}$ tangent moment at p equals one,

$$I_{H,N} \geq G_{H,N}(p) \geq 1 + q_{\max}^N - q_{\min}^N > 1.$$

Exact compatibility is therefore impossible by Theorem 34. □

F.4 Zero infimum without assuming attainment

The preceding existence theorem does not by itself exclude a sequence of inexact comparators whose objective tends to zero. The following result closes precisely that gap.

Theorem 37 (Zero infimum is exact compatibility)**.** *Under the fixed finite, positive-weight, interior assumptions of this appendix,*

$$\boxed{\mathcal{G}_N^\star = 0 \quad \Longleftrightarrow \quad I_{H,N} = 1 \quad \Longleftrightarrow \quad \textit{one exact normalized comparator exists.}} \tag{82}$$

No minimizer of (4) *is assumed to exist.*

Proof. Theorem 34 proves the second equivalence, and an exact comparator plainly gives $\mathcal{G}_N^\star = 0$. It remains to show that a zero infimum forces $I_{H,N} = 1$.

Assume $\mathcal{G}_N^\star = 0$ and choose a near-minimizing sequence $f_n \in \mathcal{F}_N$ with

$$E_n := \sum_j w_j \{B[f_n(q_j, \cdot), t_j] - p_j\} \longrightarrow 0.$$

Define its endpoint and contact deficits by

$$e_{j,n} = B[f_n(q_j, \cdot), t_j] - p_j \geq 0, \qquad d_{j,n} = t_j - f_n(q_j, p_j) \geq 0.$$

Both quantities are finite: $e_{j,n} \leq 1 - p_j$, while normalization at the interior ordinate p_j makes all binomial-grid values finite and horizontal convexity then makes $f_n(q_j, p_j)$ finite. Finiteness of the state set and $w_{\min} = \min_j w_j > 0$ imply

$$\max_j e_{j,n} \leq E_n / w_{\min} \longrightarrow 0. \tag{83}$$

For each j, fix $s_j \in (p_j, 1)$ and set

$$h_j = s_j - p_j > 0, \qquad M_j = \mathrm{kl}(q_j, s_j) - \mathrm{kl}(q_j, p_j) > 0.$$

The same convex interpolation used in Lemma 18, now state by state, gives

$$e_{j,n} \geq h_j \frac{d_{j,n}}{d_{j,n} + M_j}. \tag{84}$$

Let $h_{\min} = \min_j h_j$ and $M_{\max} = \max_j M_j$. For all sufficiently large n, (83) makes every $e_{j,n} < h_{\min}/2$. Rearranging (84) yields

$$d_{j,n} \leq \frac{e_{j,n} M_j}{h_j - e_{j,n}} \leq \frac{2M_{\max}}{h_{\min} w_{\min}} E_n, \qquad \max_j d_{j,n} \longrightarrow 0. \tag{85}$$

Write $x_j = (q_j, p_j)$ and $K(q, r) = \mathrm{kl}(q, r)$. Normalization makes each f_n finite on the open square: for every interior r, all binomial-grid values are finite, and horizontal convexity fills the intervals between them. Hence f_n is continuous and has a nonempty subdifferential at each x_j. Choose any $g_{j,n} \in \partial f_n(x_j)$. For either coordinate unit vector u, choose a fixed $\tau > 0$ for which $x_j \pm \tau u \in (0, 1)^2$. The supporting inequality, the KL ceiling, and $f_n(x_j) = K(x_j) - d_{j,n}$ give

$$\frac{K(x_j) - K(x_j - \tau u) - d_{j,n}}{\tau} \leq g_{j,n} \cdot u \leq \frac{K(x_j + \tau u) - K(x_j) + d_{j,n}}{\tau}. \tag{86}$$

Given any tolerance, differentiability of K lets us first choose τ so that the two secant slopes approximate $\nabla K(x_j) \cdot u$, and then (85) makes $d_{j,n}/\tau$ smaller than the remaining tolerance. Thus every such choice satisfies

$$g_{j,n} \longrightarrow \nabla K(x_j) = (A_j, D_j). \tag{87}$$

No subsequence or compactness extraction is used.

Let

$$L_{j,n}(x) = f_n(x_j) + g_{j,n} \cdot (x - x_j).$$

These supporting affine functions satisfy $L_{j,n} \leq f_n$. Equations (85) and (87) imply

$$\sup_{x \in [0,1]^2} |L_{j,n}(x) - \ell_j(x)| \leq d_{j,n} + \mathrm{diam}([0, 1]^2) \|g_{j,n} - \nabla K(x_j)\| \longrightarrow 0. \tag{88}$$

Since there are finitely many states,

$$H_n := \max_j L_{j,n} \longrightarrow H \quad \text{uniformly on } [0,1]^2, \qquad H_n \le f_n. \tag{89}$$

Fix $r \in [0,1]$. Normalization and (89) give

$$G_{H_n,N}(r) \le G_{f_n,N}(r) \le 1.$$

For fixed finite N, this binomial expectation contains only $N+1$ terms. Uniform convergence therefore permits termwise passage to the limit, including the zero-mass endpoint convention, and gives

$$G_{H,N}(r) = \lim_{n\to\infty} G_{H_n,N}(r) \le 1.$$

This holds for every fixed r, so taking the supremum only afterward gives $I_{H,N} \le 1$. Lemma 33 gives the reverse inequality. Hence $I_{H,N} = 1$, completing the proof. Notice that neither an infimum-attaining comparator nor an exchange of supremum and limit has entered the argument. □

The same equivalence applies to the original comparator parameterization. If $0 < I_{\Delta,N} < \infty$, the shift

$$f = \Delta - \frac{1}{N} \log I_{\Delta,N}$$

preserves every endpoint after the corresponding threshold shift, as in (26). Under the endpoint convention of Appendix B.1, an infinite-normalizer comparator has endpoint one in every state and objective $\sum_j w_j(1-p_j) > 0$. Such a branch cannot occur infinitely often in a sequence whose objective tends to zero, so discarding a finite prefix reduces the near-minimizer argument to the finite-normalizer class.

F.5 Scope of the recovery results

The conclusions in this appendix concern one comparator selected from fixed pre-data task information. The mean-Catoni construction gives a global upper bound and the tangent envelope gives an exact compatibility test plus an explicit approximate comparator; neither construction is asserted to be globally optimal outside the proved cases. The local $\Theta(\varepsilon^2)$ law fixes N before sending $\varepsilon \to 0$, and its nonzero-variance assumption is essential. The zero-infimum equivalence requires fixed finite N, finitely many strictly positive weights, and fixed interior targets. It does not extend uniformly to growing state spaces, vanishing weights, moving boundaries, or joint large-N limits.

G Discrete Information-Price Proofs

This appendix proves the fixed-family entropy law in Theorem 39. All logarithms are natural.

G.1 Comparator normalization and the common information bridge

The source and normalized formulations have the same endpoint infima by Appendix B.1. We use $\mathcal{F}_N$ from the main text. The following lemma combines the common tilted law of Appendix B with a local endpoint expansion.

Lemma 38 (Common-tilt information bridge)**.** *Fix $p \in (0,1)$ and a compact interval of empirical risks contained in $(0,p)$. For $f \in \mathcal{F}_N$ and a target risk q, write*

$$d_f(q) = \mathrm{kl}(q,p) - f(q,p), \qquad e_f(q) = B[f(q,\cdot), \mathrm{kl}(q,p)] - p. \tag{90}$$

There is a probability mass function π_f on $\{0,\dots,N\}$, common to every target at endpoint p, such that

$$N d_f(q) \ge \mathrm{KL}(\mathrm{Bin}(N,q) \| \pi_f)\,. \tag{91}$$

Moreover, uniformly over the fixed compact geometry, there are $R < \infty$ and $d_0 > 0$ such that

$$e_f(q) \ge s(q) d_f(q) - R d_f(q)^{3/2}, \qquad s(q) = \frac{p(1-p)}{p-q}, \tag{92}$$

whenever $d_f(q) \le d_0$. For any fixed $h_0 \in (0, 1-p)$, the same interpolation also implies $d_f(q) \le C_0 e_f(q)$ whenever $e_f(q) \le h_0/2$, where C_0 is uniform on the compact geometry.

Proof. Use the common law π_f in (30). The change-of-measure calculation in the proof of Lemma 16, before averaging over states, gives (91) for every target at this same ordinate.

For the endpoint conversion, let $M_q(h) = \mathrm{kl}(q, p+h) - \mathrm{kl}(q,p)$. Convex interpolation between the value of $f(q,\cdot)$ at p and the KL-ceiling point at $p+h$ gives

$$e_f(q) \geq h \frac{d_f(q)}{d_f(q) + M_q(h)}. \tag{93}$$

On a compact right neighborhood of p, Taylor's theorem gives $M_q(h) \leq h/s(q) + Lh^2$ with a uniform L. Taking $h = \sqrt{d_f(q)}$ proves (92). Taking instead $h = h_0$ and rearranging (93) gives

$$d_f(q) \leq \frac{e_f(q) M_q(h_0)}{h_0 - e_f(q)} \leq \frac{2 \sup_q M_q(h_0)}{h_0} e_f(q),$$

which is the fixed-height assertion. □

G.2 Fixed finite families

Consider a finite common-endpoint family

$$\mathsf{S} = \{(w_j, q_j, p) : j = 1, \dots, m\}, \qquad w_j > 0, \quad \sum_j w_j = 1, \quad 0 < q_j < p < 1.$$

If several indices have the same q_j, every comparator gives them the same endpoint. Merge these indices and denote the distinct risks and combined weights by $(q_c, W_c)_{c=1}^r$. This leaves $\mathcal{G}_N^\star(\mathsf{S})$ unchanged. Put

$$s_c = \frac{p(1-p)}{p - q_c}, \quad a_c = W_c s_c, \quad A = \sum_c a_c, \quad \alpha_c^\star = \frac{a_c}{A}. \tag{94}$$

Theorem 39 (A weighted entropy law)**.** *Fix r, positive weights W_c summing to one, and distinct $0 < q_c < p < 1$, with A and $\alpha^\star$ as in (94). With $H(\alpha) = -\sum_j \alpha_j \log \alpha_j$,*

$$\mathcal{G}_N^\star(\mathsf{S}) = \frac{A H(\alpha^\star)}{N} + O_{\mathsf{S}}(N^{-3/2}). \tag{95}$$

The infimum ranges over the entire normalized convex-comparator class.

Proof of Theorem 39. The case $r = 1$ is exact: the Catoni tangent at (q_1, p) is sharp for every N. Assume $r \geq 2$.

For each state, define the Catoni tangent

$$C_c(q,r) = -\beta_c q - \log(1 - r + r e^{-\beta_c}), \qquad \beta_c = \log \frac{p(1-q_c)}{q_c(1-p)}. \tag{96}$$

It is jointly convex, touches binary KL at (q_c, p), and satisfies

$$\mathbb{E}_{K \sim \mathrm{Bin}(N,r)} e^{N C_c(K/N, r)} = 1.$$

For pre-data weights $\alpha \in \Delta_r^\circ$, the single aggregate

$$F_{N,\alpha}(q,r) = \frac{1}{N} \log \sum_{c=1}^r \alpha_c e^{N C_c(q,r)} \tag{97}$$

is therefore proper, jointly convex, lower semicontinuous, and exactly normalized. Since $F_{N,\alpha} \geq C_c + N^{-1} \log \alpha_c$, the exact Catoni root formula gives

$$e_c(F_{N,\alpha}) \leq s_c(1 - \alpha_c^{1/N}) \tag{98}$$

until clipping at one. For fixed positive α_c, clipping is absent for all sufficiently large N. Hence

$$\mathcal{G}_N^\star(\mathsf{S}) \leq \frac{1}{N} \sum_c a_c \log \frac{1}{\alpha_c} + O_{\mathsf{S}}(N^{-2}). \tag{99}$$

The objective on the right is strictly convex and diverges at every face of the simplex. Its unique minimizer is $\alpha_c = a_c/A$, and its minimum is

$$C(\mathsf{S}) = \sum_c a_c \log \frac{A}{a_c} = AH(\alpha^\star). \tag{100}$$

We next prove a matching lower bound without assuming that an optimal comparator exists. For each N, choose $f_N \in \mathcal{F}_N$ with

$$E_N := \sum_c W_c e_{f_N}(q_c) \le \mathcal{G}_N^\star(\mathsf{S}) + N^{-2}.$$

The upper bound implies $E_N = O_{\mathsf{S}}(N^{-1})$. Because the number of states and all positive weights are fixed, $e_{f_N}(q_c) = O_{\mathsf{S}}(N^{-1})$ coordinatewise. The fixed-height part of Lemma 38 then gives $d_{f_N}(q_c) = O_{\mathsf{S}}(N^{-1})$. Combining its local endpoint and common-tilt inequalities yields

$$NE_N \ge \sum_c a_c \,\mathrm{KL}(\mathrm{Bin}(N, q_c) \| \pi_{f_N}) - O_{\mathsf{S}}(N^{-1/2}). \tag{101}$$

Let V have law $\alpha^\star$, and conditionally on $V = c$ let $K \sim \mathrm{Bin}(N, q_c)$. With $\bar\lambda_N = \sum_c \alpha_c^\star \,\mathrm{Bin}(N, q_c)$, the weighted KL centroid identity gives

$$\begin{aligned}\sum_c a_c \,\mathrm{KL}(\mathrm{Bin}(N, q_c) \| \pi_{f_N}) \\ = A\left\{I(V;K) + \mathrm{KL}(\bar\lambda_N \| \pi_{f_N})\right\} \ge AI(V;K).\end{aligned} \tag{102}$$

After merging, the finite set of risks has a positive minimum separation ρ. Nearest-center decoding and Hoeffding's inequality give

$$\mathbb{P}(\widehat{V} \ne V \mid V = c) \le 2e^{-N\rho^2/2}.$$

The nonuniform-prior form of Fano's inequality therefore implies $H(V \mid K) = O_{\mathsf{S}}(e^{-c_{\mathsf{S}}N})$, and hence $I(V;K) = H(\alpha^\star) + O_{\mathsf{S}}(e^{-c_{\mathsf{S}}N})$. Substituting this into (101), and accounting for the N^{-2} near-minimizer error, gives

$$N\mathcal{G}_N^\star(\mathsf{S}) \ge AH(\alpha^\star) - O_{\mathsf{S}}(N^{-1/2}).$$

Together with (99), this proves

$$\mathcal{G}_N^\star(\mathsf{S}) = \frac{AH(\alpha^\star)}{N} + O_{\mathsf{S}}(N^{-3/2}).$$

The remainder is pointwise in the frozen geometry; its constant is not uniform when risks merge, weights vanish, or targets approach the boundary. □

Why the common endpoint is necessary. A naive heterogeneous-endpoint version of the entropy law is false. Fix $A_0 < 0$ and let

$$C_{A_0}(q, r) = A_0 q - \log(1 - r + re^{A_0}).$$

This is a normalized Catoni comparator. Whenever $q/(1-q) = e^{A_0}p/(1-p)$, it touches binary KL at (q, p) and its slice is strictly increasing at that contact. Taking $e^{A_0} = 1/2$ gives the two distinct targets

$$(q_1, p_1) = (1/3, 1/2), \qquad (q_2, p_2) = (3/5, 3/4), \tag{103}$$

which the same comparator serves exactly for every N. Their commitment gap is zero even though an unqualified two-label entropy expression would be positive.

H Exact Calibration and Posterior Learning

This appendix proves Theorem 4, its binomial strengthening, and the paired application. The setting is a finite hypothesis space, binary losses, fixed N, δ, and a prior independent of the sample. No independence between hypothesis loss columns is assumed.

H.1 THE COMPARATOR-TO-CALIBRATION INTERFACE

Define $b_{N,k}(r) = \sum_{j=0}^{k} \binom{N}{j} r^j (1-r)^{N-j}$. For $k < N$ and $0 < \eta < 1$, $\mathrm{CP}_{N,\eta}(k)$ is the unique root of $b_{N,k}(r) = \eta$ in $(0,1)$. Set $\mathrm{CP}_{N,\eta}(N) = 1$ and $\mathrm{CP}_{N,0}(k) = 1$. These are continuous in $\eta \in [0,1)$ for each count. For a true risk r, $b_{N,K}(r)$ is a superuniform discrete p-value: counts with $b_{N,k}(r) \le t$ form an initial segment whose binomial probability is at most t. Consequently, $\Pr_r\{r > \mathrm{CP}_{N,\eta}(K)\} \le \eta$. Exactness denotes binomial inversion, not equality of coverage with $1-\eta$ for every true risk.

For $q = k/N \in (0,1)$, put $p = F(q, \log(1/\delta)/N) > q$. The optimal negative tilt in Chernoff's inequality is finite. Its Markov bound is strict because the binomial law puts positive mass above k, where the event indicator is zero but the exponential is positive. Thus $b_{N,k}(p) < \exp\{-N\,\mathrm{kl}(q,p)\} = \delta$. Strict monotonicity gives $\mathrm{CP}_{N,\delta}(k) < p$. The KL ceiling (Corollary 17) gives $f(q,p) \le \mathrm{kl}(q,p)$ for $f \in \mathcal{F}_N$, so p remains feasible for its endpoint. This proves (7). At $k = 0$ both endpoints equal $1 - \delta^{1/N}$; at $k = N$ both equal one.

H.2 A GLOBAL LEARNING SET

For $x_h = k_h/N$, let $\phi_h(r) = e^{N(r-x_h)_+^2}$ and

$$C(P,Q,x,\eta) = \max\left\{ Q \cdot r : r \in [0,1]^M,\ \sum_h P_h \phi_h(r_h) \le 2/\eta \right\}.$$

Set $C(P,Q,x,0) = 1$. For the true risk $r_h^\star$, Hoeffding gives $\Pr\{N(r_h^\star - x_h)_+^2 > t\} \le e^{-2t}$, hence

$$\mathbb{E}\phi_h(r_h^\star) \le 1 + \int_0^\infty e^t e^{-2t}\, dt = 2.$$

Prior averaging and Markov yield one event of probability at least $1-\eta$ on which the true risk vector is feasible, covering every posterior. For every feasible r and finite $K = \mathrm{KL}(Q\|P)$, change of measure gives

$$N\mathbb{E}_Q(r_h - x_h)_+^2 \le K + \log \sum_h P_h \phi_h(r_h) \le K + \log(2/\eta). \tag{104}$$

The first inequality follows from Jensen applied to $P_h e^{f_h}/Q_h$ on the support of Q. Cauchy–Schwarz and maximization then give $C \le \min\{1, x(Q) + \sqrt{[K + \log(2/\eta)]/N}\}$.

For the intersection used in Theorem 4, let $w_h = (2P_h - 1)_+$ and $w = \sum_h w_h$. At most one w_h is positive. Its risk set is

$$\begin{aligned} \mathcal{R}_\phi(P,k,\delta) = \big\{ r \in [0,1]^M : {} & \delta(1-w)\textstyle\sum_j P_j \phi_j(r_j) \le 2, \\ & r_h \le \mathrm{CP}_{N,\delta w_h P_h}(k_h) \quad \text{for all } h, \\ & \delta w_h \textstyle\sum_{j \ne h} P_j \phi_j(r_j) \le 2 \quad \text{for all } h \big\}. \end{aligned} \tag{105}$$

Set $U_\phi = \max_{r \in \mathcal{R}_\phi} Q \cdot r$, $\mathrm{CP}_{N,0} = 1$, and omit terms with zero coefficient. The three constraint types receive total failure budgets $\delta(1-w)$, $\delta \sum_h w_h P_h$, and $\delta \sum_h w_h (1 - P_h)$, which sum to δ.

H.3 VALIDITY, EXACTNESS, AND CONTINUITY OF THE INTERSECTION

All constraints in (105) are convex in r. At zero each evidence constraint is at most $\delta < 2$, and each CP cap is at least $c_\star = 1 - \delta^{1/N} > 0$. The set is nonempty and compact.

The global constraint fails with probability at most $\delta(1-w)$. For each h with $w_h > 0$, the CP cap fails with probability at most $\delta w_h P_h$, and the residual constraint with probability at most $\delta w_h (1 - P_h)$, by the moment bound on the unnormalized residual masses. Zero coefficients impose no constraint. The budgets sum to δ. Their union bound yields a measurable common event on which the true risk vector belongs to the set, proving simultaneous coverage for every Q.

At $P = \delta_h$, all evidence constraints vanish. The sole nontrivial cap is the original-confidence CP constraint, so

$$U_{\mathrm{sq}}(\delta_h, Q, k, \delta) = Q_h\, \mathrm{CP}_{N,\delta}(k_h) + 1 - Q_h. \tag{106}$$

This proves exactness at $Q = \delta_h$.

For continuity, write the feasible correspondence as $\Gamma(P)$. It is defined by the finite continuous constraint family

$$G_0(P, r) = \delta(1 - w) \sum_j P_j \phi_j(r_j) - 2 \le 0,$$
$$G_h^{\text{cap}}(P, r) = r_h - \text{CP}_{N, \delta w_h P_h}(k_h) \le 0,$$
$$G_h^{\text{res}}(P, r) = \delta w_h \sum_{j \ne h} P_j \phi_j(r_j) - 2 \le 0 \quad (h = 1, \dots, M).$$

Here $P \mapsto w_h = (2P_h - 1)_+$ and the CP endpoint at zero budget are continuous, so no constraint index appears or disappears at a dominant-atom transition. Compactness of $[0, 1]^M$ and the closed graph of Γ give upper hemicontinuity.

For lower hemicontinuity, shrink any feasible r to $r_t = (1 - t)r$, $t \in (0, 1)$. For each evidence constraint $F(P, r) \le 2$, convexity gives

$$F(P, r_t) \le (1 - t)F(P, r) + tF(P, 0) \le 2(1 - t) + t\delta < 2.$$

Every cap has slack at least $tc_\star$. The finite constraint family is continuous in P, so r_t stays feasible in a neighborhood of P. Letting $t \downarrow 0$ proves lower hemicontinuity. Berge's maximum theorem applied to $\max_{r \in \Gamma(P)} Q \cdot r$ proves joint continuity on the full product simplex, including changes in the dominant atom and zero prior masses. The endpoint is one-Lipschitz in total variation in Q. There are finitely many binary loss matrices; continuity on each yields Borel measurability. The simultaneous event can also be expressed on a countable dense set of rational posteriors.

H.4 The KL learning bound and its constant

Suppose a dominant atom h has mass $m \in (1/2, 1)$. Write $\varepsilon = 1 - m$, $a = Q_h$, $q = 1 - a$, and $P_j^- = P_j/\varepsilon$, $Q_j^- = Q_j/q$ for $j \ne h$. If $q = 0$, terms multiplied by q are zero. Put

$$\alpha = (1 - w, wm, w\varepsilon), \qquad \beta = (1 - w, wa, wq), \qquad K_- = \text{KL}(Q^- \| P^-).$$

The intersection endpoint is at most both

$$B_0 = C(P, Q, x, \delta\alpha_0), \qquad B_1 = a\ \text{CP}_{N, \delta\alpha_1}(k_h) + q\, C(P^-, Q^-, x^-, \delta\alpha_2).$$

It is therefore at most $(1 - w)B_0 + wB_1$. Hoeffding inversion gives $\text{CP}_{N,\eta}(k) \le x + \sqrt{\log(1/\eta)/(2N)} \le x + \sqrt{\log(1/\eta)/N}$. If CP is below x this is immediate; otherwise apply Hoeffding at its root. Together with (104) and concavity of square root, using $(K_0, K_1, K_2) = (K, 0, K_-)$ and $(c_0, c_1, c_2) = (2, 1, 2)$,

$$U_{\text{sq}} - x(Q) \le \frac{1}{\sqrt{N}} \left[\sum_{b=0}^{2} \beta_b \{ K_b + \log(1/\delta) + \log(c_b/\alpha_b) \} \right]^{1/2}.$$

The KL chain rule is $K = \text{kl}(a, m) + qK_-$. Since $\text{KL}(\beta \| \alpha) = w\, \text{kl}(a, m)$,

$$\sum_b \beta_b K_b + \text{KL}(\beta \| \alpha) = K, \qquad \sum_b \beta_b \{ K_b + \log(1/\alpha_b) \} = K + H(\beta).$$

Finally,

$$H(\beta) + \sum_b \beta_b \log c_b = \log 5 - \text{KL}\big(\beta \| (2/5, 1/5, 2/5)\big) \le \log 5. \tag{107}$$

This proves (8), with the usual zero-mass convention for zero β_b. Without a dominant atom the global bound at $\eta = \delta$ is stronger. For a point prior, finite KL forces $Q = P$ and the CP bound suffices. The conclusion holds for all $\delta \in (0, 1)$ without dependence on M or the smallest positive prior mass.

H.5 Binomial evidence: validity, convexity, and boundary limits

Take $\psi_h(r) = b_{N,k_h}(r)^{-1/2}$, with $b_{N,N} = 1$ and $\psi_h(1) = +\infty$ when $k_h < N$. Zero-coefficient products are omitted. For a superuniform p-value V,

$$\mathbb{E}V^{-1/2} = 1 + \int_1^\infty \Pr(V < t^{-2})\, dt \le 2.$$

This is standard inverse-square-root p-to-e calibration (Vovk & Wang, 2021). The same budget argument proves validity.

For $k < N$, $b_{N,k}(r)$ is the survival function of a $\mathrm{Beta}(k+1, N-k)$ density g, which is log-concave. Its reciprocal hazard on $(0,1)$ is

$$\frac{\int_r^1 g(s)\, ds}{g(r)} = \int_0^{1-r} \frac{g(r+u)}{g(r)}\, du.$$

Concavity of $\log g$ makes the integrand nonincreasing in r for fixed u, and the interval shrinks. The hazard is therefore nondecreasing, so the logarithm of the survival has nonincreasing derivative. Thus $-\log b_{N,k}$ is convex, and ψ_h is convex and increasing. The extended-value limit preserves convexity at one. Chernoff and Pinsker give

$$\psi_h(r) \ge e^{(N/2)d_+(x_h,r)} \ge e^{N(r-x_h)_+^2} = \phi_h(r). \tag{108}$$

The binomial feasible set is consequently a subset of the square set, proving (9) and the same learning guarantee.

Continuity requires care because ψ_h is unbounded. The map $a\psi_h(r)$, defined as zero at $a = 0$, is lower-semicontinuous for $a \ge 0$. Feasible limits therefore remain feasible. Shrink a feasible r to $r_t = (1-t)r$. Every coordinate is below one, and convexity gives the same strict evidence slack $2(1-t) + t\delta < 2$ and cap slack $tc_\star$ as before. All evidence values at r_t are finite and continuous in the prior coefficients, including at zero. The lower-hemicontinuity proof applies. This proves joint continuity and the specialization (106) for U_{bin}, for arbitrary joint limits of P, Q without a bounded-KL restriction.

For the low-risk guarantee, joint convexity follows from

$$d_+(x,r) = \sup_{t \le 0}\{tx - \log(1 - r + re^t)\}.$$

Change of measure using (108), followed by Jensen, bounds a global binomial support endpoint by $d_+(x(Q), C_{\mathrm{bin}}) \le 2[K + \log(2/\eta)]/N$. The CP coordinate satisfies $Nd_+(x_h, \mathrm{CP}_{N,\eta}(k_h)) \le \log(1/\eta)$. Use the same three-branch combination, relax this CP bound by a factor of two, and apply joint convexity to the weighted empirical risks and endpoints. Equation (107) gives

$$d_+(x(Q), U_{\mathrm{bin}}) \le 2[K + \log(5/\delta)]/N. \tag{109}$$

The intersection endpoint is at most that combination and d_+ is nondecreasing in its second argument. At $x(Q) = 0$, this gives $U_{\mathrm{bin}} \le 1 - \exp\{-2[K + \log(5/\delta)]/N\}$. The contribution is compatibility with exactness and continuity; low-risk PAC–Bayes rates themselves are not new.

H.6 Why the joint properties are substantive

The weighted finite-class CP rule

$$U_{\mathrm{WCP}} = \sum_h Q_h\, \mathrm{CP}_{N,\delta P_h}(k_h)$$

is simultaneously valid by a union bound, continuous under the zero-budget convention, and exact at a shared point mass. It can be tighter on small libraries, as the experiments show. It does not, however, satisfy a uniform KL bound of the form (8). Let $M_N = \lceil e^N \rceil$, $P = Q$ uniform, and all losses zero. Then $K = 0$, but $U_{\mathrm{WCP}} = 1 - (\delta/M_N)^{1/N} \to 1 - e^{-1}$, whereas the right side of (8) tends to zero.

A learning witness also separates our construction from prior-mean transport. Let $X \sim \mathrm{Ber}(\rho)$, $Y = X$, $0 < \rho < 1$, and use classifiers 0, X, and 1 with uniform prior. An empirical-risk minimizer

with ties preferring the constant-zero classifier returns it on an all-zero sample and the perfect classifier otherwise. Both posteriors occur with positive probability; empirical risk is always zero and KL is $\log 3$. At $\delta = 1/20$, direct inversion gives $U_{\mathrm{sq}} = \min\{1, \sqrt{\log 118/N}\}$. It is nonvacuous for $N \geq 5$ and vanishes as the learner selects the zero-risk classifier with probability tending to one. Prior-averaged loss is identically $1/3$ on every observation. Any valid deterministic prior-mean upper endpoint for this constant sample is at least $1/3$. Transport of only that scalar constraint to any one of the point posteriors returns one. This establishes learning compatibility, not dominance over weighted CP.

Requiring every nondegenerate $P = Q$ to return an element of the fixed integer CP grid would be a different, impossible strengthening of continuity. For fixed all-zero and all-one loss columns, continuity on $P = Q = (t, 1-t)$ would give a continuous map from a connected interval into a finite grid. It must be constant, whereas point-mass exactness requires distinct endpoint values $1 - \delta^{1/N}$ and one.

H.7 Paired improvement certification

Fix a baseline classifier h_0 before certification. Define

$$A_h(Z) = \mathbf{1}\{\ell(h, Z) = 1, \ell(h_0, Z) = 0\}, \qquad B_h(Z) = \mathbf{1}\{\ell(h, Z) = 0, \ell(h_0, Z) = 1\}.$$

The loss difference equals $A_h - B_h$ pointwise. Apply either endpoint at failure probability $\delta/2$ to binary columns A and $1 - B$. A union bound gives, simultaneously for every Q,

$$L(Q) - L(h_0) \leq U_\phi(P, Q, k^A, \delta/2) + U_\phi(P, Q, k^{1-B}, \delta/2) - 1 =: \mathcal{U}_\phi(Q) \tag{110}$$

with probability at least $1 - \delta$. The posterior may minimize this expression on the certification sample. The two columns need not be independent. A negative endpoint certifies strict improvement. Binary dominance implies $\mathcal{U}_{\mathrm{bin}} \leq \mathcal{U}_{\mathrm{sq}}$, including after minimizing over the same candidate family. This uses paired disagreements but is not an optimal multinomial certificate.

H.8 Dual computation and numerical upper bounds

There are at most two nontrivial evidence constraints. For dominant atom h, put $a_i = \delta(1-w)P_i$, $b_i = \delta w P_i \mathbf{1}\{i \neq h\}$, $c_h = \mathrm{CP}_{N, \delta w P_h}(k_h)$, and $c_i = 1$ otherwise. Without a dominant atom take $a_i = \delta P_i$, $b_i = 0$, $c_i = 1$. For either evidence E_i, convex duality gives

$$U = \inf_{\lambda, \mu \geq 0} \left\{ 2\lambda + 2\mu + \sum_i \sup_{0 \leq r \leq c_i} [Q_i r - (\lambda a_i + \mu b_i) E_i(r)] \right\}. \tag{111}$$

Zero-coefficient products are omitted. Strict feasibility holds at a sufficiently small positive risk vector, so Slater's condition applies. At zero multipliers the dual is at most one. Evaluating each inner problem at zero gives $D(\lambda, \mu) \geq (2 - \delta)(\lambda + \mu)$, allowing the restriction $\lambda + \mu \leq 1/(2-\delta)$. Point priors use (106) directly.

Each inner objective is concave, with nonincreasing derivative. Boundary checks and bisection locate a maximizer. If $[l_i, u_i]$ encloses that global maximizer and $t_i = \lambda a_i + \mu b_i \geq 0$, monotonicity yields

$$\sup_r [Q_i r - t_i E_i(r)] \leq Q_i u_i - t_i E_i(l_i). \tag{112}$$

Substitution at any nonnegative multipliers yields an upper bound by weak duality; exact outer optimization is unnecessary for this direction. At $t_i = 0$, the inner maximum is $Q_i c_i$. CP is a monotone polynomial inversion for rational inputs, and both evidence profiles admit arbitrary-precision enclosures. Primal feasible points and dual subdivision provide convergent enclosures, without a claim of polynomial running time.

Proposition 40 (Exact variable compression in a fixed support program)**.** *Fix the observed count vector and either the squared or binomial evidence. Partition the positive-prior, nondominant coordinates into groups such that k_i and Q_i/P_i are constant within each group; keep every zero-prior coordinate and the dominant atom, if present, as singletons. In the support program with its original coefficients (a_i, b_i, c_i), tie the variables in each group and replace their terms by $Q_G = \sum_{i \in G} Q_i$, $A_G = \sum_{i \in G} a_i$, and $B_G = \sum_{i \in G} b_i$. This algebraic compression leaves the optimum unchanged. It does not permit replacing a group by a new aggregated prior coordinate and recomputing the dominant-atom split.*

Proof. For a positive-prior group G, replace any feasible $(r_i)_{i\in G}$ by $\bar{r}_G = \sum_{i\in G} P_i r_i / P_G$ on every coordinate in G. Writing $Q_i = \rho_G P_i$, this preserves the group objective: $\sum_{i\in G} Q_i r_i = Q_G \bar{r}_G$. Both evidence profiles are convex in risk and share the same count inside the group, so Jensen gives

$$P_G E_{k_G}(\bar{r}_G) \le \sum_{i\in G} P_i E_{k_i}(r_i).$$

The coefficients of both global and residual evidence constraints are common multiples of P_i within a nondominant group. Hence averaging cannot violate either constraint; the only nontrivial cap belongs to the singleton dominant atom. This maps every original feasible vector to the stated tied-variable program with the same objective. Conversely, expanding a tied variable by setting all r_i in its group equal reproduces its aggregated constraint terms and objective exactly. The two optima are therefore equal. □

After this compression, the outer dual still has at most two multipliers and each evaluation uses one scalar concave maximization per effective group. Thus repeated prediction columns can reduce the practical linear factor from M to the number of distinct $(k_i, Q_i/P_i)$ groups. The original values of w, a_i, b_i, c_i must be retained: merging prior masses and then recomputing the dominant-atom split can define a different endpoint. This is an exact preprocessing result for the fixed program, not a polynomial-time guarantee.

Equation (112) gives an upper bound in exact arithmetic. The empirical comparisons in Appendix O use floating-point evaluations with conservative numerical margins; these evaluations are not interval-certified bounds on the special functions.

I General-Prior Recovery and the Cost of Exactness

This appendix specifies the randomized evaluation interpretation of Open Problem 2 used in Theorem 5. The hypothesis space is finite throughout; the data law is arbitrary and the prior is fixed before sampling. Coverage probabilities include the evaluation randomness. Exact recovery means equality to the original-confidence CP endpoint for the realized count. Grace means convergence to that endpoint as Q approaches the fixed prior in total variation. This differs from joint pathwise continuity in (P, Q) and from a learning remainder that vanishes with N.

I.1 Binomial observations and exact risk transport

For $\ell \in [0,1]$, put $g_P(z) = \sum_h P_h \ell(h, z)$. Let $V_1, \dots, V_N$ be independent uniform variables on $(0,1)$, independent of S, and define

$$Y_i(P) = \mathbf{1}\{V_i \le g_P(Z_i)\}, \qquad K_P = \sum_i Y_i(P). \tag{113}$$

Each $Y_i(P)$ has success probability $\mathbb{E} g_P(Z) = L(P)$, and the pairs (Z_i, V_i) are independent and identically distributed. Thus $K_P \sim \mathrm{Bin}(N, L(P))$ for every fixed P and data law. For binary loss, fresh independent draws $H_i \sim P$ and counts $\sum_i \ell(H_i, Z_i)$ give the same binomial law. Reusing one draw of H generally gives a mixture of binomials instead. Conditional on S, the uniform-threshold count need not be binomial; the stated law is unconditional over (S, V). Foong et al. already discuss Bernoulli randomization of stochastic losses for test-set evaluation (Foong et al., 2021, Appendix I.2).

Operational protocol. Before evaluation, fix P. For each fresh test observation Z_i, evaluate the loss vector, compute $g_P(Z_i)$, draw one fresh V_i, and retain $Y_i(P)$. After all N augmented observations have been recorded, a learner may choose Q as any measurable function of (S, V); compute $c = \mathrm{CP}_{N,\delta}(\sum_i Y_i(P))$ and then solve (114). The scalar dual displayed below is a one-dimensional convex piecewise-linear problem whose breakpoints are the ratios Q_h/P_h for $P_h > 0$; sorting these ratios gives a direct greedy implementation. Thus the protocol needs one auxiliary uniform per test example and no sampled hypothesis. The uniforms are part of the certified evaluation record and may be used by the posterior-selection rule because the coverage event is simultaneous over all such Q.

Define the finite risk transport

$$T(P, Q, c) = \max\{Q \cdot r : r \in [0,1]^M,\ P \cdot r \le c\}, \qquad 0 \le c \le 1. \tag{114}$$

This is the maximal power of a level-c randomized test between P and Q, equivalently one minus the testing trade-off function (Dong et al., 2022, Definition 2). We use this classical optimization as a transport operation, not as a new testing principle.

Proposition 41 (Greedy transport evaluation)**.** *The value $T(P, Q, c)$ is computed by first setting $r_h = 1$ on every coordinate with $P_h = 0$, then ordering the remaining coordinates by nonincreasing Q_h/P_h, filling them with $r_h = 1$ until the next coordinate would exceed the budget c, and using at most one fractional coordinate to exhaust the remaining budget. Coordinates after it receive zero. Any order within a tie has the same value. After the ratios are sorted, evaluation is linear in M; including the sort, it costs $O(M \log M)$ time.*

Proof. Zero-prior coordinates have no budget cost, so setting them to one cannot decrease the objective. For positive-prior coordinates, suppose $Q_i/P_i > Q_j/P_j$, $r_i < 1$, and $r_j > 0$. Moving an amount of P-budget from j to i strictly increases the objective until i is full or j is empty. Repeating this exchange proves the threshold form. Equal ratios make the exchange value-neutral, and a final fractional coordinate enforces the budget whenever $c < 1$. □

Proposition 42 (Exact recovery at every prior)**.** *For every finite M, $N \ge 1$, $\delta \in (0,1)$, fixed prior P, and bounded loss, set $c = \mathrm{CP}_{N,\delta}(K_P)$ and $U_{\mathrm{tr}} = T(P, Q, c)$. There is one measurable event of probability at least $1 - \delta$ on which $L(Q) \le U_{\mathrm{tr}}$ for all Q, including posteriors chosen using (S, V). For every realized augmented sample,*

$$U_{\mathrm{tr}}(P, P) = c, \qquad 0 \le U_{\mathrm{tr}}(P, Q) - c \le \mathrm{TV}(P, Q) \le \sqrt{\mathrm{KL}(Q\|P)/2}. \tag{115}$$

The last bound is understood as trivial at infinite KL.

Proof. Binomial inversion gives the measurable event $E = \{L(P) \le c\}$ with probability at least $1 - \delta$. On E, the true risk vector belongs to the feasible set in (114); hence all linear posterior risks are controlled on that same event, without a union bound over Q. The constant vector $r = c\mathbf{1}$ is feasible and attains c under P, so $T(P, P, c) = c$ and $T(P, Q, c) \ge c$. Every feasible r satisfies $Q \cdot r \le P \cdot r + \mathrm{TV}(P, Q) \le c + \mathrm{TV}(P, Q)$. Pinsker gives the last inequality.

Finite-dimensional linear programming also gives

$$T(P, Q, c) = \inf_{\lambda \ge 0} \left\{ \lambda c + \sum_h (Q_h - \lambda P_h)_+ \right\}.$$

The infimum can be taken over nonnegative rational λ, since the objective is continuous in λ. This proves Borel measurability in (P, Q, c). It also covers zero prior masses. Alternatively, at the CP values $c > 0$, shrinking a feasible risk vector supplies the slack needed for continuity of the finite program. The common event E itself avoids any measurability issue from an uncountable intersection of posterior events. □

I.2 When deterministic binomialization is possible

Proposition 43 (Deterministic count characterization)**.** *Fix P, a measurable bounded loss, and $N \ge 1$. A deterministic measurable statistic $K_N(S) \in \{0, \dots, N\}$, independent of the unknown data law, satisfies $K_N \sim \mathrm{Bin}(N, L_{\mathcal{D}}(P))$ for every $\mathcal{D}$ if and only if $g_P(z) \in \{0, 1\}$ for every z.*

Proof. If g_P is binary, use $K_N = \sum_i g_P(Z_i)$. Conversely, under $\mathcal{D} = \delta_z$, every deterministic statistic of S is a constant. The binomial law with parameter $g_P(z)$ is nondegenerate unless this parameter is zero or one. Since z is arbitrary, necessity follows. □

For binary losses, the condition says that, at every observation, the loss is P-almost surely constant across hypotheses. A non-point prior on loss-equivalent hypotheses therefore causes no obstruction. The proposition requires the full class of data laws, including point masses, and an actual binomial count. It does not rule out deterministic bounded-mean confidence bounds or their extensions to fractional observations.

I.3 A LOWER BOUND FORCED BY THE ORIGINAL CONFIDENCE BUDGET

We prove a result that applies to both evaluation experiments above, even when an endpoint uses all observed losses and additional randomization. Let $\mathcal{Z} = \{0,1\}^M$ and $\ell(h,z) = z_h$. Conditional on the full sample $s \in \mathcal{Z}^N$, an auxiliary output a is drawn from a fixed measurable kernel A_s, independent of the unknown data law. The kernel may depend on the fixed P, N, δ. Write $\omega = (s,a)$. For $r \in (0,1)^M$, let $\mathcal{D}_r$ be the independent-coordinate Bernoulli law with means r_h.

Theorem 44 (Previous-count transport lower bound)**.** *Suppose a measurable count $K(\omega) \in \{0,\dots,N\}$ has law $\mathrm{Bin}(N, P\cdot r)$ under every $\mathcal{D}_r$. Let $U(\omega,Q) \in [0,1]$ be measurable for each fixed Q and independent of the unknown data law. Assume that, under every data law, (i) one measurable event of probability at least $1-\delta$ covers all posterior risks and (ii) $U(\omega,P) = \mathrm{CP}_{N,\delta}(K(\omega))$ almost surely. Then, for each fixed Q and every data law on $\mathcal{Z}$, almost surely on $K \geq 1$,*

$$U(\omega,Q) \geq T\big(P,Q,\mathrm{CP}_{N,\delta}(K(\omega)-1)\big). \tag{116}$$

When the auxiliary record consists only of fresh draws $H_i \sim P$, this holds on every augmented sample with $H_i \in \mathrm{supp}(P)$. If further continuous randomization is added, or for uniform-threshold evaluation, the conclusion is almost sure rather than a claim about every zero-probability path.

Proof. Let μ be the augmented law obtained from $\mathcal{D}^N_{(1/2,\dots,1/2)}$ and the same kernels A_s. Each data sample s has positive μ-probability. The augmented law ν_r under an interior parameter r is equivalent to μ, with density

$$\frac{d\nu_r}{d\mu}(s,a) = \frac{\mathcal{D}^N_r(s)}{\mathcal{D}^N_{(1/2,\dots,1/2)}(s)} > 0.$$

For every fixed measurable set B, the probability $\nu_r(B) = \sum_s \mathcal{D}^N_r(s) A_s(B_s)$ is continuous in r, because the sum is finite. No continuity of U is required.

Fix Q and $k \in \{1,\dots,N\}$, and set $c = \mathrm{CP}_{N,\delta}(k-1) \in (0,1)$. Choose any interior r with $P\cdot r = c$. Suppose $B = \{K = k,\ U(\omega,Q) < Q\cdot r\}$ has positive μ-measure. Then $\nu_r(B) > 0$. Set $r_\varepsilon = (1-\varepsilon)r + \varepsilon\mathbf{1}$, so $c_\varepsilon = P\cdot r_\varepsilon > c$ and $Q\cdot r_\varepsilon \geq Q\cdot r$. Every point of B remains a posterior-coverage failure under this perturbed law.

On $A = \{K \leq k-1\}$, prior exactness forces $U(\omega,P) \leq c < c_\varepsilon = L_{\mathcal{D}_{r_\varepsilon}}(P)$ almost surely. Thus A is a prior-coverage failure and is disjoint from B. The binomial law and CP identity give

$$\nu_{r_\varepsilon}(A) = b_{N,k-1}(c_\varepsilon) \longrightarrow b_{N,k-1}(c) = \delta, \qquad \nu_{r_\varepsilon}(B) \longrightarrow \nu_r(B) > 0.$$

For small ε, even the joint coverage failure of P and this fixed Q therefore exceeds δ, a contradiction. Hence $\mu(B) = 0$. The exactness exceptional set can be taken μ-null and remains null under all interior laws by equivalence.

Take a countable dense subset of the interior slice $\{r \in (0,1)^M : P\cdot r = c\}$. Outside the union of its null sets, on $K = k$ we obtain

$$U(\omega,Q) \geq \sup_{r\in(0,1)^M : P\cdot r = c} Q\cdot r = T(P,Q,c).$$

To verify the last equality, take a maximizer r^* of the compact program (114). If $P\cdot r^* < c$, move toward $\mathbf{1}$ until equality holds; the objective cannot decrease. Then $(1-t)r^* + tc\mathbf{1}$ is interior, has prior mean c, and approaches r^* as $t \downarrow 0$.

A finite union over k proves the claim μ-almost surely for this fixed Q. Every data law on the finite alphabet, augmented by the same kernels, is absolutely continuous with respect to μ, so the conclusion holds under each such law. For the finite H-draw experiment, every path with $H_i \in \mathrm{supp}(P)$ is an atom of positive μ-probability; none can belong to an exceptional null set. □

For continuous auxiliary randomness, the null set above may depend on Q. We do not exchange this quantifier with a simultaneous almost-sure lower bound without additional regularity. The fixed-Q conclusion suffices for the following impossibility result. We first isolate the discrete quantile estimate used in its finite-sample form.

Lemma 45 (Fair-binomial quantile overshoot)**.** *Let $K \sim \mathrm{Bin}(N, 1/2)$ and let a_N be the least integer with $\Pr\{K \le a_N\} \ge \delta$. Then*

$$\Pr\{K \ge a_N + 1\} \ge 1 - \delta - \sqrt{\frac{2}{\pi N}}, \qquad \mathrm{CP}_{N,\delta}(K-1) \ge \frac{1}{2} \quad \text{on } \{K \ge a_N + 1\}. \tag{117}$$

Proof. Minimality of a_N gives $\Pr\{K \le a_N\} < \delta + \Pr\{K = a_N\}$, where the preceding CDF is understood as zero if $a_N = 0$. The largest fair-binomial atom is at most $\sqrt{2/(\pi N)}$, proving the probability bound. For $K \ge a_N + 1$, $\Pr\{\mathrm{Bin}(N, 1/2) \le K - 1\} \ge \delta$; monotonicity of this CDF in its success parameter gives the CP inequality. □

Corollary 46 (A fixed-task obstruction to vanishing learning remainders)**.** *In either evaluation experiment above, every sequence of endpoints satisfying Theorem 44 admits a fixed two-classifier task and fixed uniform prior for which the unique empirical-risk minimizer Q satisfies $L(Q) = \widehat{L}_S(Q) = 0$, $\mathrm{KL}(Q\|P) = \log 2$, but $U(\omega, Q) \to 1$ in probability and in mean. Moreover, for every N,*

$$\Pr\{U(\omega, Q) = 1\} \ge 1 - \delta - \sqrt{\frac{2}{\pi N}}. \tag{118}$$

Consequently, a universal empirical-risk remainder tending to zero at fixed KL is impossible, even with any fixed positive probability of success.

Proof. Take $\mathcal{D} = \delta_{(0,1)}$, $P = (1/2, 1/2)$, and $Q = \delta_1$. Then $K \sim \mathrm{Bin}(N, 1/2)$ in either experiment, while $L(Q) = \widehat{L}_S(Q) = 0$ and $\mathrm{KL}(Q\|P) = \log 2$. For $K \ge 1$, (116) yields

$$U(\omega, Q) \ge \min\{1, 2\,\mathrm{CP}_{N,\delta}(K-1)\}.$$

Lemma 45 shows that the displayed lower bound equals one with probability at least $1 - \delta - \sqrt{2/(\pi N)}$, proving (118). For fixed δ, Hoeffding bounds at the CP root give, for $j < N$,

$$\frac{j}{N} - \sqrt{\frac{\log(1/(1-\delta))}{2N}} \le \mathrm{CP}_{N,\delta}(j) \le \frac{j}{N} + \sqrt{\frac{\log(1/\delta)}{2N}}.$$

For the lower inequality when the root is below j/N, use $1 - \delta = \Pr\{\mathrm{Bin}(N, c) > j\} \le e^{-2N(j/N-c)^2}$; the other cases are immediate or use the lower tail. Since $K/N \to 1/2$ and $\Pr(K = 0) = 2^{-N}$, the lower bound on U tends to one in probability. Boundedness gives convergence in mean. For any deterministic $\rho_N \to 0$, $\Pr\{U \le \widehat{L}_S(Q) + \rho_N\} \to 0$.

Realize this as binary classification by taking $X \in \{0, 1\}^2$, label always zero, and classifiers $h_1(X) = X_1$, $h_2(X) = X_2$. The true feature law is the fixed point mass at $(0, 1)$, so ordinary ERM uniquely selects h_1. The alternative product-Bernoulli feature laws used in the lower bound belong to the same fixed classification problem. □

I.4 Interpretation and limits

Proposition 42 gives exact recovery by transport, whereas Corollary 46 proves the main-text obstruction in Theorem 5: randomized Gibbs evaluation permits general-prior CP recovery with a total-variation remainder, while exact recovery and common coverage force saturation on a fixed zero-risk ERM task. The lower bound applies to every rule satisfying these requirements. It requires equality to original-confidence CP and validity throughout the product-Bernoulli risk-vector model; approximate recovery is treated next. Appendix O tabulates the finite-sample saturation bound alongside the distinct learning-compatible endpoints.

J Calibration on Known Loss-Equivalence Classes

The equivalence relation in this section is fixed before the certification sample: two hypotheses are equivalent only if their losses agree on every possible observation. It cannot be inferred from matching sample columns.

Proposition 47 (Calibration on a known quotient)**.** *Fix before sampling a partition $\mathcal{E}$ of the finite hypothesis space such that $\ell(h,z) = \ell(h',z)$ for every z whenever h, h' lie in one class. Let $\bar{P}_E = \sum_{h\in E} P_h$, $\bar{Q}_E = \sum_{h\in E} Q_h$, and let k_E be the common count. Applying U_{bin} or U_{sq} to $(\bar{P}, \bar{Q}, (k_E)_E)$ gives simultaneous posterior coverage, joint continuity in (P, Q), and the stated learning bounds with $\mathrm{KL}(\bar{Q}\|\bar{P}) \le \mathrm{KL}(Q\|P)$. If P, Q both concentrate on one class E, the endpoint is $\mathrm{CP}_{N,\delta}(k_E)$. The endpoint is unchanged by redistributing mass or adding copies within classes while preserving their total prior and posterior masses.*

Proof. The representative losses define a fixed finite binary-loss problem. Its empirical and population risks under $\bar{Q}$ equal the original risks under Q. Apply Theorem 4 and its binomial strengthening on this quotient. Every original posterior maps to a quotient posterior, so the quotient common event covers all original posteriors. The two pushforward maps are linear and continuous. The log-sum inequality gives the KL comparison, including the extended-value cases. Concentration on one class becomes a shared point mass on the quotient. All computed quantities depend only on class totals and representative counts. □

J.1 Recovering learning with a split confidence budget

A split confidence budget and a calibration floor give two-sided, pathwise approximate recovery without estimating a law on the full loss-vector alphabet. Write $\hat{L}(Q) = N^{-1}\sum_i \sum_h Q_h \ell(h, Z_i)$ and let $B_\alpha(P,Q)$ be the largest $u \in [\hat{L}(Q), 1]$ satisfying

$$N\,\mathrm{kl}(\hat{L}(Q), u) \le \mathrm{KL}(Q\|P) + \log\frac{N+1}{\alpha}.$$

For a full-support prior, put $p_{\min} = \min_h P_h$.

Proposition 48 (Dimension-free approximate CP calibration)**.** *Fix finite M, a full-support prior P, $N \ge 1$, $\delta \in (0,1)$, and $\epsilon > 0$ before the augmented sample. Set*

$$g_\delta = \min\{\log(1/\delta), \tfrac{1}{2}\log(2/\delta)\}, \quad \eta = \min\{\tfrac{1}{2}, g_\delta\epsilon\}, \quad \alpha = \eta\delta,$$
$$c = \mathrm{CP}_{N,\delta}(K_P), \quad c' = \mathrm{CP}_{N,(1-\eta)\delta}(K_P).$$

Then the endpoint, continuous in Q,

$$H_\epsilon(P,Q) = \max\big\{\min\{T(P,Q,c'), B_\alpha(P,Q)\}, [c - \epsilon - p_{\min}^{-1}\,\mathrm{TV}(P,Q)]_+\big\} \tag{119}$$

has simultaneous coverage at level $1-\delta$ for every data law and every posterior selected using (S,V). On every augmented sample, $|H_\epsilon(P,P) - c| \le \epsilon$. Moreover, for every Q,

$$[H_\epsilon(P,Q) - \hat{L}(Q)]_+ \le \max\left\{\sqrt{\frac{\mathrm{KL}(Q\|P) + \log((N+1)/\alpha)}{2N}}, [c - \hat{L}(P)]_+\right\}. \tag{120}$$

Consequently, for any $K \ge 0$, $\tau \in (0,1)$ and any data law, the $(1-\tau)$-quantile of the maximum width over $\{Q : \mathrm{KL}(Q\|P) \le K\}$ is at most

$$\max\left\{\sqrt{\frac{K + \log((N+1)/\alpha)}{2N}}, \sqrt{\frac{\log(1/\delta)}{2N}} + \sqrt{\frac{\log(1/\tau)}{2N}}\right\}. \tag{121}$$

The bound is independent of M and holds at each N, including for growing finite libraries. Fixed-Q evaluation requires a scalar binomial-KL inversion and a fractional-knapsack transport on M coordinates; it does not enumerate the 2^M loss patterns.

Proof. For each h, the binary-loss binomial moment satisfies

$$\mathbb{E}_S e^{N\,\mathrm{kl}(\hat{L}(\delta_h), L(\delta_h))} \le \sum_{k=0}^{N}\binom{N}{k}(k/N)^k(1-k/N)^{N-k} \le N+1.$$

The first inequality is equality for interior population risks; at risk zero or one the moment equals one. Markov's inequality, the change-of-measure inequality, and joint convexity of kl therefore give, with probability at least $1-\alpha$ over S, $L(Q) \le B_\alpha(P,Q)$ simultaneously for all Q. This event remains

common if Q depends on V. The CP transport at level $(1-\eta)\delta$ has a simultaneous coverage event on (S, V); their intersection has probability at least $1-\delta$. Taking their minimum and then a maximum preserves coverage.

At $Q = P$, transport equals c'. The CP perturbation inequality (138) gives $c \le c' \le c + \epsilon$, while the second argument of the maximum in (119) is at least $c - \epsilon$. This proves pathwise calibration. The first argument of the maximum is at most $B_\alpha(P, Q)$, whose excess over $\hat{L}(Q)$ is at most the square-root term in (120) by Pinsker's inequality. Since $\hat{L}(P) - \hat{L}(Q) \le \mathrm{TV}(P, Q)$ and $p_{\min}^{-1} \ge 1$, the excess of the second argument is at most $[c - \hat{L}(P)]_+$. This proves (120).

Conditional on S, the randomized count has independent summands with mean $N\hat{L}(P)$. Hoeffding's inequality bounds $K_P/N - \hat{L}(P)$ by $\sqrt{\log(1/\tau)/(2N)}$ with probability $1-\tau$. The CP root bound (137) gives $c - K_P/N \le \sqrt{\log(1/\delta)/(2N)}$ on every draw. Substitution in (120) proves (121). Continuity follows from continuity of the transport, the KL inverse, and total variation on the finite simplex. The transport is a fractional-knapsack problem, solved by sorting the ratios Q_h/P_h. □

For fixed δ, τ, K and $\log(1/\epsilon_N) = o(N)$, the quantile bound vanishes even when $M = M_N$ grows. This construction need not attain the sharp local constants of Theorem 6. The full-law construction below attains them for fixed models; Appendix L removes the exponential alphabet overhead at a fixed positive exponent.

K Approximate Calibration and Post-Selection Width

This appendix proves Theorem 6 and its local and deterministic-task refinements. The randomized prior count and transport endpoint are defined in Section 5. The relative-entropy ambiguity set is a standard statistical construction; the calibration constraint supplies the additional prior-risk restriction.

K.1 Problem and main conclusions

Fix an integer $M \ge 1$, the alphabet $\mathcal{Z} = \{0,1\}^M$ of size $J = 2^M$, a prior $P \in \Delta_M$ with $P_h > 0$, a KL radius $K \ge 0$, and $\delta, \tau \in (0,1)$. All of these are fixed as $N \to \infty$. For a law $D \in \Delta_J$, write

$$r(D) = \sum_z D(z)z, \qquad b(D) = P \cdot r(D), \qquad \widehat{D}_N = N^{-1}\sum_{i=1}^N \delta_{Z_i}, \quad \hat{r} = r(\widehat{D}_N).$$

The observations are $Z_i \overset{\text{iid}}{\sim} D$ and independent $V_i \overset{\text{iid}}{\sim} \mathrm{Unif}(0,1)$. Define

$$Y_i = \mathbf{1}\{V_i \le P \cdot Z_i\}, \quad K_P = \sum_i Y_i, \quad c_{N,a}(k) = \mathrm{CP}_{N,a}(k).$$

Thus $K_P \sim \mathrm{Bin}(N, b(D))$ unconditionally on the augmented space. Conditional on S, the Y_i are independent Bernoulli variables with generally unequal success probabilities $P \cdot Z_i$.

Let $\mathcal{A}_N^+(\epsilon)$ consist of jointly Borel rules $U_N(S, V, Q) \in [0,1]$ with the following properties:

$$\text{for every } D, \text{ there is a measurable } E_D \text{ with } \mathbb{P}_{D,V}(E_D) \ge 1-\delta,$$
$$Q \cdot r(D) \le U_N(Q) \quad \text{on } E_D \text{ simultaneously for every } Q \in \Delta_M; \tag{122}$$
$$U_N(P) \le c_{N,\delta}(K_P) + \epsilon \quad \text{almost surely under every } D. \tag{123}$$

Q may depend on all of (S, V). This is a one-sided upper-calibration requirement: the rule may lie below CP at the prior. Write $\mathcal{A}_N^{\pm}(\epsilon)$ for the subclass satisfying the stronger pathwise condition $|U_N(P) - c_{N,\delta}(K_P)| \le \epsilon$. We call $\epsilon = 0$ zero-tolerance one-sided calibration; equality with CP belongs to the stronger class. No continuity in Q is imposed on lower-bound competitors. Put

$$\mathcal{Q} = \{Q \in \Delta_M : \mathrm{KL}(Q\|P) \le K\},$$
$$W_N(U) = \sup_{Q \in \mathcal{Q}} [U_N(Q) - Q \cdot \hat{r}]_+,$$
$$w_{1-\tau}(U; D) = \inf\{w \in [0,1] : \mathbb{P}_{D,V}(W_N(U) \le w) \ge 1-\tau\}, \tag{124}$$
$$\mathfrak{R}_{N,+}^{\mathrm{loc}}(D; \epsilon) = \inf_{U \in \mathcal{A}_N^+(\epsilon)} w_{1-\tau}(U; D), \qquad \mathfrak{R}_{N,+}^{\mathrm{glob}}(\epsilon) = \inf_{U \in \mathcal{A}_N^+(\epsilon)} \sup_D w_{1-\tau}(U; D). \tag{125}$$

The admissibility condition still quantifies over *every* law in the local problem. These are not expectation objectives. Since $\mathcal{Q}$ is compact, the supremum in (124) of a jointly Borel function is upper semianalytic and universally measurable; probabilities are taken in the completed observation law. One may equivalently restrict to competitors for which W_N is measurable. If only fixed-Q section measurability is stipulated, measurability of W_N must be added; it does not follow automatically. Our constructed rules are continuous in Q and cause no such issue.

K.1.1 Three deterministic functionals

The classical testing transport is

$$T(P,Q,c) = \max\{Q \cdot x : x \in [0,1]^M,\ P \cdot x \le c\}. \tag{126}$$

Define the geometry constants

$$\gamma_P(Q) = \max_{v \in [-1,1]^M :\, P\cdot v = 0} Q \cdot v, \qquad \Gamma(P,K) = \max_{Q\in\mathcal{Q}} \gamma_P(Q), \tag{127}$$

$$L(P,K) = \max_{Q\in\mathcal{Q}} \max_h \frac{Q_h}{P_h} \le p_{\min}^{-1}, \qquad p_{\min} = \min_h P_h.$$

The classical sharp KL/TV envelope, with TV equal to half the ℓ^1 distance, is

$$\nu(\lambda) = \max\{v - u : 0 \le u \le v \le 1,\ \mathrm{kl}(u,v) \le \lambda\}, \quad \nu(0) = 0, \quad \nu(\infty) = 1. \tag{128}$$

Finally, for $0 \le \lambda < \infty$ define

$$A_\lambda(D;P,K) = \max_{\substack{Q\in\mathcal{Q},\ E\in\Delta_J \\ \mathrm{KL}(D\|E)\le\lambda,\ b(E)\le b(D)}} [Q\cdot(r(E) - r(D))]_+. \tag{129}$$

For $\lambda = \infty$ remove the KL constraint. Its direction is $\mathrm{KL}(D\|E)$, not $\mathrm{KL}(E\|D)$. The finite-λ feasible set is compact by lower semicontinuity of relative entropy, and contains $E = D$.

Theorem 49 (Matched local and global quantile limits)**.** *Let $\epsilon_N > 0$, $\epsilon_N \to 0$, and*

$$\lambda_N = \frac{1}{N}\log\frac{1}{\epsilon_N} \longrightarrow \lambda \in [0,\infty].$$

For every fixed D and every fixed $\tau \in (0,1)$,

$$\mathfrak{R}^{\mathrm{loc}}_{N,+}(D;\epsilon_N) \longrightarrow A_\lambda(D;P,K), \tag{130}$$

$$\mathfrak{R}^{\mathrm{glob}}_{N,+}(\epsilon_N) \longrightarrow \Gamma(P,K)\nu(\lambda). \tag{131}$$

A single explicit sequence of rules, independent of D, K, τ, attains the local limits for every fixed D, K, τ and attains the global limit. More precisely, for every $t < A_\lambda(D;P,K)$ with $t \ge 0$,

$$\sup_{U\in\mathcal{A}_N^+(\epsilon_N)} \mathbb{P}_{D,V}\{W_N(U) < t\} \longrightarrow 0. \tag{132}$$

For zero-tolerance one-sided calibration,

$$\mathfrak{R}^{\mathrm{loc}}_{N,+}(D;0) \longrightarrow A_\infty(D;P,K) = \max_{Q\in\mathcal{Q}}\{T(P,Q,b(D)) - Q\cdot r(D)\}, \tag{133}$$

$$\mathfrak{R}^{\mathrm{glob}}_{N,+}(0) \longrightarrow \Gamma(P,K). \tag{134}$$

The transport rule attains these limits and is CP-exact at the prior; the approximate upper-bound constructions used here satisfy the stronger two-sided condition.

Corollary 50 (Deterministic tasks and nonuniform two-classifier priors)**.** *For $z \in \{0,1\}^M$, put*

$$\Psi_z(P,K) = \max_{Q\in\mathcal{Q}}\{T(P,Q,P\cdot z) - Q\cdot z\}.$$

Then

$$A_\lambda(\delta_z;P,K) = (1-e^{-\lambda})\Psi_z(P,K), \quad e^{-\infty} = 0. \tag{135}$$

In particular, for $M = 2$, $P = (\pi, 1-\pi)$, $D_0 = \delta_{(0,1)}$, and $Q_0 = (1,0)$, define

$$\ell(\pi) = \min\{1, (1-\pi)/\pi\}.$$

The optimal fixed-Q_0 certificate quantile has limit

$$\ell(\pi)(1-e^{-\lambda}) \quad \textit{under approximate calibration}, \qquad \ell(\pi) \quad \textit{under exact calibration}. \tag{136}$$

The corresponding universal directional lower bound holds with probability tending to one. A matching admissible construction converges in probability to the displayed value. For every fixed $0 < \pi < 1$, an admissible sequence with $U_N(Q_0) \to 0$ in probability exists if and only if $\lambda_N \to 0$.

The directional quantile in (136) is the quantile of $U_N(Q_0)$, whose empirical risk on D_0 is zero. Section K.6 separately computes the KL-ball width.

K.2 CALIBRATION FACTS AND THE PREVIOUS-COUNT BOUND

For $0 \le k < N$, set $b_{N,k}(p) = \mathbb{P}\{\mathrm{Bin}(N,p) \le k\}$, and define $c_{N,a}(k)$ by $b_{N,k}(c_{N,a}(k)) = a$. Set $c_{N,a}(N) = 1$. Let

$$h_a^+(N) = \sqrt{\frac{\log(1/a)}{2N}}, \quad h_a^-(N) = \sqrt{\frac{\log(1/(1-a))}{2N}}, \quad h_a(N) = \max(h_a^+, h_a^-).$$

Lemma 51 (CP inversion and confidence-level perturbation)**.** *For every $N \ge 1$, $a \in (0,1)$, and $k \in \{0,\dots,N\}$,*

$$\frac{k}{N} - h_a^-(N) \le c_{N,a}(k) \le \frac{k}{N} + h_a^+(N). \tag{137}$$

The endpoint is valid at failure probability a. If $0 < \eta \le 1/2$ and

$$g_a = \min\{\log(1/a), \tfrac{1}{2}\log(2/a)\},$$

then

$$0 \le c_{N,(1-\eta)a}(k) - c_{N,a}(k) \le \eta/g_a. \tag{138}$$

Also $|b'_{N,k}(p)| \le N$ for $k < N$.

Proof. Monotonicity of the binomial lower tail gives coverage by inversion. If the CP root lies above k/N, Hoeffding's lower-tail inequality at that root gives the upper bound in (137). If it lies below k/N, apply the upper-tail inequality to $1 - a = \mathbb{P}\{\mathrm{Bin}(N,c) > k\}$ for the lower bound. The remaining cases are immediate.

For the perturbation bound, differentiation gives

$$-b'_{N,k}(p) = N\binom{N-1}{k}p^k(1-p)^{N-k-1} =: f(p),$$

which is a log-concave beta density and is at most N. Its survival b has increasing hazard $H = f/b$: indeed $b(p)/f(p) = \int_0^{1-p} f(p+u)/f(p)\,du$ is nonincreasing by log concavity. Thus $-\log b(p) = \int_0^p H(t)dt \le H(p)$ and $-b'(p) \ge b(p)\log(1/b(p))$. Between the two roots, $b(p) \in [a/2, a]$, on which $b\log(1/b) \ge ag_a$ by concavity and inspection of the endpoints. Integrating gives $\eta a \ge ag_a(c_{N,(1-\eta)a} - c_{N,a})$. The case $k = N$ is constant. □

Lemma 52 (Transport properties)**.** *The function $T(P,Q,c)$ is continuous and is L_Q-Lipschitz in c, where $L_Q = \max_h Q_h/P_h$. It satisfies*

$$T(P,P,c) = c, \qquad c \le T(P,Q,c) \le c + \mathrm{TV}(P,Q).$$

$U_{\mathrm{tr}}(Q) = T(P,Q,c_{N,\delta}(K_P))$ has common coverage and exact calibration.

Proof. On $\{b(D) \le c_{N,\delta}(K_P)\}$, the true risk vector is feasible in every transport problem. This gives the common event. The constant vector $c\mathbf{1}$ is feasible; $Q \cdot x \le P \cdot x + \mathrm{TV}(P,Q)$ proves the two inequalities. The fractional-knapsack solution fills coordinates in decreasing order of Q_h/P_h; its slopes in c lie in $[0, L_Q]$, proving Lipschitz continuity. Equivalently,

$$T(P,Q,c) = \min_{s \ge 0}\{sc + \sum_h (Q_h - sP_h)_+\},$$

with $s \le L_Q$ sufficient. Continuity in Q follows directly by taking a maximum of linear functions over a common compact set. □

Lemma 53 (Previous-count bound)**.** *Let $U \in \mathcal{A}_N^+(0)$ and fix Q. Almost surely under every law D, on $K_P \ge 1$,*

$$U(Q) \ge T(P,Q,c_{N,\delta}(K_P - 1)). \tag{139}$$

The null set may depend on Q.

Proof. Let μ be the augmented law from the full-support product Bernoulli law with all means $1/2$. For interior $r \in (0,1)^M$, its product Bernoulli law induces an augmented law μ_r equivalent to μ. For every measurable augmented event B, $\mu_r(B)$ is continuous in r: it is a finite sum over S of product-Bernoulli sample probabilities times fixed conditional probabilities over V.

Fix $k \geq 1$, $c = c_{N,\delta}(k-1)$, and an interior r with $P \cdot r = c$. Suppose $B = \{K_P = k, U(Q) < Q \cdot r\}$ has positive μ-measure. Under $r_t = (1-t)r + t\mathbf{1}$ it remains a posterior failure. On $A = \{K_P \leq k-1\}$, the one-sided prior constraint forces a prior failure since $P \cdot r_t > c$. These disjoint failures have total probability at most δ, but as $t \downarrow 0$ their probabilities converge to $\delta + \mu_r(B) > \delta$. This is impossible.

Take a countable dense subset of the interior slice $P \cdot r = c$. Outside a null set, $U(Q)$ dominates $Q \cdot r$ for every point in that subset and hence its supremum, namely $T(P, Q, c)$. For the equality of suprema, take a maximizer in the closed cube, first increase its prior mean to c if necessary, and then mix it with $c\mathbf{1}$ to enter the interior. A finite union over k completes the result. Every augmented data law is absolutely continuous with respect to μ. No uncountable union over posterior directions is taken. □

For full-support P, extend the right side of (139) at $K_P = 0$ by $T(P, Q, 0) = 0$, which is a trivial lower bound. Denote the resulting scalar endpoint by $c_N^-(0) = 0$ and $c_N^-(k) = c_{N,\delta}(k-1)$ for $k \geq 1$. Then

$$c_N^-(k) \geq k/N - 1/N - h_\delta^-(N). \tag{140}$$

K.3 Geometry, sharp KL/TV, and a finite-sample upper construction

Lemma 54 (Balanced posterior geometry)**.** *For every $Q \in \Delta_M$,*

$$\gamma_P(Q) = \min_{s \in \mathbb{R}} \sum_h |Q_h - sP_h|, \quad \textit{a minimizer lies in } [0, L_Q], \tag{141}$$

$$Q \cdot v \leq \gamma_P(Q) + L_Q(P \cdot v)_+ \quad (v \in [-1,1]^M). \tag{142}$$

Moreover,

$$\Gamma(P, K) = \max_{z \in \{0,1\}^M} \Psi_z(P, K) = \max_{r \in [0,1]^M, Q \in \mathcal{Q}} \{T(P, Q, P \cdot r) - Q \cdot r\}. \tag{143}$$

$0 \leq \Gamma \leq \min\{1, \sqrt{2K}\}$ *and* $1 \leq L(P, K) \leq 1/p_{\min}$.

Proof. Finite-dimensional linear programming duality for the cube with one equality gives (141). More concretely, the right side is the weighted absolute-deviation objective $\sum_h P_h |Q_h/P_h - s|$, whose minimizers are weighted medians of nonnegative numbers in $[0, L_Q]$. Thus $Q \cdot v = (Q - sP) \cdot v + sP \cdot v \leq \gamma_P(Q) + L_Q(P \cdot v)_+$.

If $x, r \in [0,1]^M$ and $P \cdot x \leq P \cdot r$, increase coordinates of x until equality holds, without reducing $Q \cdot x$. The contrast $v = x - r$ is then balanced, so its objective is at most $\gamma_P(Q)$. Conversely, for any balanced v, let $z_h = 1$ when $v_h < 0$ and $z_h = 0$ otherwise; then $x = z + v \in [0,1]^M$, $P \cdot x = P \cdot z$, and $Q \cdot (x - z) = Q \cdot v$. This proves the identities. Finally $Q \cdot v \leq 1$ and, on $P \cdot v = 0$, $Q \cdot v = (Q - P) \cdot v \leq \|Q - P\|_1 \leq \sqrt{2\,\mathrm{KL}(Q\|P)}$. Since $P \in \mathcal{Q}$, $L \geq 1$. □

Lemma 55 (Classical sharp KL-to-TV reduction)**.** *For any two laws D, E on a finite alphabet,*

$$\mathrm{TV}(D, E) \leq \nu(\mathrm{KL}(D\|E)).$$

The bound is attained by binary laws. The function ν is continuous on $[0, \infty]$ and

$$1 - e^{-\lambda} \leq \nu(\lambda) \leq \min\{1, \sqrt{\lambda/2}\}, \qquad \nu(\lambda) \sim \sqrt{\lambda/2} \quad (\lambda \downarrow 0). \tag{144}$$

For finite λ, $\nu(\lambda) < 1$.

Proof. For $B = \{z : E(z) > D(z)\}$, put $u = D(B)$ and $v = E(B)$. Then $v - u = \mathrm{TV}(D, E)$ and the log-sum inequality gives $\mathrm{kl}(u, v) \leq \mathrm{KL}(D\|E)$. Binary laws give equality. Compactness and lower semicontinuity give existence of the maximum and upper semicontinuity. Mixing v towards u makes any positive-divergence feasible pair strictly feasible with arbitrarily small objective loss,

proving lower semicontinuity at positive finite λ. At zero use $\mathrm{kl}(u,v) \geq 2(u-v)^2$, obtained by differentiating twice with respect to u. The pair $u = 0, v = 1 - e^{-\lambda}$ gives the first lower bound and continuity at infinity. For the small-λ lower estimate take $u = 1/2$ and $v = (1 + \sqrt{1 - e^{-2\lambda}})/2$. Its divergence is λ and its variation is asymptotic to $\sqrt{\lambda/2}$. A variation of one would require $(u, v) = (0, 1)$, of infinite divergence. □

The sharp function in Lemma 55 is classical; see (Fedotov et al., 2003).

K.3.1 The admissible rule

For any $\epsilon > 0$, define

$$\eta = \min\{1/2, g_\delta \epsilon\}, \quad \alpha = \eta\delta, \quad \delta' = (1-\eta)\delta,$$

$$a_N = \frac{J \log(N+1) + \log(1/\alpha)}{N}, \quad c = c_{N,\delta}(K_P), \quad c' = c_{N,\delta'}(K_P), \tag{145}$$

$$\mathcal{C}_N = \{E \in \Delta_J : \mathrm{KL}(\widehat{D}_N \| E) \leq a_N,\ b(E) \leq c'\}, \tag{146}$$

$$B_N(Q) = \sup_{E \in \mathcal{C}_N} Q \cdot r(E), \quad B_N(Q) = 0 \text{ if } \mathcal{C}_N = \varnothing,$$

$$U_N^*(Q) = \max\{B_N(Q), [c - \epsilon - p_{\min}^{-1} \mathrm{TV}(P,Q)]_+\}. \tag{147}$$

The coefficient $p_{\min}^{-1} \geq 1$ suppresses the calibration floor at pure posterior directions when $M \geq 2$. Coefficient one would also suffice for all limiting KL-ball statements.

Proposition 56 (Common coverage, pathwise calibration, and finite global bound)**.** *The rule* (147) *belongs to* $\mathcal{A}_N^{\pm}(\epsilon) \subseteq \mathcal{A}_N^{+}(\epsilon)$ *for every finite* N*. It is continuous in* Q*. Let*

$$e_N(a, \tau) = \sqrt{\frac{\log(1/a)}{2N}} + \sqrt{\frac{\log(1/\tau)}{2N}}. \tag{148}$$

For every D*,*

$$w_{1-\tau}(U_N^*; D) \leq \min\{1, \Gamma(P,K)\nu(a_N) + L(P,K) e_N(\delta', \tau)\}. \tag{149}$$

For exact transport replace $\nu(a_N)$ *by one and* δ' *by* δ*.*

Proof. There are at most $(N+1)^J$ empirical types. The probability of any type F under D is at most $e^{-N\,\mathrm{KL}(F\|D)}$: the multinomial coefficient is at most $e^{NH(F)}$, and multiplying by its likelihood gives this bound. Summing types proves

$$\mathbb{P}_D\{\mathrm{KL}(\widehat{D}_N \| D) > a_N\} \leq \alpha.$$

Together with the CP event of failure at most δ', this puts D in $\mathcal{C}_N$ with probability at least $1 - \alpha - \delta' = 1 - \delta$. Hence B_N, and therefore U_N^*, covers all posteriors on one event. This argument does not require independence of the two events.

At $Q = P$, $B_N(P) \leq c'$, also for the empty set. Lemma 51 gives $c' \leq c + \epsilon$. The floor lies between $c - \epsilon$ and c after clipping at zero. This proves pathwise two-sided calibration. Compact support functions are continuous in Q; for each finite observed type/count pair, the feasible set is fixed, so joint measurability also follows.

For any two laws F, E, write their signed difference as $E - F = t(E_+ - E_-)$, where $t = \mathrm{TV}(F, E)$ and E_+, E_- are probability laws with disjoint supports. The vector $v = r(E_+) - r(E_-) \in [-1, 1]^M$ obeys, by (142),

$$Q \cdot (r(E) - r(F)) \leq \gamma_P(Q)\,\mathrm{TV}(F, E) + L_Q[b(E) - b(F)]_+. \tag{150}$$

This is trivial for $t = 0$. Use $F = \widehat{D}_N$ and $E \in \mathcal{C}_N$, then Lemma 55, to obtain

$$B_N(Q) - Q \cdot \hat{r} \leq \Gamma\nu(a_N) + L[c' - P \cdot \hat{r}]_+.$$

The floor has excess at most $(c - P \cdot \hat{r})_+$, because $p_{\min}^{-1} \geq 1$ and $(P - Q) \cdot \hat{r} \leq \mathrm{TV}(P, Q)$. Since $L \geq 1$ and $c' \geq c$, the same displayed upper bound controls the width.

Conditional on S, Hoeffding gives

$$\mathbb{P}\{K_P/N - P \cdot \hat{r} > \sqrt{\log(1/\tau)/(2N)} \mid S\} \leq \tau.$$

Combine this with $c' \leq K_P/N + h_{\delta'}^{+}(N)$ to prove (149). For transport, apply (142) to $v = x - \hat{r}$ directly, yielding the analogous bound with Γ. □

Finite-sample evaluation and computation. For the fixed task $D_0 = \delta_{(0,1)}$, $P = (1/2, 1/2)$, and $Q_0 = (1, 0)$, the randomized count has the exact law $K_P \sim \mathrm{Bin}(N, 1/2)$. With $\delta = 0.05$ and $\epsilon = 0.01$, enumerating its $N + 1$ values gives the following median upper endpoints:

N	exact transport	full-law rule	split hybrid	preselected atom
100	1	0.224690	0.115710	0.067501
400	1	0.074491	0.033612	0.017320
1024	1	0.033336	0.014170	0.006802
4096	1	0.009781	0.003899	0.001705

The last column uses the atom $z = (0, 1)$ fixed before sampling as in (165); the full-law column uses (145). The split hybrid is (119), with $\eta = \min\{1/2, g_\delta \epsilon\}$; at Q_0 its PAC–Bayes-kl branch is $1 - \exp\{-[\log 2 + \log((N + 1)/(\eta\delta))]/N\}$. These are evaluations of admissible rules, not finite-N minimax values. The pathwise prior error is at most ϵ, and exact enumeration gives a joint-coverage failure probability below δ on this task.

Although $\mathcal{C}_N$ is expressed with $J = 2^M$ law coordinates, a fixed-Q support calculation has a smaller dual in strictly feasible cases. Write $F = \widehat{D}_N$, $v_z = Q \cdot z$, $b_z = P \cdot z$, and $d_z = \mu b_z + \xi - v_z$. For $\lambda > 0$, $\mu \geq 0$, and $d_z \geq 0$ on every z, with $d_z > 0$ whenever $F(z) > 0$, its dual objective is

$$\lambda a_N + \mu c' + \xi + \lambda(\log \lambda - 1) - \lambda \sum_{z : F(z) > 0} F(z) \log d_z.$$

It follows by dualizing the KL, prior-mean, and simplex constraints and maximizing independently over nonnegative law masses. The constraints over unobserved states have the separation formula

$$\xi \geq \max_{z \in \{0,1\}^M} (Q - \mu P) \cdot z = \sum_{h=1}^{M} (Q_h - \mu P_h)_+.$$

Thus a fixed-Q dual evaluation can use the observed support and an $O(M)$ separation check; equality with the primal follows under Slater feasibility. This does not remove the $2^M \log(N + 1)/N$ term from a_N, so computational simplification alone does not make the full-law certificate informative for growing libraries.

K.4 The core converse: saturated cutoffs force a reverse-KL barrier

We first record a budget accounting bound and then prove the matching lower-bound mechanism.

Lemma 57 (Failure-budget accounting)**.** *If A is a measurable prior-failure event under an alternative augmented law P_E, then for fixed Q,*

$$\mathsf{P}_E(\{U(Q) < Q \cdot r(E)\} \cap A^c) \leq \delta - \mathsf{P}_E(A).$$

For $\mathsf{P}_D \ll \mathsf{P}_E$ and $s > 1$, Hölder gives

$$\mathsf{P}_D\{U(Q) < Q \cdot r(E)\} \leq \mathsf{P}_D(A) + e^{(s-1)D_s(\mathsf{P}_D \| \mathsf{P}_E)/s} (\delta - \mathsf{P}_E(A))^{(s-1)/s}.$$

Proof. The two failures on A and on its complement are disjoint and contained in the joint failure event. For the second inequality integrate $d\mathsf{P}_D / d\mathsf{P}_E$ over the latter event and apply Hölder, then add $\mathsf{P}_D(A)$. □

The sharp converse uses the CP cutoff refinement in Lemma 58 and the reverse-KL barrier in Theorem 59. Together with the upper construction in Proposition 56, these yield the matching limits in Theorem 49.

Lemma 58 (Finite cutoff leakage at prescribed alternatives)**.** *Suppose $U \in \mathcal{A}_N^+(\epsilon)$, $0 \leq j < N$, and an alternative law E satisfies*

$$b(E) = c_{N,\delta}(j) + \epsilon \in (0, 1), \qquad Q \cdot r(E) > t.$$

Then

$$\mathbb{P}_{E,V}\{K_P > j,\ U(Q) < t\} \leq N\epsilon. \tag{151}$$

The same bound holds for any measurable subset of that event.

Proof. Perturb E continuously by mixing it with the all-one law, so that its prior risk becomes $b(E) + \xi$. For all sufficiently small $\xi > 0$, its Q risk still exceeds t. On $A = \{K_P \le j\}$, upper calibration and monotonicity of CP give $U(P) \le c_{N,\delta}(j) + \epsilon < b(E) + \xi$. The event in (151) is disjoint from A and is another failure. Its probability under the perturbed law is at most $\delta - b_{N,j}(b(E) + \xi)$. For any fixed measurable augmented event, its probability is a finite polynomial in the data-law masses with fixed bounded conditional coefficients, and hence is continuous. Let $\xi \downarrow 0$ and use $|b'_{N,j}| \le N$. □

Theorem 59 (Calibration-constrained KL indistinguishability)**.** *Fix D with $b = b(D) \in (0,1)$, a full-support alternative E, and a fixed Q. Assume*

$$b(E) \le b, \quad I = \mathrm{KL}(D\|E) < \lambda, \quad \lambda_N \to \lambda \in (0, \infty].$$

For every $t < Q \cdot r(E)$,

$$\sup_{U \in \mathcal{A}_N^+(\epsilon_N)} \mathbb{P}_{D,V}\{U(Q) < t\} \longrightarrow 0. \tag{152}$$

The result also holds under exact calibration, with no finite upper restriction on I. It uses only the upper half of the calibration condition.

Proof. Write $m = b(E) \in (0,1)$ and choose $s > 0$ with $I + 2s < \lambda$; for $\lambda = \infty$ any fixed sufficiently small s works. For $u \in (0,1)$ close to m, let $E[u]$ be E mixed with the all-one law if $u \ge m$, or with the all-zero law if $u < m$, in the unique amount giving prior mean u. Then

$$\mathrm{TV}(E[u], E) \le \frac{|u - m|}{\min(m, 1 - m)}. \tag{153}$$

As $u \to m$ these laws remain uniformly positive on the finite alphabet, their log masses converge uniformly, and their Q risks exceed t eventually.

Case 1: $m < b$. Put $j_N = \lfloor Nm \rfloor$ and $u_N = c_{N,\delta}(j_N) + \epsilon_N$. For large N, $j_N < N$, $u_N \in (0,1)$, and $u_N \to m$ by Lemma 51. Lemma 58 applied to $E_N = E[u_N]$ gives

$$\mathbb{P}_{E_N,V}\{K_P > j_N, U(Q) < t\} \le N\epsilon_N.$$

Under D, $K_P/N \to b > m$, so $\mathbb{P}_D(K_P \le j_N) \to 0$. On the event

$$H_N = \left\{ \frac{1}{N} \sum_i \log \frac{D(Z_i)}{E(Z_i)} \le I + s \right\},$$

whose D-probability tends to one, uniform convergence of the log masses implies, for large N,

$$\frac{D^N(S)}{E_N^N(S)} \le e^{N(I+2s)}.$$

The auxiliary kernel cancels exactly in this likelihood ratio. Therefore the probability in (152) is at most

$$\mathbb{P}_D(K_P \le j_N) + \mathbb{P}_D(H_N^c) + N\epsilon_N e^{N(I+2s)} \longrightarrow 0. \tag{154}$$

All terms on this upper bound are independent of U.

Case 2: $m = b$. Put

$$J_N = \{k \in \{1, \dots, N\} : |k/N - b| \le N^{-1/4}\}, \quad u_{N,k} = c_{N,\delta}(k-1) + \epsilon_N.$$

Uniformly for $k \in J_N$, $u_{N,k} \to b$; thus $E_{N,k} = E[u_{N,k}]$ is well-defined and $Q \cdot r(E_{N,k}) > t$ for all sufficiently large N. With cutoff $j = k - 1$, Lemma 58 implies

$$\mathbb{P}_{E_{N,k},V}\{K_P = k, U(Q) < t\} \le N\epsilon_N.$$

On the same H_N, the likelihood ratios from D to all $E_{N,k}$ are at most $e^{N(I+2s)}$ uniformly. Summing over at most N count values gives

$$\mathbb{P}_{D,V}\{U(Q) < t\} \le \mathbb{P}_D(K_P \notin J_N) + \mathbb{P}_D(H_N^c) + N^2\epsilon_N e^{N(I+2s)} \longrightarrow 0. \tag{155}$$

Here the count term is at most $2e^{-2\sqrt{N}}$ for large N. The log-likelihood summands on the support of D have a finite range, so the second term can also be bounded exponentially by Hoeffding (or by zero if that range is zero). The last term vanishes because $I + 2s < \lambda$.

If $\epsilon_N = 0$, the last terms in (154) and (155) vanish identically; no condition on finite I is needed. All alternatives and likelihood events were chosen independently of U, proving uniformity. Null sets were used only inside individual probability statements and finite sums; no posterior-dependent null sets were interchanged. □

Remark 60 (Why the boundary $b(E) = b(D)$ matters). Using a single cutoff at the limiting prior mean can discard a nonvanishing part of the target count distribution. The separate cutoff $k - 1$ on each typical count layer in Case 2 removes this loss. In particular, the converse has probability tending to one; it is not limited to the $1 - \delta$ lower bound obtained by keeping only counts above a CP quantile. This is what permits matching *every* fixed $\tau \in (0, 1)$.

K.5 Proof of the local and global phase theorem

K.5.1 Local matching

For finite λ, along almost every infinite iid augmented sample,

$$\widehat{D}_N \to D, \qquad c' \to b(D), \qquad a_N \to \lambda.$$

The last assertion follows from $\eta_N = g_\delta \epsilon_N$ eventually. The first two follow from the law of large numbers and the uniform root bound with $\delta' \geq \delta/2$.

Consider a subsequence of the rule (147) on which its width has a positive limiting upper value. The calibration floor has excess tending to zero, since it is bounded by $(c - P \cdot \hat{r})_+$. For the support term, take maximizing pairs (Q_N, E_N) in the compact product $\mathcal{Q} \times \mathcal{C}_N$ and extract a convergent subsequence (Q, E). Joint lower semicontinuity gives $\mathrm{KL}(D\|E) \leq \lambda$ and the mean constraint gives $b(E) \leq b(D)$. The limiting excess is at most $A_\lambda(D; P, K)$. Empty confidence sets contribute zero instead. This proves

$$\limsup_N W_N(U_N^*) \leq A_\lambda(D; P, K) \quad \text{almost surely.} \tag{156}$$

For $\lambda = \infty$ simply omit the KL step; the same argument gives A_∞. For exact transport this argument is even simpler, with the limiting cube constraint $P \cdot x \leq b(D)$.

For the reverse direction, suppose first $b(D) \in (0, 1)$, $\lambda > 0$, and choose $0 \leq t < A_\lambda(D; P, K)$. A maximizing feasible pair can be approximated by a fixed pair (Q, E) having $Q \cdot (r(E) - r(D)) > t$, full-support E, $b(E) \leq b(D)$, and $\mathrm{KL}(D\|E) < \lambda$. To see this for finite λ, first mix E slightly with D to make its divergence strictly below λ, preserving the mean inequality and losing arbitrarily little objective. Then mix slightly with the full-support product-Bernoulli law whose common coordinate mean is $b(D)$. This keeps the mean inequality; continuity on the support of D keeps the strict KL inequality. At infinite λ, the second mixing alone supplies finite divergence.

Choose t' strictly between $Q \cdot r(D) + t$ and $Q \cdot r(E)$. Theorem 59 makes $U(Q) \geq t'$ with probability tending to one, uniformly over all admissible rules. Since $Q \cdot \hat{r} \to Q \cdot r(D)$, this proves (132). Under exact calibration use the exact version of that theorem, or Lemma 53 for a fixed maximizing Q. At $\lambda = 0$, $A_0 = 0$ and a nonnegative width needs no lower argument. If $b(D) = 0$, full support of P forces $D = \delta_{\mathbf{0}}$ and the mean constraint forces zero posterior risk; if $b(D) = 1$, then $D = \delta_{\mathbf{1}}$ and every excess is nonpositive. These cases also have $A_\lambda = 0$.

Uniformity in U is essential for taking its infimum. For any $t < A_\lambda$, the probability of $W_N < t$ is eventually smaller than $1 - \tau$, so every admissible quantile is at least t. For any $t > A_\lambda$, (156) makes the constructed quantile at most t eventually. This proves (130) and (133), including the infimum and quantile order in the task.

K.5.2 Evaluation of the worst distribution

Lemma 61 (Exact factorization of the variational maximum). *For every* $\lambda \in [0, \infty]$,

$$\sup_D A_\lambda(D; P, K) = \Gamma(P, K)\nu(\lambda). \tag{157}$$

Proof. If $\mathrm{KL}(D\|E) \le \lambda$ and $b(E) \le b(D)$, (150) and Lemma 55 give an excess at most $\Gamma\nu(\lambda)$.

For the converse assume $\Gamma > 0$. Choose Q_* attaining Γ and an extreme-point maximizer v_* of its balanced-cube linear program. At least one coordinate of v_* equals 1 or -1 (for $M \ge 2$, all but at most one coordinate is at a cube boundary). Form the binary z and $x = z + v_*$ as in Lemma 54. Let $R_0 = \delta_z$ and take R_1 to be any law with risk vector x, for example the product Bernoulli law with means x. A coordinate with $|v_{*,h}| = 1$ is deterministically opposite to z_h under R_1, hence R_0 and R_1 have disjoint supports. They have equal prior risk and

$$Q_* \cdot (r(R_1) - r(R_0)) = \Gamma.$$

For a maximizing pair (u, v) in (128), put

$$D = (1-u)R_0 + uR_1, \qquad E = (1-v)R_0 + vR_1.$$

Then $b(D) = b(E)$, $\mathrm{KL}(D\|E) = \mathrm{kl}(u, v)$ by disjointness, and the posterior risk difference is $(v-u)\Gamma$. This proves the finite-λ result. At infinity use $u = 0, v = 1$; no divergence constraint remains. The case $\Gamma = 0$ is immediate. □

The global upper limit follows *uniformly over* D from (149), since $a_N \to \lambda$ and ν is continuous. The global lower limit follows by fixing a law arbitrarily close to the supremum in (157) and applying the uniform-in-U local lower bound. There is no exchange of a nonuniform limit with $\sup_D$. The same argument and the transport bound prove the exact limit (134). This finishes Theorem 49.

Corollary 62 (What the global minimax problem does and does not degenerate to)**.** *For $M \ge 2$ and $K > 0$, $\Gamma(P, K) > 0$. Let*

$$p_* = \max\{P(A) : \varnothing \ne A \subseteq \{1, \dots, M\},\ P(A) \le 1/2\}.$$

Then

$$\Gamma(P, K) = 1 \quad \Longleftrightarrow \quad K \ge \log(1/p_*). \tag{158}$$

Consequently exact-calibration global minimax width tends to the vacuous value one precisely in this region. Outside it, it tends to a strictly smaller constant. For finite positive λ, the approximate minimax limit is $\Gamma\nu(\lambda) < 1$, not identically one. If $K = 0$ or $M = 1$, it is zero.

Proof. A small nonzero perturbation of P lies in any positive-radius KL ball and has a nonzero balanced contrast. For $\Gamma = 1$, compactness gives Q, v with $Q \cdot v = 1$, so $v_h = 1$ on $A = \mathrm{supp}(Q)$. Balance and $v_h \ge -1$ require $P(A) \le 1/2$. The KL chain rule gives $\mathrm{KL}(Q\|P) = \log(1/P(A)) + \mathrm{KL}(Q\|P(\cdot \mid A)) \ge \log(1/p_*)$. Conversely take $Q = P(\cdot \mid A)$ for a maximizing subset and choose $v = 1$ on A, $v = -P(A)/(1-P(A))$ elsewhere. This is balanced and has objective one. The remaining claims use Lemma 55. □

Remark 63 (Sequences without a limiting exponent). The global formula also yields

$$\mathfrak{R}^{\mathrm{glob}}_{N,+}(\epsilon_N) - \Gamma(P, K)\nu(\lambda_N) \to 0$$

for every positive $\epsilon_N \to 0$, without assuming λ_N converges. Otherwise a subsequence with a nonzero discrepancy has a further subsequence converging in the compactified exponent interval $[0, \infty]$, contradicting Theorem 49. The deterministic-task analogue replaces $\Gamma\nu$ by $\Psi_z(1 - e^{-\lambda_N})$. This proves the stated necessary-and-sufficient subexponential criterion without a monotonicity assumption on the tolerance.

K.6 Explicit deterministic-task bounds and the nonuniform phase

K.6.1 The local deterministic geometry

If $D = \delta_z$, then $\mathrm{KL}(D\|E) = -\log E(z)$. For $h = 1 - e^{-\lambda} > 0$, the constraint is equivalent to writing

$$E = (1-h)\delta_z + hR$$

for some probability law R. Its prior constraint is $P \cdot r(R) \le P \cdot z$. Every vector in $[0,1]^M$ occurs as the risk vector of some law. Thus taking the support over R and then over Q proves (135). At $h = 0$ the sole feasible law is δ_z.

Proposition 64 (Finite deterministic-task lower bound with probability tending to one)**.** *Fix a nonconstant z, a fixed Q, and*

$$p = P \cdot z \in (0,1), \quad m = \min(p, 1-p), \quad C = T(P, Q, p) - Q \cdot z > 0.$$

Choose $u \in (0,1)$ and $0 < t < uC$, and set

$$\rho = \frac{m(uC - t)}{4}.$$

For every N such that

$$\epsilon + 1/N + h_\delta(N) \le \rho, \tag{159}$$

every $U \in \mathcal{A}_N^+(\epsilon)$ satisfies

$$\mathbb{P}_{\delta_z, V}\{U(Q) - Q \cdot z \ge t\} \ge 1 - 2e^{-2N\rho^2} - \frac{N^2\epsilon}{(1-u)^N}. \tag{160}$$

In particular the same lower bound holds for $\mathbb{P}(W_N \ge t)$ when $Q \in \mathcal{Q}$. If its right side is greater than τ, the corresponding $(1-\tau)$-quantile is at least t.

Proof. Choose x_* attaining $T(P, Q, p)$ with $P \cdot x_* = p$ (increase unused coordinates if necessary). For k with $|k/N - p| \le \rho$, let $m_k = c_{N,\delta}(k-1) + \epsilon$. Here $k \ge 1$, since $\rho \le m/4$. The CP bound and (159) imply $|m_k - p| \le 2\rho$. Adjust x_* towards $\mathbf{0}$ or $\mathbf{1}$ to obtain x_k with

$$P \cdot x_k = p + (m_k - p)/u, \qquad \|x_k - x_*\|_\infty \le |m_k - p|/(um).$$

This is possible because $2\rho/u \le m/2$. Let $E_k = (1-u)\delta_z + uR_k$, where $r(R_k) = x_k$. Then $b(E_k) = m_k$, $E_k(z) \ge 1 - u$, and

$$Q \cdot r(E_k) - Q \cdot z \ge uC - 2\rho/m = (uC + t)/2 > t.$$

Use Lemma 58 with cutoff $k-1$ on the subset

$$\{S = z^N, K_P = k, U(Q) < Q \cdot z + t\}.$$

Conditional on $S = z^N$, the auxiliary law is the same under E_k and δ_z, so its target probability is at most $N\epsilon/(1-u)^N$. Summing at most N count layers and using $K_P \sim \mathrm{Bin}(N, p)$ gives (160). The quantile assertion follows because all $w < t$ have CDF less than $1 - \tau$. □

For $\lambda_N \to \lambda > 0$ and $t < (1 - e^{-\lambda})C$, choose $t/C < u < 1 - e^{-\lambda}$. Both error terms in (160) vanish. This also gives a direct finite-sample proof of the deterministic-task lower phase.

K.6.2 Two classifiers

Write $a = \pi$, $b = 1 - \pi$, and $Q = (q, 1-q)$. A two-coordinate knapsack calculation gives

$$T(P, Q, b) - Q \cdot (0,1) = \frac{(q-a)_+}{\max(a,b)}. \tag{161}$$

Indeed, if $q \le a$, the bad coordinate has the larger likelihood ratio and fills the entire budget b, giving zero excess. If $q > a$, fill the good coordinate first; for $b \le a$ the excess is $(q-a)/a$, and for $b > a$ it is $(q-a)/b$.

Let q_- and q_+ be the endpoints of the compact binary-KL interval

$$\{q \in [0,1] : \mathrm{kl}(q, a) \le K\}.$$

Then

$$\Psi_{(0,1)}(P, K) = \frac{q_+ - a}{\max(a,b)}, \tag{162}$$

$$\Gamma(P, K) = \frac{\max\{q_+ - a, a - q_-\}}{\max(a,b)}. \tag{163}$$

The first is the local KL-ball factor; the second is the worst-distribution factor. The point direction Q_0 belongs to the ball if and only if $K \geq \log(1/\pi)$. In that case $\Psi_{(0,1)} = \ell(\pi)$, but the global factor can still be larger.

For the explicit rule (147), on D_0 let $h_N = 1 - e^{-a_N}$. Its confidence set has $E(0,1) \geq 1 - h_N$, so the pure-direction support is exactly

$$U_N^*(Q_0) = \min\left\{h_N, \left[\frac{c' - (1-h_N)b}{a}\right]_+\right\}. \tag{164}$$

The calibration floor vanishes because $p_{\min}^{-1}\,\mathrm{TV}(P, Q_0) = b/\min(a,b) \geq 1$. If $c' < (1-h_N)b$, the confidence set is empty and the displayed value is zero; otherwise the maximum is attained by taking the second risk equal to $1 - h_N$. Since $c' \to b$, (164) converges in probability to

$$\min\{1, b/a\}(1 - e^{-\lambda}).$$

Together with Proposition 64, this proves (136) for arbitrary fixed τ, with matching probability limits. Notice the multiplicative prior factor when $\pi > 1/2$. The generally larger expression $\min\{\ell(\pi), 1 - e^{-\lambda}\}$ is *not* the optimum.

If only this fixed deterministic task is evaluated, the $J\log(N+1)$ term in the construction can be removed: before sampling fix $z = (0,1)$, form a one-sided CP lower bound $\underline{p}_z$ for the probability of observing z at failure probability α, and replace the type constraint by $E(z) \geq \underline{p}_z$. This is valid for every D on the same finite alphabet. On $S = z^N$, $\underline{p}_z = \alpha^{1/N}$, so (164) holds with

$$a_N^{\mathrm{atom}} = \frac{\log(1/(\eta\delta))}{N}. \tag{165}$$

In particular $U_N(Q_0) \leq 1 - e^{-a_N^{\mathrm{atom}}}$ *for every realization of* V. Thus a subexponential tolerance yields vanishing certificates, with the finite upper bound $O_\delta((\log(1/\epsilon_N)+1)/N)$ on this zero-risk direction. Appendix M gives a sharper construction and a matching logarithmic-rate converse for the complete local width when $P = (1/2, 1/2)$ and $K = \log 2$.

For exact calibration, Lemma 53 and the transport construction give

$$\min\{1, c_N^-(K_P)/\pi\} \leq U_N(Q_0), \qquad U_{\mathrm{tr}}(Q_0) = \min\{1, c_{N,\delta}(K_P)/\pi\}.$$

Since $K_P/N \to 1 - \pi$, the lower benchmark and the transport upper both tend to $\ell(\pi)$, proving the exact branch of (136). This is an optimal benchmark; an arbitrary admissible rule may be larger when $\ell(\pi) < 1$.

K.7 Scope

The deterministic-task factor $1 - e^{-\lambda}$ is local; the worst-law frontier is $\Gamma(P,K)\nu(\lambda)$ from Theorem 49. These results fix finite M, positive P, K, δ, and τ before the limit, require validity under every binary-vector law, and impose one-sided pathwise upper approximation to original-confidence CP; the matching construction meets the stronger two-sided condition. Appendix L strengthens the positive finite-exponent branch to uniform local and global optimality over growing finite libraries and changing positive priors, at fixed K and the stated fixed confidence levels. Its logarithmic mixture rule replaces the 2^M law variables of the method-of-types construction used here. Exact finite-N approximate-calibration minimax values, general-law second-order corrections, and sharp subexponential rates beyond the fixed task of Appendix M remain open here. Proposition 48 remains a simpler feasible bound. The uniform theorem does not cover varying confidence levels or joint vanishing-exponent limits.

L A Uniform Local Frontier for Growing Finite Libraries

This section removes the fixed-library restriction at a fixed positive calibration exponent. We retain the full binary alphabet, the randomized prior count, and the local width criterion of Appendix K. Coordinate losses may be arbitrarily dependent. Throughout this section $\delta = 1/20$ and $\tau = 1/10$. The prior has positive coordinates but need not have a uniform positive lower bound. Supremums

over M, P, D below range over all finite M, all such priors, and all laws on $\{0,1\}^M$. For Borel competitors, posterior supremums are interpreted on the completed probability space; they are upper semianalytic and hence universally measurable. The continuous constructed rule has a Borel width.

Theorem 65 (Uniform local calibration frontier)**.** *Fix $0 < \lambda < \infty$ and $0 \le K < \infty$, and set $\epsilon_N = e^{-\lambda N}$. For both the one-sided class and its two-sided subclass,*

$$\sup_{M,P,D} \left|\mathfrak{R}^{\mathrm{loc}}_{N,\star}(D;\epsilon_N) - A_\lambda(D;P,K)\right| \longrightarrow 0, \qquad \star \in \{+,\pm\}. \tag{166}$$

There is one explicit family U_N, independent of D, K, τ, with finite-sample simultaneous coverage and pathwise two-sided error at most $\epsilon_N/4$, such that, for every $a > 0$,

$$\sup_{M,P,D} \Pr_{D,V}\{|W_N(U_N) - A_\lambda(D;P,K)| > a\} \longrightarrow 0. \tag{167}$$

Consequently the global criterion satisfies

$$\sup_{M,P} \left|\mathfrak{R}^{\mathrm{glob}}_{N,\star}(\epsilon_N) - \Gamma(P,K)\nu(\lambda)\right| \longrightarrow 0. \tag{168}$$

Thus M, P, D may change arbitrarily with N, subject to finiteness and positive prior coordinates at each N. The convergence is additive and first-order, at fixed λ, K; it does not assert a relative error when the frontier vanishes or a joint limit with $\lambda_N \to 0$. Under a uniform prior, a point posterior has KL $\log M$, so growing-library point-mass ERM need not fall within this fixed budget.

L.1 A COMMON LOGARITHMIC TEST FOR MIXTURES

Write $\widehat{D}f = N^{-1}\sum_i f(Z_i)$ and $Df = \mathbb{E}_D f(Z)$. The following test retains the logarithm of a mixture, rather than replacing it by a mixture of coordinate logarithms.

Lemma 66 (Logarithmic mixture bounds)**.** *Let X_j be a finite family of $[0,1]$-valued functions with a positive prior Π. For a probability vector R, write $X_R = \sum_j R_j X_j$. Fix an integer $m \ge 1$, a finite positive grid $\mathcal{T}$ of cardinality J, and $\alpha \in (0,1)$. With probability at least $1-\alpha$, simultaneously for every R and $t \in \mathcal{T}$,*

$$\widehat{D}\log(t + X_R) - \log D(t + X_R) \le a_N(R,t) := \frac{m\,\mathrm{KL}(R\|\Pi) + \log(J/\alpha)}{N} + \frac{1}{8mt^2}. \tag{169}$$

For any $s > 0$ and $\rho \in (0,1)$, another event of probability at least $1-\rho$ satisfies, for all the same R, t,

$$|(\widehat{D} - D)\log(t + X_R)| \le \frac{m\,\mathrm{KL}(R\|\Pi) + \log(2J/\rho)}{Ns} + \frac{s}{8}\log^2(1 + 1/t) + \frac{1}{8mt^2}. \tag{170}$$

Proof. For a tuple $H = (H_1, \dots, H_m)$, put $X_H = m^{-1}\sum_l X_{H_l}$. At every fixed H, t, $e_{H,t} = \prod_i (t + X_H(Z_i))/D(t + X_H)$ has expectation one. Mixing under Π^m, applying Markov's inequality, and taking a union over t gives $\mathbb{E}_{\Pi^m} e_{H,t} \le J/\alpha$ for every grid point on one event. Change of measure with R^m then gives

$$\mathbb{E}_{R^m}\{\widehat{D}\log(t + X_H) - \log D(t + X_H)\} \le \{m\,\mathrm{KL}(R\|\Pi) + \log(J/\alpha)\}/N$$

simultaneously for all R. For each z, the tuple average has mean $X_R(z)$ and variance at most $1/(4m)$. Taylor's theorem therefore gives

$$0 \le \log(t + X_R(z)) - \mathbb{E}_{R^m}\log(t + X_H(z)) \le 1/(8mt^2).$$

The population Jensen gap is nonnegative and enters with the favorable sign, proving (169). For (170), apply the two-sided Hoeffding exponential bound to each fixed-tuple function, whose range length is at most $\log(1 + 1/t)$, and repeat change of measure. The difference between the empirical and population Jensen gaps is at most $1/(8mt^2)$. No independence between the coordinate features is used. □

The tuples are a proof device; neither sampling nor enumerating them is part of evaluating the rule.

Use $2M$ features $X_{h,+}(z) = 1 - z_h$, $X_{h,-}(z) = z_h$, with $\Pi_{h,+} = \Pi_{h,-} = P_h/2$. For $\beta \geq 0$ and $Q \in \Delta_M$, let

$$d = Q - \beta P, \quad \gamma = \|d\|_1, \quad a_+ = \sum_h (d_h)_+, \qquad R_{h,+} = \frac{(d_h)_+}{\gamma}, \quad R_{h,-} = \frac{(-d_h)_+}{\gamma} \tag{171}$$

when $\gamma > 0$. Then $\gamma X_R(z) = a_+ - Qz + \beta Pz$.

Lemma 67 (Contrast entropy). *For $0 \leq \beta \leq B$,*

$$\gamma \,\mathrm{KL}(R\|\Pi) \leq \mathrm{KL}(Q\|P) + C_B, \qquad C_B = (1+B)\log(2(1+B)) + 2/e. \tag{172}$$

Proof. Set $t_h = Q_h/P_h$ and $a_h = |t_h - \beta| \leq t_h + \beta$. For $f(x) = x \log x$, $0 \leq a \leq b$ implies $f(a) \leq f(b) + 1/e$. Convexity of KL applied to $(Q + \beta P)/(1+\beta)$ yields

$$\sum_h P_h(t_h + \beta)\log(t_h + \beta) \leq \mathrm{KL}(Q\|P) + (1+\beta)\log(1+\beta).$$

Now $\gamma \,\mathrm{KL}(R\|\Pi) = \sum_h P_h a_h \log a_h - \gamma \log \gamma + \gamma \log 2$. Use $-\gamma \log \gamma \leq 1/e$ and $\gamma \leq 1 + \beta$. □

L.2 An explicit finite-sample endpoint

For $0 < \epsilon \leq 1$, set $\alpha = \epsilon/80$, $\delta' = \delta - \alpha$, $c = \mathrm{CP}_{N,\delta}(K_P)$, and $c' = \mathrm{CP}_{N,\delta'}(K_P)$. The CP sensitivity bound (138) gives $c \leq c' \leq c + \epsilon/4$, since $\delta' \geq 3\delta/4$ and $\min\{\log(1/\delta), \frac{1}{2}\log(2/\delta)\} > 1$. Fix the grids before sampling:

$$\begin{aligned} L &= \max\{2, \lceil \log_2(N+1) \rceil\}, \qquad m = \lceil \sqrt{N} \rceil, \qquad \mathcal{B}_N = \{j/L : 0 \leq j \leq L^2\}, \\ \mathcal{T}_N &= \{L^{-1}(1+L^{-1})^j : 0 \leq j \leq j_N\}, \qquad j_N = \min\{j : L^{-1}(1+L^{-1})^j \geq L\}. \end{aligned} \tag{173}$$

Here $J = |\mathcal{T}_N| = O(L \log L)$ and $t \leq L + 1$. For each β, t and $\gamma > 0$, define

$$B_{\beta,t}(Q) = \beta c' + a_+ + \gamma t - \gamma \exp\{\widehat{D} \log(t + X_R) - a_N(R,t)\}. \tag{174}$$

The case $\gamma = 0$ means $\beta = 1, Q = P$; define that branch as c'. Writing $[x]_{[0,1]} = \min\{1, \max\{0, x\}\}$, set

$$\begin{aligned} V_N(Q) &= \left[\min\left\{ T(P,Q,c'), \min_{\beta \in \mathcal{B}_N, t \in \mathcal{T}_N} B_{\beta,t}(Q) \right\} \right]_{[0,1]}, \\ U_N(Q) &= \max\{V_N(Q), [c - \epsilon/4 - \mathrm{TV}(P,Q)]_+\}. \end{aligned} \tag{175}$$

Proposition 68 (Coverage and pathwise calibration). *For every finite N, M, positive P, and $0 < \epsilon \leq 1$, the rule (175) has simultaneous posterior coverage at level $1 - \delta$ under every binary-vector law. It is continuous in Q, jointly measurable, and satisfies $c - \epsilon/4 \leq U_N(P) \leq c + \epsilon/4$ on every sample.*

Proof. Intersect the event in (169) with $\{Pr(D) \leq c'\}$. Their failure probabilities sum to $\alpha + \delta' = \delta$. On the intersection, simultaneously for all Q, β, t,

$$Qr(D) = \beta Pr(D) + a_+ + \gamma t - \gamma D(t + X_R) \leq B_{\beta,t}(Q), \qquad Qr(D) \leq T(P,Q,c').$$

Thus taking their minimum is valid on the same event. Clipping and adding the lower floor preserve upper coverage. At $Q = P$, transport equals c' and the floor gives the lower calibration bound. All branches are continuous where $\gamma > 0$. As $\gamma \to 0$, the subtracted exponential term is between zero and $\gamma(t+1)$, so the branch tends to c'. Finite minima, maxima, and transport preserve continuity. The finite loss alphabet and measurable count give joint measurability. □

L.3 The oracle dual and a uniform upper bound

For $b = Pr(D) \in (0,1)$ and fixed Q, put $g_\beta(z) = (Q - \beta P)z$ and $h_\beta = \max_{z \in \{0,1\}^M} g_\beta(z) = \sum_h (Q_h - \beta P_h)_+$. The reverse-KL support function has the dual representation

$$H_{\lambda,D}(Q) := \max_{\substack{\mathrm{KL}(D\|E) \leq \lambda \\ Pr(E) \leq b}} Qr(E) = \inf_{\beta \geq 0,\, \eta > h_\beta} \{\beta b + \eta - e^{-\lambda} \exp D \log(\eta - g_\beta)\}. \tag{176}$$

This is the reverse-KL dual used in distributionally robust optimization (Van Parys et al., 2021), with an additional prior-mean constraint. For completeness, mixing D slightly with the zero vector, then adding sufficiently little full-support mass, gives a strictly feasible law. Strong convex duality applies. With multipliers $\beta, \kappa \geq 0$ and normalization multiplier η, maximizing over atom masses gives

$$\beta b + \eta + \kappa\{\lambda + \log \kappa - 1 - D \log(\eta - g_\beta)\}.$$

Minimizing over κ gives $\kappa = e^{-\lambda} \exp D \log(\eta - g_\beta)$ and (176). Boundary dual values are limits of points with $\eta > h_\beta$; attainment in this open domain is not needed. Crucially, h_β uses the full cube, including unobserved atoms.

Fix $b_0 \in (0, 1/2)$ and restrict temporarily to $b \in [b_0, 1 - b_0]$. Let

$$u = 1 - e^{-\lambda/2}, \quad B = \frac{2}{ub_0}, \quad H = 1 + \frac{4e^\lambda}{\lambda}. \tag{177}$$

Every strict dual point with objective at most $H_{\lambda,D}(Q) + \xi$, $0 < \xi \leq 1$, satisfies $\beta \leq B$ and $0 < \eta \leq H$. Indeed, $E_s = (1 - u)D + u\delta_0$ has KL at most $\lambda/2$ and prior slack ub. The dual objective is at least $Qr(E_s) + \beta ub + \kappa\lambda/2$. Since it is at most two, $\beta \leq B$ and $\kappa \leq 4/\lambda$. If $\eta > 1$, then $\kappa \geq e^{-\lambda}(\eta - 1)$ because $g_\beta \leq 1$; otherwise the bound on η is immediate.

Here are explicit errors for the upper bound. Put $\bar{B} = B + 1$, $G = B + 2$, $\rho_N = (N + 1)^{-2}$, $s_N = N^{-1/4}$, and

$$\begin{aligned}
H_{R,N} &= (K + C_{\bar{B}})L/(2H), \\
r_N &= \frac{mH_{R,N} + \log(80J)}{N} + \frac{L^2}{8m}, \\
d_N &= \frac{mH_{R,N} + \log(2J/\rho_N)}{Ns_N} + \frac{s_N}{8} \log^2(1 + L) + \frac{L^2}{8m}, \\
v_N &= \sqrt{\frac{\log(80/3)}{2N}} + \sqrt{\frac{\log(2/\rho_N)}{2N}}, \\
h_N &= \sqrt{\frac{K + \log((N + 1)/\rho_N)}{2N}}, \\
e_N &= \frac{2H + 2G + 4}{L} + \bar{B}v_N + h_N + G(L + 2)(r_N + d_N).
\end{aligned} \tag{178}$$

The binomial CP root bound and Hoeffding give $|c' - b| \leq v_N$ except on probability ρ_N. The usual binary PAC–Bayes moment bound, $\mathbb{E}e^{N \,\mathrm{kl}(\hat{r}_h, r_h)} \leq N + 1$, gives $|Q\hat{r} - Qr(D)| \leq h_N$ simultaneously on the KL ball, except on probability ρ_N. This moment inequality also holds at risks zero and one, where the moment is one. Intersect these events with (170), at failure probability ρ_N.

We claim that whenever $L \geq \max\{2, B + 1\}$,

$$\Pr\{W_N(U_N) > A_\lambda(D; P, K) + e_N\} \leq 3\rho_N. \tag{179}$$

On the preceding intersection, take any Q in the KL ball and a strict dual point within $1/L$ of its infimum. Round β_* upward to $\beta_g \in \mathcal{B}_N$, with error at most $1/L$. If $\gamma_g \leq 2H/L$, transport itself gives

$$T(P, Q, c') - Q\hat{r} \leq 2H/L + \bar{B}v_N + h_N.$$

Otherwise put $t_* = (\eta_* - a_g)/\gamma_g$. Upward rounding of β decreases a_g and g_β, so $0 < t_* < L/2$. Round $\max\{t_*, 1/L\}$ upward on $\mathcal{T}_N$, obtaining t_g, and set $\eta_g = a_g + \gamma_g t_g$. Then

$$0 \leq \eta_g - \eta_* \leq (H + 2G)/L, \qquad \eta_g - g_{\beta_g}(z) \geq \eta_* - g_{\beta_*}(z) > 0.$$

Monotonicity of the geometric mean shows that the population dual value at this grid point is at most $H_{\lambda,D}(Q) + (H + 2G + 2)/L$. This argument does not require a derivative bound near zero dual slack. Lemma 67 gives $\mathrm{KL}(R_g \| \Pi) \leq H_{R,N}$. Thus $a_N(R_g, t_g) \leq \lambda + r_N$ and the log empirical error is at most d_N. Since $\exp D \log(t_g + X_R) \leq L + 2$, the empirical branch exceeds its population dual value by at most $\bar{B}v_N + G(L + 2)(r_N + d_N)$, using $1 - e^{-x} \leq x$. Adding the empirical-risk error proves the claim for V_N. The floor in (175) exceeds $Q\hat{r}$ by at most $[c - P\hat{r}]_+ \leq v_N + h_N$, so the same bound holds for U_N. All choices of Q use the same three events.

Every constant in (179) is independent of M, the smallest prior coordinate, and the smallest positive atom mass. At fixed λ, K, b_0, $e_N \to 0$; the displayed, nonoptimal error is

$$O_{\lambda,K,b_0}\big((\log N)^{-1} + (\log N)^2 N^{-1/4} + (\log N)^3 N^{-1/2}\big).$$

L.4 A LOWER BOUND UNIFORM OVER ALL ADMISSIBLE RULES

Two facts make the likelihood argument uniform in the alphabet. First, if U obeys one-sided calibration, $j < N$, and $Pr(E) = c_{N,\delta}(j) + \epsilon \in (0,1)$, then, for fixed Q and $a \le Qr(E)$,

$$\Pr_{E,V}\{K_P > j,\ U(Q) < a\} \le N\epsilon. \tag{180}$$

To see this, mix E with an arbitrarily small positive mass at the all-one vector. On $K_P \le j$ the prior then fails coverage. The event displayed in (180) is disjoint and also forces failure at Q. Joint coverage bounds their total probability by δ. Letting the mixture mass decrease to zero yields $\delta - B_{N,j}(c_{N,\delta}(j) + \epsilon) \le N\epsilon$, since $|B'_{N,j}| \le N$. Event probabilities are continuous in the law on a finite alphabet, even for merely measurable rules on the augmented sample. No lower calibration inequality is used.

Second, if $E_\theta = (1-\theta)E_0 + \theta D$, $0 < \theta < 1$, and $\ell(Z) = \log(D(Z)/E_\theta(Z))$ under D, then

$$\ell \le \log(1/\theta), \quad \Pr_D\{\ell < -t\} \le e^{-t}, \quad \mathbb{E}_D\ell^2 \le C_\theta := \log^2(1/\theta) + 2. \tag{181}$$

The negative-tail bound follows by Markov from $\mathbb{E}_D(E_\theta/D) \le 1$; integration bounds its second moment by two. Consequently the empirical log likelihood exceeds its mean by $s > 0$ with probability at most $C_\theta/(Ns^2)$, without an atom-mass condition.

Fix $\theta \in (0,1)$ and $b \in [b_0, 1-b_0]$. Let $t_N = N^{-1/4}$, $s = \theta\lambda/4$, and

$$\begin{aligned} h_\delta(N) &= \max\left\{\sqrt{\frac{\log(1/\delta)}{2N}}, \sqrt{\frac{\log(1/(1-\delta))}{2N}}\right\}, \\ \rho_N^{\mathrm{low}} &= \frac{t_N + 2/N + h_\delta(N) + e^{-\lambda N}}{\theta b_0}, \\ p_N^{\mathrm{low}} &= 3e^{-2Nt_N^2} + \frac{C_\theta}{Ns^2} + N^2 e^{-\theta\lambda N/2}. \end{aligned} \tag{182}$$

For all sufficiently large N, uniformly over every one-sided admissible rule U,

$$\Pr_{D,V}\{W_N(U) < (1-\theta)A_\lambda(D;P,K) - \rho_N^{\mathrm{low}} - t_N\} \le p_N^{\mathrm{low}}. \tag{183}$$

Here it suffices that $t_N < b_0/2$, $\rho_N^{\mathrm{low}} < 1$, and $-\log(1-\rho_N^{\mathrm{low}}) \le \theta\lambda/4$.

For the proof, the case $A_\lambda = 0$ is immediate. Otherwise choose fixed Q, E_0 attaining A_λ; compactness and lower semicontinuity of KL ensure attainment. Put $E_\theta = (1-\theta)E_0 + \theta D$ and $m_\theta = Pr(E_\theta)$. Then

$$\mathrm{KL}(D\|E_\theta) \le (1-\theta)\lambda, \quad \theta b_0 \le m_\theta \le b \le 1-b_0, \quad Qr(E_\theta) - Qr(D) = (1-\theta)A_\lambda.$$

On a typical count layer $|k/N - b| \le t_N$, take

$$j(k) = \min\{k-1, \lfloor Nm_\theta \rfloor\}. \tag{184}$$

This choice handles alternatives with strictly smaller prior risk: $0 \le j < N$, $k > j$, and $|j/N - m_\theta| \le t_N + 1/N$. The CP root bounds imply $u_j = c_{N,\delta}(j) + \epsilon_N \in (0,1)$ and $|u_j - m_\theta| \le t_N + 2/N + h_\delta(N) + \epsilon_N$ for large N. Mix E_θ with the zero or one vector to obtain a law E_k with prior mean exactly u_j. Its mixture mass is at most ρ_N^{low}, so $E_k \ge (1-\rho_N^{\mathrm{low}})E_\theta$ and $Qr(E_k) \ge Qr(E_\theta) - \rho_N^{\mathrm{low}}$. Equation (180) bounds the failure event in layer k by $N\epsilon_N$ under E_k.

On the common log-likelihood event from (181), its probability under D is at most its probability under E_k times

$$\exp\{N[(1-\theta)\lambda + s - \log(1-\rho_N^{\mathrm{low}})]\}.$$

The auxiliary uniform kernel is identical and cancels from the likelihood ratio. Summing over at most N typical layers gives $N^2 e^{-\theta\lambda N/2}$. The atypical-count probability is at most $2e^{-2Nt_N^2}$. Finally, $\Pr\{Q\hat{r} > Qr(D) + t_N\} \le e^{-2Nt_N^2}$ for this fixed Q. These bounds prove (183). The probability estimates and large-N threshold depend only on λ, b_0, θ, not on M, P, D, U.

L.5 Boundary means and completion of the proof

Define $J_K(x) = \max\{v \in [x,1] : \mathrm{kl}(v,x) \le K\}$, with $J_K(0) = 0$. For any vector $x_h \in [0,1]$, data processing through the kernel $h \mapsto \mathrm{Bernoulli}(x_h)$ gives

$$Qx \le J_K(Px), \qquad J_K(x) \le \min\{1, (K + \log 2)/\log(1/x)\} \longrightarrow 0 \quad (x \downarrow 0). \tag{185}$$

The function J_K is nondecreasing. The floor in (175) is at most c', while $T(P,Q,c') \ge c'$ by the constant feasible vector. Hence $U_N(Q) \le T(P,Q,c')$ and $W_N(U_N) \le J_K(c')$. If $b \le b_0$, all oracle alternatives have prior mean at most b_0; therefore $A_\lambda \le J_K(b_0)$ and, outside probability ρ_N, $W_N(U_N) \le J_K(\min\{1, b_0 + v_N\})$. If $b \ge 1 - b_0$, then $A_\lambda \le J_K(b_0)$, and for any $[0,1]$-valued rule,

$$W_N(U) \le \sup_{\mathrm{KL}(Q\|P)\le K} Q(1-\hat{r}) \le J_K(P(1-\hat{r})) \le J_K(\min\{1, b_0 + h_N\})$$

outside probability ρ_N.

Proof of Theorem 65. Given $a > 0$, first choose b_0 with $J_K(2b_0) < a/4$. For large N, $v_N, h_N \le b_0$, so the preceding boundary estimates make both the oracle and the constructed width smaller than $a/4$ except on a uniformly vanishing event. In the interior, (179) bounds the upper deviation uniformly. Choose $\theta < a/4$ in (183) and then N large enough that $\rho_N^{\mathrm{low}} + t_N < a/4$. Since $A_\lambda \le 1$, this bounds the lower deviation uniformly as well, proving (167).

The upper-deviation probability eventually falls below τ, while the lower-deviation probability in (183) eventually falls below $1 - \tau$ uniformly over all one-sided competitors. Thus the corresponding quantiles, and their infima over rules, obey the same additive sandwich. At boundary means use non-negativity of width and the small oracle value. The constructed rule belongs to the two-sided subclass, so it proves the upper bound for both classes. This establishes (166). For the global claim, the single rule supplies the uniform upper bound, whereas $\inf_U \sup_D q_D(W_N(U)) \ge \sup_D \inf_U q_D(W_N(U))$ supplies the lower bound. The identity $\sup_D A_\lambda(D; P, K) = \Gamma(P,K)\nu(\lambda)$ from Appendix K gives (168). □

L.6 Joint information and fixed-posterior computation

An explicit example distinguishes local from worst-law sharpness. Let $P = (1/2, 1/2)$ and $K = \log 2$. On the same full four-vector alphabet, consider $D_s = (\delta_{00} + \delta_{11})/2$ and $D_o = (\delta_{10} + \delta_{01})/2$. Their coordinate risks agree, but

$$A_\lambda(D_s; P, K) = \frac{1 - e^{-\lambda}}{2}, \qquad A_\lambda(D_o; P, K) = \frac{\sqrt{1 - e^{-2\lambda}}}{2}. \tag{186}$$

For D_s, write the alternative masses at $00, 11, 10, 01$ as a, b, c, d. The constraints give $ab \ge e^{-2\lambda}/4$ and $b \le a$, hence $a \ge e^{-\lambda}/2$ and each coordinate risk is at most $1 - a$. Equality is attained by $a = b = e^{-\lambda}/2$, $c = 1 - e^{-\lambda}$, $d = 0$, and a point-mass posterior. For D_o, coordinate data processing gives $\mathrm{kl}(1/2, r_h(E)) \le \lambda$. Its upper root is $1/2 + \sqrt{1 - e^{-2\lambda}}/2$, attained by putting that mass at 10 and the remainder at 01. Subtracting the common true risk $1/2$ proves (186). The rule in (175) attains both limits without knowing which law generated the data.

For fixed Q, the dual needs the law of (PZ, QZ) and the known full-cube support h_β, rather than all 2^M atom probabilities. Evaluating (175) uses the two sample projections, one contrast entropy per β, and scalar logarithms over the grid. Its real-arithmetic cost is

$$O\big(NM + M \log M + ML^2 + NL^3 \log L\big),$$

in addition to two scalar CP inversions. The terms account respectively for projections, transport sorting, contrast entropies, and branch evaluation. The all-posterior validity event allows evaluating a data-dependent Q; this complexity statement does not provide a global optimizer over Q. For rigorous numerical upper endpoints, logarithms and CP inversions must be enclosed with outward rounding; ordinary optimizer output is not such an enclosure.

The full-cube support is essential. For D_s, $Q = (1, 0)$ and $\beta = 1$, all observed vectors have $g_\beta = 0$, but $h_\beta = 1/2$ on the full cube. Replacing it by the empirical maximum would let $\eta \downarrow 0$ in (176), incorrectly producing zero width. Thus retaining joint sample information does not license identifying hypotheses merely because their observed loss columns coincide.

M A Sharp Local Rate for One-Sided Calibration

The first-order limit in Theorem 6 gives zero for subexponential precision. The fixed task below admits a sharper local description without strengthening the one-sided calibration requirement. We use the admissible rule class and quantile convention of Section 5.2: $S \sim D^N$, the auxiliary uniforms have a law-independent kernel, and each $[0,1]$-valued rule is jointly measurable in the augmented sample and posterior. Fix $z_0 = (0,1)$, $D_0 = \delta_{z_0}$, $P = (1/2, 1/2)$, $Q_0 = (1,0)$, and $K = \log 2$. The posterior KL ball is the whole two-point simplex. Write $b_{N,j}(p) = \Pr\{\mathrm{Bin}(N,p) \le j\}$ and $c_{N,a}(j) = \mathrm{CP}_{N,a}(j)$, with $c_{N,a}(N) = 1$. We abbreviate $R_N^+(\epsilon) = \mathfrak{R}_{N,+}^{\mathrm{loc}}(D_0;\epsilon)$ and use $R_N^{\pm}$ for the same objective under the two-sided condition $|U_N(P) - c_{N,\delta}(K_P)| \le \epsilon$. Let $k_{1-\tau} = q_{1-\tau}(\mathrm{Bin}(N,1/2))$.

$$D_0 = \delta_{(0,1)}, \qquad P = (1/2,1/2), \qquad K = \log 2. \tag{187}$$

M.1 A one-sided finite-sample converse

Theorem 69 (A calibrated single-cutoff converse)**.** *Fix N, $\delta, \tau \in (0,1)$, $\epsilon > 0$, and an integer $0 \le j < N$. Set*

$$c_j = c_{N,\delta}(j), \quad m_j = c_j + \epsilon, \quad a_j = b_{N,j}(1/2), \quad s_j = \delta - b_{N,j}(m_j), \quad d_j = 2(m_j - 1/2).$$

Suppose $1/2 \le m_j < 1$. For every $U \in \mathcal{A}_N^+(\epsilon)$ and every $d_j \le u < 1$,

$$\mathbb{P}_{D_0,V}\{W_N(U) < u\} \le a_j + \frac{s_j}{(1-u)^N}. \tag{188}$$

If $a_j < 1 - \tau$, $0 < s_j < 1 - \tau - a_j$, and

$$\ell_j = 1 - \left(\frac{s_j}{1 - \tau - a_j}\right)^{1/N} > d_j,$$

then $R_N^+(\epsilon) \ge \ell_j$. The same conclusion holds with s_j replaced by any proved upper bound $\bar{s}_j \ge s_j$, provided its displayed conditions hold.

Proof. For $u \in [d_j, 1)$ define the alternative law

$$E_u(0,1) = 1 - u, \quad E_u(1,0) = u - d_j, \quad E_u(1,1) = d_j, \quad E_u(0,0) = 0.$$

These are nonnegative masses summing to one, and

$$P \cdot r(E_u) = m_j, \qquad Q_0 \cdot r(E_u) = u.$$

Let $E_{u,t} = (1-t)E_u + t\delta_{(1,1)}$, $t > 0$. Its prior risk is strictly greater than m_j. On $A = \{K_P \le j\}$, upper calibration alone therefore forces a prior-coverage failure, almost surely under $E_{u,t}$. The event

$$B = \{S = z_0^N,\ K_P > j,\ U_N(Q_0) < u\}$$

is disjoint from A and is another coverage failure under $E_{u,t}$. Distribution-free simultaneous coverage implies

$$\mathbb{P}_{E_{u,t},V}(B) \le \delta - b_{N,j}(P \cdot r(E_{u,t})).$$

Probabilities of fixed augmented events are continuous in the law: they are finite sums of sample probabilities times fixed conditional probabilities over V. Sending $t \downarrow 0$ gives $\mathbb{P}_{E_u,V}(B) \le s_j$.

Conditional on $S = z_0^N$, the auxiliary distribution is exactly the same under E_u and D_0. Since $E_u(z_0)^N = (1-u)^N$,

$$\mathbb{P}_{D_0,V}\{K_P > j, U_N(Q_0) < u\} \le s_j(1-u)^{-N}.$$

Under D_0, $K_P \sim \mathrm{Bin}(N,1/2)$ and $W_N(U) \ge U_N(Q_0)$, because the latter posterior has zero empirical risk. Adding the event $K_P \le j$ proves (188).

For $d_j < u < \ell_j$, its right side is strictly smaller than $1 - \tau$. For each $w < u$, $\mathbb{P}(W_N \le w) \le \mathbb{P}(W_N < u) < 1 - \tau$, hence the quantile is at least u. Send $u \uparrow \ell_j$ and take the infimum over all admissible rules. No lower calibration inequality or continuity in Q is used. □

Evaluating the cutoff leakage. Rather than subtracting two nearly equal CDF values, use

$$s_j = \int_{c_j}^{c_j+\epsilon} f_{N,j}(p)\,dp \le \epsilon \max_{p\in[c_j,c_j+\epsilon]} f_{N,j}(p), \qquad f_{N,j}(p) = N\binom{N-1}{j}p^j(1-p)^{N-j-1}. \tag{189}$$

For $N > 1$, the maximum is at the projection of $j/(N-1)$ onto this interval. The formula can be evaluated in logarithms even for exponentially small tolerances.

Lemma 70 (A coverage-only reference lower bound)**.** *If $\delta + \tau < 1$, then for every $\epsilon > 0$,*

$$R_N^+(\epsilon) \ge 1 - \left(\frac{\delta}{1-\tau}\right)^{1/N}.$$

Proof. Use $E_u = (1-u)\delta_{(0,1)} + u\delta_{(1,0)}$. Under this law Q_0 has risk u, so coverage alone implies

$$(1-u)^N \mathbb{P}_{D_0,V}\{U_N(Q_0) < u\} \le \delta.$$

As $W_N(U) \ge U_N(Q_0)$, quantile inversion proves the claim. This bound is independent of any calibration property. □

M.2 A CALIBRATED UPPER CONSTRUCTION

The elementary perturbation bound in Lemma 51 suffices for the first-order limit. The next lemma gives a uniform $N^{-1/2}$ refinement.

Lemma 71 (Density at an interior quantile of a unimodal law)**.** *Let a continuous distribution have a unimodal density f, mean μ, and standard deviation $\sigma > 0$. If both $F(x)$ and $1 - F(x)$ are at least $\theta > 0$, then*

$$f(x) \ge \frac{\theta^{3/2}}{2(1+\sqrt{2})\sigma}.$$

Proof. Chebyshev's inequality and the two tail masses imply $|x-\mu| \le \sigma/\sqrt{\theta}$. Put $y_\pm = \mu \pm \sigma\sqrt{2/\theta}$. If x lies to the right of a mode, the interval $[x, y_+]$ has probability at least $\theta/2$, length at most $(1+\sqrt{2})\sigma/\sqrt{\theta}$, and density at most $f(x)$. These facts give the inequality. If x lies to the left of a mode, use $[y_-, x]$ instead. Flat modes and a mode at a support endpoint cause no change. □

Lemma 72 (Uniform CP confidence sensitivity)**.** *Set*

$$\theta_\delta = \min\{\delta/2, 1-\delta\}, \qquad \kappa_\delta = \frac{\theta_\delta^{3/2}}{1+\sqrt{2}}.$$

For every $N \ge 1$, $0 < \alpha \le \delta/2$, and $0 \le k \le N$,

$$0 \le c_{N,\delta-\alpha}(k) - c_{N,\delta}(k) \le \frac{\alpha}{\kappa_\delta\sqrt{N+2}}. \tag{190}$$

Proof. For $k < N$, $-b'_{N,k}$ is the density of $\mathrm{Beta}(k+1, N-k)$. It is unimodal and has variance at most $1/[4(N+2)]$. Between the two roots in (190), its survival probability is in $[\delta/2, \delta]$. Lemma 71 therefore bounds the density below by $\kappa_\delta\sqrt{N+2}$. Integrating that lower bound over the root interval gives

$$\alpha \ge \kappa_\delta\sqrt{N+2}\{c_{N,\delta-\alpha}(k) - c_{N,\delta}(k)\}.$$

For $k = N$ both endpoints are one. □

Proposition 73 (Preselected-atom one-sided construction)**.** *Fix the atom $z_0 = (0,1)$ before sampling. For each N and $\epsilon > 0$, put*

$$\alpha_N = \min\{\delta/2, \kappa_\delta\epsilon\sqrt{N+2}\}, \qquad \delta'_N = \delta - \alpha_N.$$

Let $T_0 = \sum_i \mathbf{1}\{Z_i = z_0\}$ and let $l_{N,\alpha_N}(T_0)$ be the one-sided CP lower bound for $D(z_0)$ with failure probability α_N. Define

$$\mathcal{C}_N^{\mathrm{atom}} = \{E : E(z_0) \ge l_{N,\alpha_N}(T_0),\ P \cdot r(E) \le c_{N,\delta'_N}(K_P)\},$$

and $U_N^{\mathrm{atom}}(Q) = \sup_{E \in \mathcal{C}_N^{\mathrm{atom}}} Q \cdot r(E)$*, with value zero for an empty set. This rule is in* $\mathcal{A}_N^+(\epsilon)$ *and is valid under every law on* $\{0,1\}^2$*. On* D_0*,*

$$W_N(U_N^{\mathrm{atom}}) \le 1 - \alpha_N^{1/N} \quad \textit{on every augmented sample}, \qquad R_N^+(\epsilon) \le 1 - \alpha_N^{1/N}. \tag{191}$$

Proof. The atom-probability and prior-CP confidence events have failure probabilities at most α_N and δ'_N. Their intersection places the true law in $\mathcal{C}_N^{\mathrm{atom}}$ and covers every posterior, with total failure probability at most δ. No independence of these two events is needed. At $Q = P$, the support is at most $c_{N,\delta'_N}(K_P) \le c_{N,\delta}(K_P) + \epsilon$ by Lemma 72. The set is compact and depends on a finite pair of counts; its support function is continuous in Q, including the empty-set convention, which also proves joint measurability.

Under D_0, $T_0 = N$ and $l_{N,\alpha_N}(N) = \alpha_N^{1/N}$. Every feasible law has at most $1 - \alpha_N^{1/N}$ probability away from z_0. For every posterior, its loss belongs to $[0,1]$, so

$$Q \cdot r(E) - Q \cdot z_0 \le 1 - \alpha_N^{1/N}.$$

Maximize over E, Q and use the empty-set convention to conclude. □

Remark 74. The choice of z_0 is fixed before sampling, and the resulting rule covers every data law. Its width statement concerns the local objective at D_0.

M.3 A LOCAL LOGARITHMIC LAW

The cutoff converse and the preselected-atom construction give a matching logarithmic scale on the fixed task. All constants below may depend on δ, τ but not on ϵ.

Lemma 75 (Uniform central CP cutoff geometry)**.** *Fix* $\delta \in (0,1)$*. There are constants* $C_\delta < \infty$ *and* N_δ *such that for every* $N \ge N_\delta$ *and* $0 < \epsilon \le 1/16$*, the index*

$$j_N = \min\{0 \le j < N : c_{N,\delta}(j) + \epsilon \ge 1/2\}$$

exists, is interior, and satisfies

$$\begin{aligned} 0 \le c_{N,\delta}(j_N) + \epsilon - 1/2 &\le C_\delta/N, \\ b_{N,j_N}(1/2) &\le \delta + C_\delta/\sqrt{N}, \\ \delta - b_{N,j_N}(c_{N,\delta}(j_N) + \epsilon) &\le C_\delta \epsilon \sqrt{N}. \end{aligned}$$

If additionally $\epsilon_N \sqrt{N} \to 0$*, then* $b_{N,j_N}(1/2) \to \delta$*.*

Proof. Lemma 51 gives $c_{N,\delta}(j) = j/N + O_\delta(N^{-1/2})$, uniformly in j. Thus for the stated tolerance range the crossing index is interior and the adjacent roots lie in a fixed compact subset of $(0,1)$, for all sufficiently large N.

We record an elementary mass estimate. Fourier inversion and $|1-p+pe^{it}|^2 = 1-4p(1-p)\sin^2(t/2)$ give

$$\max_k \mathbb{P}\{\mathrm{Bin}(n,p) = k\} \le \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-2np(1-p)\sin^2(t/2)}\, dt \le \frac{\sqrt{\pi}}{2\sqrt{2np(1-p)}}.$$

In the last step use $\sin(|t|/2) \ge |t|/\pi$ and extend the Gaussian integral to the real line. Hence central binomial masses are $O(n^{-1/2})$, and $f_{N,j}(p) = N\mathbb{P}\{\mathrm{Bin}(N-1,p) = j\} = O(\sqrt{N})$ uniformly for central p.

Write $c_- = c_{N,\delta}(j_N - 1)$ and $c_+ = c_{N,\delta}(j_N)$. The binomial recurrence gives

$$b_{N,j_N}(c_-) = \delta + \mathbb{P}\{\mathrm{Bin}(N, c_-) = j_N\} = \delta + O(N^{-1/2}), \qquad b_{N,j_N}(c_+) = \delta.$$

On $[c_-, c_+]$, both the beta CDF and its survival probability are bounded away from zero. Lemma 71, together with its beta variance bound, therefore gives $f_{N,j_N} \ge c_\delta \sqrt{N}$ on this interval. Integrating shows $c_+ - c_- = O_\delta(N^{-1})$.

By the definition of the crossing, $c_- < 1/2 - \epsilon \le c_+$. This proves the first conclusion. In particular $c_+ \le 1/2 + C_\delta/N$. Since the central derivative is $O(\sqrt{N})$ and the binomial CDF is decreasing in its probability parameter,

$$b_{N,j_N}(1/2) \le \delta + O_\delta(N^{-1/2}).$$

Integrating the same derivative bound over $[c_+, c_+ + \epsilon]$ proves the leakage estimate. Finally, if $\epsilon_N\sqrt{N} \to 0$, then $|c_+ - 1/2| \le \epsilon_N + C_\delta/N = o(N^{-1/2})$, so the difference between $b_{N,j_N}(1/2)$ and $b_{N,j_N}(c_+) = \delta$ tends to zero. □

Theorem 76 (Sharp log-scale characterization on the fixed task)**.** *Assume (187) and $\delta + \tau < 1$. There exist constants $C < \infty$ and N_0, depending only on δ, τ, such that for all $N \ge N_0$ and $0 < \epsilon \le 1/16$,*

$$\left| -N \log\big(1 - R_N^+(\epsilon)\big) - [\log \frac{1}{\epsilon\sqrt{N}}]_+ \right| \le C. \tag{192}$$

In particular, if $\epsilon_N\sqrt{N} \to 0$ and $L_N = \log(1/(\epsilon_N\sqrt{N}))$, then

$$R_N^+(\epsilon_N) = 1 - \exp\left[-\frac{L_N + O_{\delta,\tau}(1)}{N} \right]. \tag{193}$$

The statement concerns the local quantile on this fixed two-classifier task.

Proof. The upper construction gives $R_N^+ \le 1 - \alpha_N^{1/N} < 1$, hence

$$\begin{aligned} -N \log(1 - R_N^+) &\le -\log \alpha_N \\ &= \max\left\{ \log\frac{2}{\delta}, \log\frac{1}{\epsilon\sqrt{N}} - \log\kappa_\delta - \frac{1}{2}\log(1 + 2/N) \right\} \\ &\le [\log \frac{1}{\epsilon\sqrt{N}}]_+ + C_1. \end{aligned}$$

For a lower bound choose the crossing j_N of Lemma 75. For all sufficiently large N,

$$1 - \tau - a_{j_N} \ge b_0 := \frac{1 - \delta - \tau}{2} > 0, \qquad d_{j_N} \le C_2/N, \qquad s_{j_N} \le C_3\epsilon\sqrt{N}.$$

Let $L = \log(1/(\epsilon\sqrt{N}))$ and choose $C_4 \ge 1$ with $C_4 \ge C_3/b_0$. When L exceeds a sufficiently large fixed constant, put

$$\ell = 1 - \exp[-(L - \log C_4)/N].$$

Then $\ell > d_{j_N}$ for all sufficiently large N. To see this uniformly even for very large L, write $x = L - \log C_4$: if $x \le N$, then $N(1 - e^{-x/N}) \ge x/2$; if $x > N$, then $N(1 - e^{-x/N}) \ge (1 - e^{-1})N$.

For every $d_{j_N} < u < \ell$, the CDF bound in Theorem 69 is strictly below $1 - \tau$. Therefore $R_N^+ \ge \ell$ and

$$-N \log(1 - R_N^+) \ge L - \log C_4.$$

When L is below that fixed threshold, nonnegativity of the transformed width already gives the required lower inequality after increasing the constant. This proves (192). Lemma 70 additionally shows $R_N^+ > 0$ in this quantile range. Equation (193) follows directly; no optimizer attainment is assumed. □

Corollary 77 (Schedules in Figure 5)**.** *Under Theorem 76, for fixed $a > 1/2$,*

$$R_N^+(N^{-a}) = (a - 1/2)\frac{\log N}{N} + O_{a,\delta,\tau}(N^{-1}).$$

For $0 < a \le 1/2$, $R_N^+(N^{-a}) = \Theta_{\delta,\tau}(N^{-1})$. Moreover,

$$\begin{aligned} R_N^+(e^{-\sqrt{N}}) &= \frac{1}{\sqrt{N}} - \frac{\log N}{2N} + O_{\delta,\tau}(N^{-1}), \\ R_N^+(e^{-\lambda N}) &= 1 - e^{-\lambda} - \frac{e^{-\lambda}\log N}{2N} + O_{\lambda,\delta,\tau}(N^{-1}), \qquad \lambda > 0. \end{aligned}$$

Proof. Substitute the tolerance schedules into (192), then expand the exponential. For $a \le 1/2$, combine the constant transformed upper bound with Lemma 70. □

M.4 Comparison with two-sided calibration

Proposition 78 (Two-sided polynomial-tolerance optimum)**.** *Under* (187)*,* $\delta+\tau<1$*, and* $\epsilon_N = N^{-a}$ *for fixed* $a > 1/2$*, let* z_b *denote the standard normal* b*-quantile. Then*

$$R_N^{\pm}(\epsilon_N) = \frac{z_{1-\delta} + z_{1-\tau}}{2\sqrt{N}} + o(N^{-1/2}), \tag{194}$$

whereas $R_N^{+}(\epsilon_N) = (a - 1/2)(\log N)/N + O(N^{-1})$.

Proof. For every two-sided rule, its width is at least $[c_{N,\delta}(K_P) - \epsilon_N - 1/2]_+$ because $Q = P$ belongs to the ball. Therefore $[c_{N,\delta}(k_{1-\tau}) - \epsilon_N - 1/2]_+$ is a valid lower bound for $R_N^{\pm}$.

Use the two-sided full-law rule (147). On D_0, its complete width is given by (198). Both are nondecreasing functions of K_P. At $k_{1-\tau}$, the support term is h_N for all sufficiently large N, because $c_{N,\delta'}(k_{1-\tau}) > 1/2$. For polynomial tolerance, $h_N = O(\log N/N) = o(N^{-1/2})$.

The binomial CLT and inversion at fixed confidence give

$$k_{1-\tau} = N/2 + z_{1-\tau}\sqrt{N}/2 + o(\sqrt{N}), \qquad c_{N,\delta}(k_{1-\tau}) = 1/2 + \frac{z_{1-\delta} + z_{1-\tau}}{2\sqrt{N}} + o(N^{-1/2}).$$

The sum of the two normal quantiles is positive exactly when $\delta + \tau < 1$. Thus, eventually, the prior-floor quantile exceeds h_N. The two-sided construction then attains the lower bound $[c_{N,\delta}(k_{1-\tau}) - \epsilon_N - 1/2]_+$ itself. Subtracting $\epsilon_N = o(N^{-1/2})$ proves (194). This argument concerns the two-sided infimum only. □

At $\delta = .05$ and $\tau = .1$, the coefficient in (194) is approximately 1.4632.

N Deterministic Bounded-Mean Calibration

Unlike the randomized binomial count, the observation here is the original fractional prior loss $X_i^P = P \cdot \ell(\cdot, Z_i)$.

For $x = (x_1, \ldots, x_N) \in [0,1]^N$, let $(D_0, \ldots, D_N) \sim \mathrm{Dirichlet}(1, \ldots, 1)$ independently of the data and define the deterministic endpoint

$$G_{N,\delta}(x) = Q_{1-\delta}\left(D_0 + \sum_{i=1}^{N} D_i x_i\right),$$

where the quantile is over the Dirichlet weights. This is the one-sided Gaffke–Learned-Miller–Thomas upper endpoint. Its distribution-free coverage for independent bounded observations follows by inversion of the Gaffke test (Vlassis & Thomas, 2026; Ming et al., 2026); for binary observations it equals the original-confidence CP upper endpoint.

Proposition 79 (Deterministic prior recovery by transport)**.** *Fix a finite hypothesis library, a prior* P *before sampling, bounded losses,* $N \geq 1$*, and* $\delta \in (0,1)$*. With* $X_i^P = \sum_h P_h \ell(h, Z_i)$*, put*

$$U_N^G(P, Q, S) = T\big(P, Q, G_{N,\delta}(X_{1:N}^P)\big).$$

For every data law, one event of probability at least $1 - \delta$ *satisfies* $L(Q) \leq U_N^G(P, Q, S)$ *simultaneously for every posterior* Q*. Pathwise,* $U_N^G(P, P, S) = G_{N,\delta}(X_{1:N}^P)$*, and*

$$0 \leq U_N^G(P, Q, S) - G_{N,\delta}(X_{1:N}^P) \leq \mathrm{TV}(P, Q) \leq \sqrt{\mathrm{KL}(Q\|P)/2}.$$

No auxiliary evaluation randomness is needed.

Proof. The Gaffke coverage event is $E = \{L(P) \leq G_{N,\delta}(X_{1:N}^P)\}$ and has probability at least $1 - \delta$. On E, the true vector of hypothesis risks is feasible in the definition of T, proving simultaneous posterior coverage. For $c = G_{N,\delta}(X_{1:N}^P)$, the constant vector $c\mathbf{1}$ is feasible, while every feasible vector has P-mean at most c; hence $T(P, P, c) = c$. For every feasible $r \in [0,1]^M$, $Q \cdot r \leq P \cdot r + \mathrm{TV}(P, Q) \leq c + \mathrm{TV}(P, Q)$. The same constant vector gives $T(P, Q, c) \geq c$, and Pinsker's inequality gives the final bound. □

The lower bound has a different mechanism from the layerwise CP previous-count inequality. Let $\mathcal{Z} \subset [0,1]^M$ be finite, D have full support on $\mathcal{Z}$, and write $x_P(z) = P \cdot z$ and $x_Q(z) = Q \cdot z$. A direction h is invisible to the prior scalar observation if its D-weighted sum is zero on each level set of x_P.

Theorem 80 (A fiber-invisible posterior direction)**.** *Fix $\delta \in (0,1)$, a prior P, and a posterior Q with $\mathrm{KL}(Q\|P) < \infty$. Suppose $\sigma_P^2 = \mathrm{Var}_D(x_P(Z)) > 0$ and there is a function $h : \mathcal{Z} \to \mathbb{R}$ such that*

$$\sum_{z : x_P(z) = x} D(z)h(z) = 0 \quad \text{for every } x \in x_P(\mathcal{Z}), \qquad \kappa := \mathbb{E}_D[h(Z)x_Q(Z)] > 0. \tag{195}$$

For each N, let $U_N(S, Q') \in [0,1]$ be a measurable rule independent of the unknown data law. Assume that under every law E on $\mathcal{Z}$, one event of probability at least $1-\delta$ covers $L_E(Q') \le U_N(S, Q')$ simultaneously for all Q'. Assume also, almost surely under every such E, that

$$U_N(S, P) \le G_{N,\delta}(x_P(Z_1), \dots, x_P(Z_N)) + \epsilon_N, \qquad \epsilon_N \ge 0, \quad \sqrt{N}\epsilon_N \to 0.$$

Then, with $W_N(Q) = [U_N(S,Q) - \widehat{L}_S(Q)]_+$, for every $\tau \in (0, 1-\delta)$,

$$\sqrt{N}\, q_{1-\tau, D}(W_N(Q)) \longrightarrow \infty.$$

Proof. Fix $t > 0$. Since the fiber condition implies $\mathbb{E}_D h = 0$, for all sufficiently large N the local alternative

$$D_{N,t}(z) = D(z)\left(1 + \frac{th(z)}{\sqrt{N}}\right)$$

is a probability law with positive masses. Summing this law over each level set in (195) shows that $x_P(Z)$ has exactly the same distribution under $D_{N,t}$ and D. Consequently $L_{D_{N,t}}(P) = L_D(P) =: \mu_P$, whereas

$$L_{D_{N,t}}(Q) = L_D(Q) + \frac{t\kappa}{\sqrt{N}}.$$

Write $G_N = G_{N,\delta}(x_P(Z_1), \dots, x_P(Z_N))$. The one-sided endpoint expansion of Ming et al. (2026, Theorem 6.2) and the ordinary central limit theorem, using $\sigma_P > 0$, give

$$\sqrt{N}(G_N - \mu_P) \;\Rightarrow\; N(\sigma_P z_{1-\delta}, \sigma_P^2) \qquad \text{under } D.$$

The whole scalar sample has the same law under $D_{N,t}$, so this limit holds there too. Hence the event $A_N = \{G_N + \epsilon_N < \mu_P\}$ obeys

$$\Pr_{D_{N,t}}(A_N) = \Pr_D(A_N) \longrightarrow \delta. \tag{196}$$

On A_N, the one-sided calibration condition makes $U_N(P) < \mu_P$, so the prior fails coverage under $D_{N,t}$. The disjoint event

$$B_N = \{U_N(Q) < L_{D_{N,t}}(Q)\} \cap A_N^c$$

is another simultaneous-coverage failure. Distribution-free coverage under the N-dependent law $D_{N,t}$ and (196) imply $\Pr_{D_{N,t}}(B_N) \to 0$.

For fixed t, the likelihood ratio from $D_{N,t}^N$ to D^N has bounded second moment:

$$\mathbb{E}_{D_{N,t}^N}\left(\frac{dD^N}{dD_{N,t}^N}\right)^2 = \left[\sum_{z \in \mathcal{Z}} \frac{D(z)}{1 + th(z)/\sqrt{N}}\right]^N \longrightarrow \exp\{t^2 \mathbb{E}_D h^2\} < \infty.$$

Indeed the linear term vanishes because $\mathbb{E}_D h = 0$, and the next term is $t^2 \mathbb{E}_D h^2 / N + O(N^{-3/2})$. Cauchy–Schwarz therefore gives $\Pr_D(B_N) \to 0$. It follows that

$$\Pr_D\left\{U_N(Q) < L_D(Q) + \frac{t\kappa}{\sqrt{N}}\right\} \le \Pr_D(A_N) + \Pr_D(B_N) \le \delta + o(1). \tag{197}$$

Finally, $Y_N = \sqrt{N}\{\widehat{L}_S(Q) - L_D(Q)\}$ is tight by the central limit theorem. For any fixed $w > 0$,

$$\Pr_D\{\sqrt{N}W_N(Q) < w\} \le \Pr_D\left\{U_N(Q) < L_D(Q) + \frac{t\kappa}{\sqrt{N}}\right\} + \Pr_D\{Y_N > t\kappa - w\}.$$

Choose the fixed t sufficiently large that the limiting upper probability on the right is strictly below $1-\tau$; this is possible because $\tau < 1 - \delta$ and Y_N is tight. Thus the $(1-\tau)$-quantile of $\sqrt{N}W_N(Q)$ is eventually at least w. Since w is arbitrary, the asserted divergence follows. □

Corollary 81 (A fixed low-risk bounded-mean obstruction)**.** *Fix $\delta \in (0,1)$ and $0 < \rho < 1/2$. Let D_ρ assign probabilities $\rho, \rho, 1-2\rho$ to loss vectors $(1,0),(0,1),(0,0)$, respectively. Set $P = (1/2, 1/2)$ and $Q = \delta_{h_0}$. Any rule satisfying the distribution-free simultaneous coverage and one-sided prior-calibration conditions of Theorem 80, with $\epsilon_N = o(N^{-1/2})$, obeys, for every $\tau \in (0, 1-\delta)$,*

$$\sqrt{N}\, q_{1-\tau, D_\rho}\big([U_N(S,Q) - \widehat{L}_S(Q)]_+\big) \longrightarrow \infty.$$

Here $L_{D_\rho}(Q) = \rho$ and $\mathrm{KL}(Q\|P) = \log 2$.

Proof. For the three loss vectors $a = (1,0)$, $b = (0,1)$, and $c = (0,0)$, the prior scalar loss is $x_P(a) = x_P(b) = 1/2$ and $x_P(c) = 0$. Under D_ρ its mean is ρ and variance is $\rho(1/2 - \rho) > 0$. Set $h(a) = 1/\rho$, $h(b) = -1/\rho$, and $h(c) = 0$. The weighted sum of h is zero on both prior-loss fibers, while $\mathbb{E}_{D_\rho}[h(Z)x_Q(Z)] = 1$. Moreover $L_{D_\rho}(Q) = \rho$ and $\mathrm{KL}(Q\|P) = \log 2$. Theorem 80 applies. □

The lower bound concerns a fixed posterior, not the data-dependent ERM posterior, and does not identify a matching bounded-mean frontier.

O Supplementary Numerical Illustrations

These illustrations measure evidence refinement and coordinate allocation under specified protocols. Conditional resampling uses a fixed empirical population; matched-posterior comparisons evaluate the same predictor under each formula. The endpoint optimality calculations and their numerical verification are given separately in Appendix P. These studies illustrate mechanisms, not a general certificate ranking.

O.1 Conditional resampling from real-data pools

We use a stratified split with 600 training and 1197 held-out examples for Digits, and 220 training and 349 held-out examples for Breast Cancer. Digits uses label 8 as positive; Breast Cancer uses malignancy as positive. A logistic model with standardized features, ℓ_2 regularization, and $C = 0.3$ is fitted on the training split. Training predictions determine the baseline threshold, an anchor, and a library of 101 thresholds in $[0.02, 0.995]$. To create a paired improvement task, the baseline is chosen above the training-optimal risk, with target excess 0.02 for Digits and 0.025 for Breast Cancer. The anchor targets one quarter of the baseline's excess training risk. The prior assigns mass 0.9 to the anchor and spreads the remainder uniformly over the other 100 classifiers.

For each sample, residual classifiers are ranked by empirical error. The nine candidate posteriors combine residual group sizes $9, 33, 100$ with anchor weights $0.5, 0.9, 0.97$. The library and candidate-generation rule are fixed before certification; candidate membership and posterior selection depend on the certification sample. All methods use the same candidate family to compare squared and binomial intersection evidence.

The held-out pool is treated as a fixed empirical population. We report all 100 repetitions, each drawing $N = 4000$ observations independently with replacement; sample size selection used the training-side objective. Population risk differences are evaluated on the complete pool. The resulting observations are conditionally i.i.d. This experiment measures variation in certification for the fixed empirical population, rather than uncertainty about an unknown natural population or classifier retraining. Panels (a)–(b) of Figure 3 report unsmoothed empirical CDFs of the upper endpoints. Table 2 gives the numerical summary. Its 95% Wilson intervals for squared and binomial evidence are $[0.502, 0.691]$ and $[0.646, 0.816]$ on Digits, and $[0.404, 0.596]$ and $[0.668, 0.833]$ on Breast Cancer. Binomial evidence gives a smaller bound than squared evidence on every one of the 200 repetitions. Its additional negative decisions number 14 and 26, with none in the opposite direction, as expected from pointwise dominance; the corresponding 95% Wilson intervals for the paired incremental rates are $[0.085, 0.221]$ and $[0.184, 0.354]$.

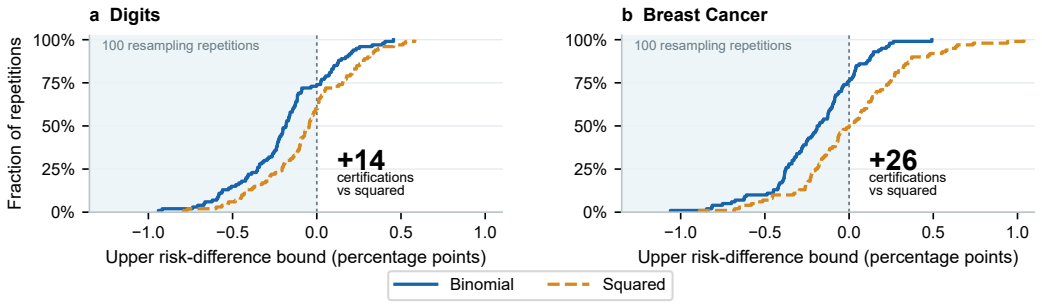


Figure 3: Certification decisions on fixed empirical populations. The curves are empirical CDFs of upper risk-difference bounds; negative values certify improvement. Binomial evidence adds 14 and 26 certifications over squared evidence, illustrating how evidence refinement changes certification decisions.

Table 2: Paired improvement certifications out of 100, with median upper risk-difference bounds in parentheses. Negative bounds certify improvement.

Certificate	Digits	Breast Cancer
Squared evidence	60 (-0.000448)	50 ($+0.000082$)
Binomial evidence	74 (-0.001831)	76 (-0.001834)

O.2 Coordinate allocation in a distinct classifier library

We compare binomial evidence and coordinatewise weighted CP at the same posterior as a genuinely distinct classifier library grows. The data model is $X \sim \mathcal{N}(0, I_{20})$ and $Y = \mathbf{1}\{X_1 \geq 0\}$. Before drawing the certification samples, we fix a master library of 16,384 linear classifiers. We normalize independent Gaussian vectors in the subspace orthogonal to the first coordinate and combine them with that coordinate direction at independent uniform angles in $[0.02, 0.30]$ radians. The library uses distinct directions rather than prediction-equivalent copies.

For each of 30 independent samples of size $N = 2000$, we evaluate every dyadic prefix from $M = 32$ to $M = 16384$ at confidence level $1 - \delta = 0.95$. The prior is uniform on the prefix, and Q is uniform on its empirically best quarter. Simultaneous coverage permits this sample-dependent choice; $\mathrm{KL}(Q\|P) = \log 4$ at every size. Both certificates evaluate this same Q. Coordinatewise weighted CP assigns failure probability δ/M to each classifier, whereas the binomial construction has a KL-controlled learning guarantee.

Figure 4 shows the full scale curve and paired differences. Weighted CP is tighter in all 30 runs at $M = 32$. The median ordering first reverses at $M = 2048$, and binomial evidence is tighter in all 30 runs at $M = 16384$, where the respective median endpoints are 0.03148 and 0.03398. This crossover illustrates the cost of coordinate allocation for a controlled synthetic library and a common posterior.

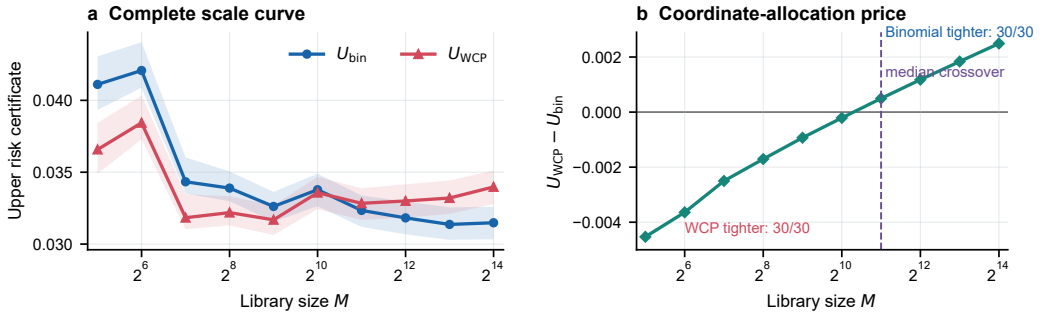


Figure 4: Binomial evidence versus coordinatewise weighted CP at a common posterior. For nested prefixes of distinct Gaussian linear classifiers, Q is the empirically best quarter and $\mathrm{KL}(Q\|P) = \log 4$. Panel (a) gives medians and interquartile ranges over 30 samples; (b) gives the paired difference, with positive values favoring binomial evidence. The full dyadic scale is shown.

O.3 A LIBRARY OF TRAINED DIGITS CLASSIFIERS

We repeat the common-posterior comparison with trained multiclass linear classifiers. In each of 12 stratified splits of the same Digits dataset, 700 observations train the models, 300 form an independent pilot block, 400 form the certification block, and 397 are held out. Bootstraps, feature masks, and regularization strengths generate up to 16,384 classifiers; pilot deduplication and the uniform prior on each nested prefix are fixed before certification. All 16,384 classifiers in each split have distinct prediction vectors on its certification block. These repeated splits do not constitute independent draws of 12 natural populations.

At each library size, Q is uniform on the empirically best quarter of the library, so $\mathrm{KL}(Q\|P) = \log 4$. Table 3 evaluates the displayed certificates at this same sample-dependent Q; entries are medians over the 12 paired splits. The binomial and weighted-CP ordering reverses between the small and large libraries: binomial is tighter in all 12 splits at $M = 2048$ and $M = 16384$, but in none at $M = 16$. This comparison isolates the coordinate-allocation crossover between binomial evidence and weighted CP for a fixed learned predictor.

Table 3: Median upper risk certificates on trained Digits classifiers, all evaluated at the same posterior.

Certificate	$M = 16$	$M = 1024$	$M = 16384$
Squared evidence	.17649	.17802	.17890
Binomial evidence	.12558	.12754	.12870
Weighted CP	.10487	.12673	.13948

O.4 NO-IMPROVEMENT CONTROLS

For each fixed held-out empirical distribution, select its minimum-risk library threshold as baseline before any evaluation resampling. Every candidate posterior is a mixture of that library, so every true risk difference is nonnegative. This defines a conditional null on the fixed empirical population. The original model, library, prior and candidate-generation rule remain fixed.

Each dataset has 200 resamples. The first 100 index sequences are shared with the improvement experiment, so the two conditions must not be pooled as independent trials. Within each condition, the repetitions are independent. Squared intersection, binomial intersection, and weighted CP each give zero false negative decisions out of 200 per dataset. For any one specified method and dataset, the one-sided 95% CP upper bound on its conditional false-decision frequency is $1 - 0.05^{1/200} = 0.01487$. This is not a simultaneous confidence statement across methods or datasets.

The minimum population gap among selected posteriors is 0.00102 on Digits and 0.00478 on Breast Cancer. For an exactly zero gap, we add a conditional control. Write the baseline loss vector on the fixed empirical population as $b \in \{0,1\}^{n_{\mathrm{pool}}}$. Starting from b, the candidate changes the first

$s = \min\{\sum_j b_j, n_{\text{pool}} - \sum_j b_j\}$ error indices to zero and the first s correct indices to one. It therefore has exactly the same number of errors as the baseline and nonzero paired disagreement. We use $P = (1/2, 1/2)$, $Q = \delta_{\text{candidate}}$, $N = 4000$, $\delta = 0.05$, and draw each paired count directly from the resulting three-category empirical-population law. The one-sided disagreement rates are 4.76% and 4.30%. In 20,000 resamples per dataset, binomial and square intersection make zero false negative decisions; weighted CP makes 13 and 20 (0.065% and 0.10%). This is a conditional diagnostic on the fixed empirical population.

O.5 Exact enumeration of general-prior transport

The fixed task in Corollary 46 permits an exact finite-sample calculation. Here $P = (1/2, 1/2)$, the loss vector is always $(0, 1)$, and the unique ERM posterior is $Q = \delta_1$ on the zero-loss classifier. The prior's empirical risk $\widehat{L}_S(P)$ is identically $1/2$, whereas the randomized count has K_P/N standard deviation $1/(2\sqrt{N})$. We enumerate all $N + 1$ values of $K_P \sim \text{Bin}(N, 1/2)$, compute $c = \text{CP}_{N,0.05}(K_P)$, and use the exact identity $U_{\text{tr}} = \min\{1, 2c\}$.

N	$\text{sd}(K_P/N)$	$\Pr(U_{\text{tr}} = 1)$	$\mathbb{E}U_{\text{tr}}$	universal lower bound
100	.0500	.95569	.99834	.87021
400	.0250	.95059	.99907	.91011
1000	.0158	.95316	.99939	.92477
4000	.0079	.95157	.99968	.93738

The last column is the theorem-wide previous-count guarantee $1 - \delta - \sqrt{2/(\pi N)}$, which applies to every endpoint satisfying the assumptions of Theorem 44; the two preceding columns describe the transport endpoint itself. Enumeration avoids Monte Carlo sampling error: despite zero empirical and population risk for the selected classifier, the expected transport endpoint approaches one.

O.6 Fixed-task cost and calibration-frontier experiments

Figure 5 evaluates the fixed low-risk task of Theorem 1 and the deterministic local case of Theorem 6. All tasks and parameters are fixed before sampling. We compute expectations and quantiles from count laws, with no Monte Carlo error; numerical summation and CP inversion use floating-point arithmetic.

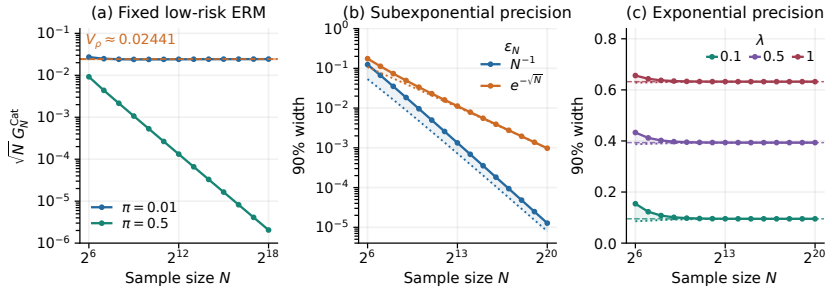


Figure 5: Fixed-task certificates ($\delta = 0.05$). (a) Scaled excess of one pre-data Catoni comparator: risk 0.1, $\pi \in \{0.01, 0.5\}$; orange marks the certified asymmetric limit interval. (b)–(c) 90th-percentile one-sided width on $D = \delta_{(0,1)}$, $P = (1/2, 1/2)$, $K = \log 2$. Solid: upper bounds from a preselected-atom rule; dotted: lower bounds for every admissible rule; shading: brackets. Dashed: limits $1 - e^{-\lambda}$ for $\epsilon_N = e^{-\lambda N}$. Protocols and numerical values: Appendix O.6.

O.6.1 A common formula on the low-risk ERM task

Set $\rho = 0.1$, $\delta = 0.05$, and $\pi \in \{0.01, 0.5\}$ in Theorem 1; ties always select h_0. Write $v = \rho(1 - \rho)$, $\alpha_j = \log(1/(\delta P(h_j)))$, and $\bar{\alpha} = (\alpha_0 + \alpha_1)/2$. For each $N \in \{2^6, 2^7, \ldots, 2^{18}\}$, use the single comparator

$$f_N(q, r) = -b_N q - \log(1 - r + re^{-b_N}), \qquad b_N = \sqrt{\frac{2\bar{\alpha}}{vN}}.$$

Its binomial exponential moment equals one for every r. Its endpoint is $U_N(q, \alpha_j/N) = \min\{1, (1 - e^{-\alpha_j/N - b_N q})/(1 - e^{-b_N})\}$. The same b_N is used for both ERM branches and every sample state. The plotted G_N^{Cat} is the expectation of this endpoint minus $F(q, \alpha_j/N)$, so it bounds the full-class optimum from above.

For $l = 0, \ldots, \lfloor N/2 \rfloor$, let $B_l = \Pr\{\mathrm{Bin}(N, \rho) = l\}$ and $H_l \sim \mathrm{Bin}(N - l, \rho/(1 - \rho))$. The state $q = l/N$ has branch weights $B_l \Pr(H_l \geq l)$ for h_0 and $B_l \Pr(H_l > l)$ for h_1. Summing these weights times the corresponding endpoint excess evaluates the full expectation, including ties.

The construction's scaled limit is $\sqrt{2v\bar{\alpha}} - (\sqrt{2v\alpha_0} + \sqrt{2v\alpha_1})/2$. For $\pi = 0.01$, interval evaluation of the analytic one-plane upper bound and a Gaussian exclusion lower bound give

$$0.0244073995 \leq V_\rho \leq 0.0244173996.$$

For the lower bound, retain only the necessary Gaussian normalizer constraint at $r = \rho$ and express the relaxation at variance $1/4$; conversion back to variance 0.09 multiplies its value by $3/5$. Round gaps upward and thresholds downward, then apply convex tangent lower bounds in the two contact slopes over a covering partition of endpoint allocations. Interval arithmetic verifies infeasibility on a complete covering partition. The relaxation makes this a lower bound for the full Gaussian program. At $N = 2^{18}$, the asymmetric and symmetric scaled construction costs are 0.024371129 and 0.000002054, respectively. The latter illustrates the zero first-order limit without asserting an optimal symmetric decay rate.

O.6.2 One-sided local-width brackets under approximate calibration

Fix $D_0 = \delta_{(0,1)}$, $P = (1/2, 1/2)$, $K = \log 2$, $\delta = 0.05$, and $\tau = 0.1$. The KL ball is the whole posterior simplex. For $N \in \{2^6, \ldots, 2^{20}\}$, use tolerance N^{-1}, $e^{-\sqrt{N}}$, or $e^{-\lambda N}$ for $\lambda \in \{0.1, 0.5, 1\}$. The solid curves in Figure 5(b)–(c) evaluate the one-sided preselected-atom rule of Proposition 73; dotted curves use the maximum of the single-cutoff converse and the elementary coverage lower bound. For comparison, the following calculation uses the full-law rule (147) with $J = 4$; its stronger two-sided calibration is not used as a one-sided lower bound. Let $c(k) = c_{N,\delta}(k)$, $c'(k) = c_{N,\delta'}(k)$, and $h_N = 1 - e^{-a_N}$, with a_N, δ' from (145). The support-function branch minus empirical risk is convex in Q and is maximized at $(1, 0)$. The calibration floor contributes positive excess $[c(k) - \epsilon_N - 1/2]_+$ at $Q = P$. Consequently the complete width is

$$W_N(k) = \max\{\min\{h_N, [2c'(k) - 1 + h_N]_+\},\ [c(k) - \epsilon_N - 1/2]_+\}. \tag{198}$$

Both terms increase in k. Since $K_P \sim \mathrm{Bin}(N, 1/2)$, its 90% quantile $k_{.9}$ satisfies $c'(k_{.9}) > 1/2$, yielding

$$q_{.9,D_0}(W_N) = \max\{h_N, [c(k_{.9}) - \epsilon_N - 1/2]_+\}.$$

This identity evaluates the quantile directly from the binomial count law. We evaluate the radius using $\log(1/\epsilon_N)$ directly, so exponentially small positive tolerances retain their intended exponent.

The floor term is a property of this two-sided construction, not a lower bound for the one-sided admissible class. Indeed, removing the floor from (147) leaves the support rule B_N. The proof of Proposition 56 shows that B_N alone covers every posterior on one event, and $B_N(P) \leq c' \leq c + \epsilon_N$; hence $B_N \in \mathcal{A}_N^+(\epsilon_N)$. On D_0, its complete width satisfies

$$W_N(B_N; k) = \min\{h_N, [2c'(k) - 1 + h_N]_+\} \leq h_N.$$

The 90% quantile equals h_N for the plotted counts. For $\epsilon_N = N^{-1}$, this gives a valid one-sided upper bound of order $\log N/N$, whereas the two-sided construction's prior floor is of order $N^{-1/2}$ on this task.

A finite-sample lower bound for every one-sided-admissible rule follows from the calibrated single-cutoff converse in Theorem 69. For $N = 2^{20}$, $\delta = .05$, and $\tau = .1$, three resulting lower and upper enclosures for the fixed-law local optimum are

Tolerance	Lower	Upper
N^{-1}	0.00000796	0.00001273
$e^{-\sqrt{N}}$	0.00097083	0.00097560
$e^{-0.1N}$	0.09515782	0.09516214

The lower values use a single CP cutoff; the upper values use the preselected-atom rule in Proposition 73. The three displayed brackets were checked with outward interval arithmetic. The global worst-law limit remains $\nu(\lambda)$; these enclosures evaluate only the fixed law D_0.

P Computing the Endpoint Optimum

We distinguish three quantities: the endpoint of an explicit normalized comparator, the optimum over all normalized convex comparators at finite N, and the Gaussian limiting optimum in Theorem 3. Their numerical bounds serve different purposes and have different accuracies.

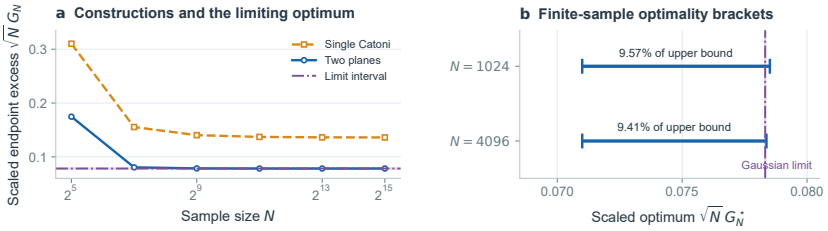


Figure 6: Endpoint cost for $\delta = 0.05$-feasible local gaps $(5/4, 11/4)$. (a) Normalized constructions and the Gaussian limit for target sequences. (b) Finite-sample optimum brackets; percentages are $(U-L)/U$. The targets form a designed state family; at $N = 1024, 4096$, each pair admits a common prior. All quantities are selection excesses.

P.1 Target families and numerical results

All targets have equal weights and $q_{j,N} = 1/2 + a_j/\sqrt{N}$, $p_{j,N} = 1/2 + b_j/\sqrt{N}$, $t_{j,N} = \mathrm{kl}(q_{j,N}, p_{j,N})$. We evaluate $N \in \{32, 128, 512, 2048, 8192, 32768\}$. The moderate and separated families have $b = (0, 0)$ and respectively $a = (-3/4, -3/2)$ and $a = (-5/4, -11/4)$. The equal-gap control has $a = (-3/4, 1/4)$ and $b = (3/20, 23/20)$, so both local gaps equal $9/10$. The reported price is the average right-endpoint excess multiplied by $\sqrt{N}$.

Local gaps	Gaussian optimum V	Construction upper, $N = 32768$
$(3/4, 3/2)$	$[0.0608341225, 0.0608541226]$	0.0608634184
$(5/4, 11/4)$	$[0.0783137482, 0.0783337483]$	0.0783351829
$(9/10, 9/10)$	0 (exact)	$2.683 \cdot 10^{-10}$

The first and third upper bounds use a single Catoni comparator; the second uses two planes.

The separated family is compatible with the standard confidence threshold at $\delta = 0.05$. Indeed, Pinsker's inequality gives $Nt_{j,N} \geq 2d_j^2$, where $(d_1, d_2) = (5/4, 11/4)$, and hence both thresholds exceed $\log 20$. At $N = 1024$ and 4096, the two $q_{j,N}$ are binomial-grid values. Put $\kappa_{j,N} = Nt_{j,N} - \log 20 \geq 0$ and assign common-prior masses $P(h_j) = e^{-\kappa_{j,N}}$; their sum is below one, so the remaining mass can be placed on an auxiliary hypothesis. Then $Q_j = \delta_{h_j}$ has $\mathrm{KL}(Q_j \| P) = \kappa_{j,N}$, and binary losses with the displayed counts realize the two empirical risks. Thus the finite brackets in Figure 6 concern bona fide PAC–Bayes states, while remaining a controlled target-design calculation rather than a fixed-task learning experiment.

P.2 Valid finite-sample constructions

A single Catoni comparator $f(q, r) = -\beta q - \log(1 - r + re^{-\beta})$, $\beta > 0$, has normalizer one. Its endpoint at threshold t is $\min\{1, (1 - e^{-t-\beta q})/(1 - e^{-\beta})\}$. We search for a suitable β and enclose this expression by interval arithmetic; this validates the candidate, not global optimality of its parameter within the Catoni family.

For a two-plane envelope $H(q, r) = \max_i(A_i q + D_i r + C_i)$ with $D_i > 0$, an upper bound $B \geq I_{H,N}$ makes $H - N^{-1} \log B$ normalized. The right endpoint is the minimum of its two affine roots, clipped

at one. The coefficients are interpreted as exact rationals; interval arithmetic bounds the moments and endpoints.

To bound the normalizer on the entire population interval, put $x = \operatorname{logit}(r)$ and write $G(x) = G_{H,N}(r)$. The function $\log G$ is semiconvex with constant

$$K = N\left(\frac{1}{4} + \frac{\max_i |D_i|}{6\sqrt{3}}\right).$$

Indeed, the sigmoid has second derivative bounded in magnitude by $1/(6\sqrt{3})$; maximum and log-sum-exp preserve a common lower curvature bound. The remaining term $-N\log(1+e^x)$ contributes at worst $-N/4$. Consequently, on every interval $[x_l, x_u]$,

$$\log G(x) \le \max\{\log G(x_l), \log G(x_u)\} + \frac{K}{8}(x_u - x_l)^2.$$

A finite partition covers $r \in [10^{-6}, 1-10^{-6}]$. Interval arithmetic encloses each endpoint moment and chord correction on this partition. On the two remaining boundary intervals, the sum of the single-plane moment maxima is a valid upper bound. Each single-plane log moment is concave in r, so its maximum is enclosed using the endpoints and its stationary point.

At each fixed r, the two affine pieces cross at an exactly computed rational count threshold. Exponential tilting reduces each piece to a scalar moment times a binomial tail. The probability at a mode is enclosed using an exact integer binomial coefficient and interval powers; recurrences with outward-rounded ratios generate the tails. Product and positive-sum roundoff bounds and an absolute underflow allowance complete the enclosure. Interval evaluations use at least 50 decimal digits.

P.3 Bounds on the Gaussian optimum

For $q_0 = 1/2$, take a centered Gaussian X of variance $1/4$. A two-plane candidate with zero tangential parameters has normalizer $\mathbb{E}\max_i \exp(\beta_i X + c_i)$. Its maximum over population parameters occurs at $r = 1/2$: the envelope is convex in X, and the Gaussian variance $r(1-r)$ is largest there. Normalizing this candidate yields an upper bound on V. Gaussian integrals are enclosed using the alternating error-function series with a remainder bound, and Mills' inequality in the tails.

For a lower bound, suppose mean endpoint excess is at most g. Partition $E_1 \in [0, 2g]$ into rational cells $[l, u]$ and relax each pair to $y_1 = d_1 + u$, $y_2 = d_2 + 2g - l$. Necessary feasibility is

$$\min_{\beta_1,\beta_2>0} F(\beta; y) \le 1, \qquad F(\beta; y) = \mathbb{E}\max_i \exp\{\beta_i(X - y_i) + 2d_i^2\}. \tag{199}$$

The function is convex in β. Its individual moments confine feasible slopes to $[0, 8y_1] \times [0, 8y_2]$. At a proposed point, interval evaluation of the value and gradient gives an affine lower support, whose minimum over that whole box bounds F from below. If the bound exceeds one, the endpoint cell is excluded; otherwise it is subdivided. The accepted cells cover the complete rational interval. Numerical optimizer termination is not an infeasibility certificate.

At a coincident-plane point, we instead construct a valid zero subgradient. Let $\beta_* = (2d_2^2 - 2d_1^2)/(y_2 - y_1)$, $\mu = \beta_*/4$ and $p = (y_2 - \mu)/(y_2 - y_1)$. Under $N(\mu, 1/4)$, choose $h(X) = a\mathbf{1}\{X \le \mu + z/2\} + b$, where $a = p(\mu - y_1)/[\varphi(z)/2]$ and $b = p - a\Phi(z)$. When intervals verify $0 \le b \le a + b \le 1$, this split satisfies $\mathbb{E}h = p$ and $\mathbb{E}[Xh] = py_1$. Both components of the associated subgradient vanish. The common moment is then a global lower bound. Here φ, Φ are the standard normal density and CDF.

P.4 Finite-sample optimality bounds

Certified finite-sample endpoint brackets. Consider the two equally weighted targets

$$q_{1,N} = \frac{1}{2} - \frac{5}{4\sqrt{N}}, \qquad q_{2,N} = \frac{1}{2} - \frac{11}{4\sqrt{N}}, \qquad p_{1,N} = p_{2,N} = \frac{1}{2}, \qquad t_{i,N} = \mathrm{kl}(q_{i,N}, 1/2).$$

The following computation bounds the finite-sample optimum over the full normalized convex-comparator class, rather than only over a chosen parametric family:

Table 4: Finite-sample bounds $L \le \sqrt{N}\,\mathcal{G}_N^\star \le U$ for the separated two-target instance. Displayed upper bounds are rounded up.

N	Lower L	Upper U	$(U-L)/U$
1024	0.071	0.078511122410	0.09567
4096	0.071	0.078374923822	0.09410

The upper bounds come from explicit two-plane comparators. Their binomial normalizers are bounded uniformly over the population parameter by outward interval arithmetic, after which both right endpoints are inverted with outward rounding.

For completeness, we describe the full-class lower certificate. Fix a candidate scaled mean excess g and suppose that a normalized convex comparator f has endpoint excesses $e_1, e_2 \ge 0$ with $\sqrt{N}(e_1+e_2)/2 \le g$. Partition the possible values of e_1 into rational cells. If $[\ell, u]$ is one such cell, both endpoints are bounded by the relaxed corner

$$u_1 = \frac{1}{2} + u, \qquad u_2 = \frac{1}{2} + \frac{2g}{\sqrt{N}} - \ell.$$

Each relaxed corner is interior and satisfies $u_i > q_{i,N}$. Continuity of f implies $f(q_{i,N}, u_i) \ge t_{i,N}$: a strict reverse inequality would make a larger risk feasible. Take a supporting plane at that point and lower its intercept to attain $t_{i,N}$. It remains below f. Its vertical slope is positive, since the original supporting slope is at least $f(q_{i,N}, u_i)/(u_i - q_{i,N}) > 0$ by $f(q_{i,N}, q_{i,N}) \le 0$. Thus we obtain

$$L_i(x,r) = A_i x + D_i r + C_i \le f(x,r), \qquad L_i(q_{i,N}, u_i) = t_{i,N}, \qquad D_i > 0.$$

Because each L_i is individually normalized, its moment at $r = u_i$ implies

$$A_i q_{i,N} - \log(1 - u_i + u_i e^{A_i}) \ge t_{i,N}. \tag{200}$$

The superlevel set in (200) is a compact interval between two negative roots. Moreover, every rational probe $r > u_i$ supplies the necessary upper bound

$$D_i \le \frac{A_i q_{i,N} - t_{i,N} - \log(1 - r + r e^{A_i})}{r - u_i}. \tag{201}$$

We cover the Cartesian product of the two root intervals by rational boxes. On each box, interval evaluation of (201) and the fact that $1/2 - u_i \le 0$ give a piecewise-affine lower bound, in the count k, on

$$N \max_i L_i(k/N, 1/2).$$

The resulting positive binomial sum is bounded from below by exponential tilting and outward tail recurrences. Every accepted box satisfies

$$\mathbb{E}_{K \sim \mathrm{Bin}(N,1/2)} \exp\{N \max_i L_i(K/N, 1/2)\} > 1, \tag{202}$$

contradicting $\max_i L_i \le f$ and $I_{f,N} \le 1$.

The endpoint cells and slope boxes cover the complete feasible region. The smallest outward-rounded lower bounds in (202) are 1.00002356884 for $N = 1024$ and 1.00002037412 for $N = 4096$. Thus the exclusion includes asymmetric allocations of the endpoint error and proves a finite-N lower bound independently of the Gaussian limit.

The slope intervals enclose both roots of the concave constraint in (200); endpoint derivative signs distinguish the two roots. Exact rational subtraction of tail probabilities followed by interval exponentiation gives the lower bound on each box. These bounds certify the displayed instances without asserting a general efficient algorithm. The relative widths $(U-L)/U$ differ from the relative suboptimality guarantees $(U-L)/L$, which are below 10.6% in both instances.

Q Closest Prior Guarantees and the Joint Contribution

Section 1.1 gives the broader context. Gastpar et al. (2025) study risk estimability for a learning algorithm and a collection of distributions. They give instability-based obstructions, loss-stability

sufficient conditions, and a conditional-variance characterization of average squared-error estimability (Proposition 5.2). That characterization concerns estimation error averaged over the specified distribution family and samples. Our criterion instead measures a width quantile under each law, with simultaneous posterior coverage under every law and pathwise CP calibration. We determine the optimal width under these additional requirements; their estimators need not satisfy them.

Within PAC–Bayes comparators, Hellström & Guedj (2024, Theorem 4) study an average-generalization objective, whereas their Section 3 and Theorem 8 study high-probability PAC–Bayes bounds. The latter identify a Cramér-comparator benchmark and give valid near-optimal bounds with normalization accounted for. We study a different optimization: one normalized comparator must serve a fixed family of sample states, and the objective is the weighted average of their inverse risk endpoints. This yields an exact compatibility criterion and sharp excess laws for the specified target regimes, beyond identifying a pointwise comparator.

Uniform parameter selection is also distinct from attainment of the unmodified statewise benchmark. For bounded binary losses, the comparison with the simultaneously valid temperature bound of Rodríguez-Gálvez et al. (2024) is

$$t_0(Q) = \frac{\mathrm{KL}(Q\|P) + \log(1/\delta)}{N}, \qquad t_{\mathrm{SL}}(Q) = t_0(Q) + \frac{\log \xi(N)}{N},$$

where $\xi(N)$ is the Seeger–Langford binomial moment. Their parameter may be selected from the sample on the common event at t_{SL}, without a further parameter-search union bound. Our selection price is relative to t_0 within the normalized convex-comparator class. Casado et al. (2024) extend exact free-parameter optimization through Cramér–Chernoff bounds to more general losses.

Other routes refine the certificate itself. The coin-betting bound of Jang et al. (2023) is time-uniform, can be evaluated through optimization over possible mean functions, and recovers Bernoulli-KL and empirical-Bernstein forms by relaxation. Kuzborskij et al. (2024) replace KL-based complexity with a ZCP-divergence bound. These works are not formulated with original-confidence CP recovery at a specified prior; our calibration obstruction is conditional on that requirement. The numerical comparisons examine evidence refinement and coordinatewise budget allocation under the protocols in Appendix O.

For calibration, binomial randomization is already described by Foong et al. (2021, Appendix I.2). The transport $T(P, Q, c)$ is a classical testing power function, equal to one minus the type-II-error trade-off function of Dong et al. (2022, Definition 2). Our results embed CP in one posterior-valid rule and quantify the learning cost of exact or pathwise approximate recovery. Shared-point-mass recovery is a weaker requirement than every-prior recovery and permits joint continuity and KL-controlled learning. Table 5 separates these joint properties from their inherited ingredients.

Reverse-KL ambiguity sets and geometric-mean duals are established tools in distributionally robust optimization (Van Parys et al., 2021). Here the exponent comes from the permitted pathwise deviation from a CP endpoint, while the coverage probability remains fixed. The uniform construction in Appendix L uses logarithmic mixture tests to attain the resulting prior-constrained local frontier without estimating the full joint law. Its contribution is this common calibrated guarantee and uniform attainability, rather than a new reverse-KL duality principle.

Table 5: Closest ingredients to the calibration results and the additional properties established here.

Ingredient	Inherited role	Additional property established here
Clopper–Pearson (Clopper & Pearson, 1934)	Exact inversion for one fixed binomial parameter	Embedding in one common posterior-valid endpoint.
p-to-e calibration (Vovk & Wang, 2021)	Converts a binomial-tail p-value into mean-bounded evidence	Compatibility with point-mass CP equality, boundary continuity, and a KL learning rate.
Vector confidence regions (Adams et al., 2024)	Simultaneous vector control followed by linear optimization	Coordinates are hypothesis risks and may be optimized by a data-dependent posterior while retaining point-mass CP equality.
DP classifier selection (Oneto et al., 2017)	A binomial bound for a privately selected classifier	Original-confidence common-event recovery, with its unavoidable transport lower bound.
Exact bounded-mean methods (Vlassis & Thomas, 2026; Ming et al., 2026)	Calibration of individual bounded observations	Deterministic prior recovery by transport and a fixed-posterior obstruction under fiber-invisible alternatives.