COMPRESS NOW, CONTAMINATE LATER: ONE QUANTIZER FOR EVERY HUBER LEVEL

Anonymous authors
Paper under double-blind review

ABSTRACT

In binary hypothesis testing with compressed observations, the encoder may have to be fixed before anyone knows a bound on the contamination. We study this setting where the decoder learns a common Huber contamination bound only after encoding, and can then choose its test but not undo the encoder's merges. For known finite laws, $O(1 + \log(1/J))$ output symbols suffice to keep a constant fraction of robust Hellinger information at every feasible level, where J is comparable to clean squared Hellinger distance. We also characterize the cost of using fewer symbols, which can be steep. On a geometric family with L evidence scales, the best uniform sample overhead $\mathcal{R}_K$ satisfies $\log(1 + \mathcal{R}_K) \asymp 1 + L/K$, even though a bit selected separately for each level has constant overhead. An adjustable-growth version of our construction attains this order. The exponential penalty applies to one identical memoryless channel where a known schedule of different one-bit encoders has $O(L)$ overhead on the same family. Learning the encoder from clean observations is harder still. No finite training budget depending only on clean difficulty preserves every raw-feasible level with a fixed output budget. A fitted test does succeed above an estimation-error threshold, and independent calibration certifies tests through simultaneous moment bounds. Experiments measure the alphabet–information trade-off, compare designs that target clean information, and quantify what certification costs.

1 INTRODUCTION

Suppose each observation in a binary testing problem must be compressed before a bound on contamination is available. Once the bound arrives, the decoder is free to choose its test, but whatever the encoder merged stays merged. Ideally, compression would cost at most a constant factor in sample size relative to testing on the raw data, whichever feasible bound is announced.

To see the difficulty, consider how the robust test changes with contamination. Huber's test clips the nominal likelihood ratio at thresholds set by the contamination bound. Without contamination, weak evidence contributes to the information only quadratically in its strength, so an encoder tuned for clean testing can often afford to lump it together. Clipping takes away the advantage of strong evidence: a symbol with overwhelming evidence counts no more than one of the same sign just past the threshold. As contamination grows, the threshold moves down into the weak evidence, where the distinctions a clean encoder may have erased can account for much of the robust information that remains. On one of our dense stress-test pairs, a partition with 146 outputs that keeps 99.84% of clean mutual information retains only .001707% of the raw robust information at its worst tested level, an intermediate clipping scale (Section 6). We work with known laws P, Q on a finite alphabet, apply the same channel independently to each observation, and compare with uncompressed testing at the same bound ε.

Existing quantization methods establish logarithmic alphabet bounds for preserving clean information and testing power (Bhatt et al., 2019; Pensia et al., 2023), and robust decentralized detection has been studied for several decades (Poor & Thomas, 1978; Gül, 2017). If the contamination bound is known in advance, one can quantize its least-favorable laws, and the encoder changes whenever the bound does. In our setting the encoder is fixed once and has to serve every level, and we compare it with uncompressed testing at the same bound as even a constant-factor change in the bound can change the sample complexity substantially (Vallinayagam et al., 2026).

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.

We introduce an encoder that first separates positive and negative nominal likelihood evidence. Within each sign, it bins strong evidence geometrically by strength and pools weak evidence into bins whose probability mass grows geometrically, without ever looking at ε. Even so, the binning preserves clipped first and second moments at every threshold simultaneously; the first moments let us compare raw and compressed least-favorable laws at exactly the same ε, and the second moments control how much information survives. The resulting representation uses $O(1 + \log(1/J))$ symbols, where $J = \mathbb{E}_{(P+Q)/2}[(P-Q)^2/(P+Q)^2]$.

The communication budget determines how much reuse costs. On a family with $L \asymp \log(1/J)$ evidence scales, the optimal uniform sample overhead satisfies $\log(1 + \mathcal{R}_K) \asymp 1 + L/K$. Thus a small alphabet can incur an exponential penalty in L/K, even though a bit chosen for each supplied level has constant overhead. The construction attains this order by increasing the growth of both its geometric intervals and its lower-bin masses. We also show that exact total-variation preservation and constant-fraction clean Hellinger retention can coexist with arbitrarily large robust sample overhead.

Learning introduces a further difficulty: even at constant clean difficulty, preserving feasibility at every level can require identifying arbitrarily weak signs. We give a fitted-test guarantee when robust separation exceeds nominal-estimation error, then use independent calibration to certify the fitted test directly.

2 Problem formulation and related work

Let $P \neq Q$ be probability vectors on $[d]$. We model contamination using the Huber class $\mathcal{H}_\varepsilon(P) = \{(1-\varepsilon)P + \varepsilon R : R \in \Delta_d\}$, where Δ_d is the probability simplex on $[d]$. Under each hypothesis the observations are iid from a single law in its class, and nothing ties the contaminant under P to the one under Q. Let $n_\varepsilon(P, Q)$ be the smallest sample size at which some test keeps both worst-case errors at most $1/4$. Testing becomes infeasible when the classes overlap, which occurs precisely when

$$\varepsilon \geq \varepsilon_* = \frac{V}{1+V}, \qquad V = \mathrm{TV}(P, Q). \tag{1}$$

We consider the feasible range $0 \leq \varepsilon < \varepsilon_*$. All results use a *common upper bound* on contamination, while allowing the actual proportions under the two hypotheses to differ below that bound.

We represent the encoder by a column-stochastic K-by-d matrix T. Each observation X is encoded independently as $Y \sim T(\cdot \mid X)$, with private randomness that the decoder never sees. The channel may depend on P and Q, but not on the contamination level, the test's sample size, or the position of the observation. Let $n_\varepsilon^T(P, Q)$ denote the testing cost for the encoded input classes $T\mathcal{H}_\varepsilon(P), T\mathcal{H}_\varepsilon(Q)$. These classes are contained in $\mathcal{H}_\varepsilon(TP), \mathcal{H}_\varepsilon(TQ)$ and can be strictly smaller. A randomized channel need not realize every contaminant allowed by the full output classes.

We use the unhalved Hellinger metric $h(P, Q) = \|\sqrt{P} - \sqrt{Q}\|_2$ and define

$$\mathcal{I}_\varepsilon(P, Q) = \inf_{U \in \mathcal{H}_\varepsilon(P),\, W \in \mathcal{H}_\varepsilon(Q)} h^2(U, W). \tag{2}$$

The classical least-favorable-pair characterization of Huber testing (Huber, 1965) gives $n_\varepsilon(P, Q) \asymp 1/\mathcal{I}_\varepsilon(P, Q)$, with universal constants at the fixed error target. Thus, robust Hellinger information controls the number of observations needed for testing. Appendix A gives the explicit least-favorable pair, a direct exponential test bound, and the product-Hellinger lower bound used in our analysis.

For symbols with positive nominal mass, define

$$m_i = \frac{P_i + Q_i}{2}, \quad z_i = \frac{P_i - Q_i}{P_i + Q_i}, \quad J = \sum_i m_i z_i^2. \tag{3}$$

Set $z_i = 0$ when $m_i = 0$. These coordinates satisfy $\mathbb{E}_m z = 0$, $V = \mathbb{E}_m|z|$, and $J \leq h^2(P, Q) \leq 2J$. We measure alphabet size in *output symbols*, corresponding to $\lceil \log_2 K \rceil$ bits per observation under fixed-length communication.

Clean-information quantization. Logarithmic alphabet bounds for preserving clean information and testing power are established in Bhatt et al. (2019); Pensia et al. (2023). The constructive

proof of Theorem 3 in Bhatt et al. (2019) groups posterior divergence geometrically above a clean-information cutoff and merges values below it. Kazemi et al. (2025) develop reverse data-processing inequalities and communication-constrained testing bounds for general priors. Our guarantee preserves robust information simultaneously across contamination levels, with an alphabet controlled by clean difficulty. Theorem 3 gives an exponential-order trade-off for reuse. For comparison, the clean testing overhead in Pensia et al. (2023) is at most $O(\max\{1, \log(1/J)/K\})$.

For binary-input channels, Kurkoski & Yagi (2014) optimize clean mutual information using contiguous likelihood-ratio partitions. Carpi et al. (2021) study per-observation compression for the clean type-II error exponent, leading to a KL objective. Jiao et al. (2026) characterize degradations that preserve one-sample decoding error and study capacity optimization. These clean objectives motivate the baselines in Section 6.

Robust quantization at a supplied level. Example 4.5 of Pensia et al. (2023) shows that a clean-optimal binary channel can lose identifiability under total-variation perturbations. Their Theorem 4.6 uses reverse data processing to quantize the least-favorable pair at a supplied radius. We study Huber mixture neighborhoods with one encoder across radii. Proposition 4 shows that even exact total-variation preservation and constant-fraction clean Hellinger retention can coexist with failure of uniform robust sample efficiency. In particular, retaining signs below the clean-information cutoff does not suffice.

Robust quantizer-detectors also appear in the small-signal asymptotic analysis of Poor & Thomas (1978), while Gül (2017) derives saddle-point conditions for quantized robust likelihood ratios under specified uncertainty classes.

Learning and calibration. Learning a quantizer together with a decision rule has been studied through joint quantizer/classifier consistency (Nguyen et al., 2009); Gerber & Polyanskiy (2024) study training/test tradeoffs and estimation followed by robust testing, and Saha & Ramdas (2025) develop robust composite and sequential tests, including plug-in constructions. Our learning guarantees concern one finite-output map, learned once, and a fixed-sample decoder. On the certification side, Shi et al. (2026) infer entire Neyman–Pearson trade-off functions from finite samples using attainable threshold classes, DKW bounds, and VC dimension. Our calibration corollary uses the same standard DKW method (Massart, 1990; Shi et al., 2026), applied to monotone fitted scores so that all of them are controlled on one confidence event. Sequential experiment simulation under endogenous stopping (Li, 2026) permits processing rules beyond a single memoryless kernel, which lie outside our model.

3 One encoder at every common contamination bound

Our first result shows that, given enough outputs, reuse costs only a constant factor in sample size.

Theorem 1 (Reusable representation)**.** *For every distinct finite pair P, Q and growth factor $q \geq 2$, there is a memoryless Markov encoder T_q with*

$$K_q \leq \min\left\{d,\ 5 + \frac{3\log 8 + 6\log(1/J)}{\log q}\right\}$$

outputs that preserves $\mathrm{TV}(T_qP, T_qQ) = V$ and satisfies

$$\mathcal{I}_\varepsilon(T_qP, T_qQ) \geq \frac{1}{4q^3}\mathcal{I}_\varepsilon(P, Q) \quad (0 \leq \varepsilon < \varepsilon_*). \tag{4}$$

Consequently $n_\varepsilon^{T_q}(P, Q) \leq 80q^3 n_\varepsilon(P, Q)$ throughout this range. The information on the left uses the full output Huber classes, so the bound also holds for input contamination.

Taking $q = 2$ gives $O(1 + \log(1/J))$ outputs, information retention $1/32$, and sample overhead at most 640. The alphabet is independent of d and of the distance to ε_*. Its fixed-length communication cost is $O(\log(2 + \log(1/J)))$ bits per observation. We use identity when the constructed alphabet exceeds d.

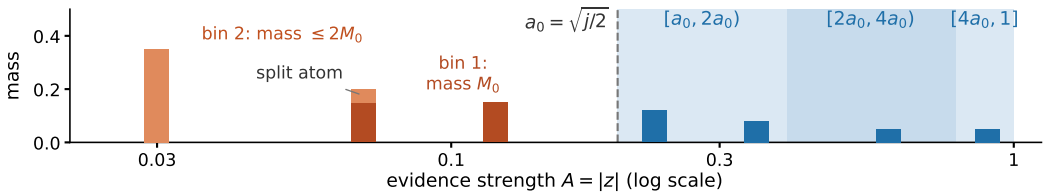


Figure 1: The encoder on one sign of a toy law ($q = 2$; bar height is conditional mass). Strong evidence ($A \geq a_0 \approx 0.198$, blue) is binned by strength into dyadic intervals and has mass $M_0 = 0.3$. Weak evidence (orange) is binned by mass, strongest first: M_0, then at most $2M_0$, splitting the atom at 0.07. Seven atoms become five outputs.

Encoder construction. We construct the bins separately for each nonzero sign of z (Figure 1). Condition m on that sign and define $A = |z|$, $j = \mathbb{E}A^2$, and $a_0 = \sqrt{j/2}$. For values $A \geq a_0$, use geometric intervals $[q^k a_0, \min\{q^{k+1} a_0, 1\})$ for $k \geq 0$, closing the final interval at 1. Let M_0 be their total conditional mass. Traverse the remaining values from largest to smallest. If the mass already assigned is H, give the next bin capacity $(q-1)H$. A complete bin increases the assigned mass to qH. The last bin contains any remaining mass. When an atom crosses a bin boundary, we allocate its remaining mass to successive bins until it is exhausted. Allocating nominal mass b_{yi} from atom i to bin y gives the private-randomness transition $T(y \mid i) = b_{yi}/m_i$. The encoder outputs the sign and bin index, with one additional output shared by all values for which $z = 0$. Conditional bin means are used in the analysis. A sparse implementation takes $O(d \log d + K)$ arithmetic operations and $O(d + K)$ storage, where K counts the constructed bins before fallback. Appendix I.1 gives pseudocode, tie conventions, and precision checks.

Lemma 2 (Clipped-moment quantization)**.** *For the preceding conditional construction, let $\bar{A}$ be the bin mean and define $F(t) = \mathbb{E}\min(A, t)$ and $G(t) = \mathbb{E}\min(A, t)^2$, with $\bar{F}, \bar{G}$ defined analogously. There are $O(1 + \log(1/j)/\log q)$ bins and, simultaneously for every $t \geq 0$,*

$$\bar{F}(t/q) \leq F(t) \leq \bar{F}(t), \qquad G(t) \leq q^3 \bar{G}(t/q). \tag{5}$$

Why can one binning serve every threshold? Clipping at t replaces A by $\min(A, t)$. An upper bin spans a factor of at most q in strength, so clipping its mean at t/q never gives more than clipping its values at t; this is the t/q in (5). A lower bin can span many strengths but holds at most $q-1$ times the mass above it. At most one lower bin straddles t, and each unit of mass above it contributes t to $F(t)$ but at most t/q to $\bar{F}(t/q)$. This slack covers the straddling bin. For second moments, each lower bin is at most q times heavier than its predecessor and no stronger, so it is charged to the predecessor with factor q^3 (Appendix B). For sign mass m_s, Jensen's inequality gives $j_s \geq [V/(2m_s)]^2 \geq J^2/4$, which turns the bin count into the bound in Theorem 1.

Comparing information at the same contamination level. Compression can change which contamination pair is least favorable. We therefore solve the raw and compressed clipping equations separately. Let $k = \varepsilon/(1-\varepsilon) < V$. For each sign s, the raw least-favorable pair has clipping threshold t_s characterized by

$$2\mathbb{E}_m[(|z| - t_s)_+ \mathbf{1}\{\operatorname{sign} z = s\}] = k(1 + t_s). \tag{6}$$

Each signed first moment equals $V/2$, and the construction preserves both exactly. We can therefore express the raw and compressed clipping equations with the same right side:

$$F_s(t_s) + kt_s/2 = (V - k)/2, \tag{7}$$

where F_s includes the unnormalized mass of sign s. Lemma 2 implies $t_s/q \leq \bar{t}_s \leq t_s$. For both least-favorable pairs, squared Hellinger information lies between $(1-\varepsilon)$ and $4(1-\varepsilon)$ times the sum of signed clipped second moments. The common factor cancels in the comparison, and (5) gives $1/(4q^3)$. At $\varepsilon = 0$, we compare the second moments directly. Appendix C gives the complete proof.

4 THE COST OF REUSE AND THE FAILURE OF CLEAN OBJECTIVES

Write

$$R_T = \sup_{0\le\varepsilon<\varepsilon_*} \frac{n_\varepsilon^T(P,Q)}{n_\varepsilon(P,Q)}, \qquad \mathcal{R}_K(P,Q) = \inf_{T:\,|\mathcal{Y}|\le K} R_T.$$

So R_T is the worst factor, over feasible levels, by which T increases the required sample size, and $\mathcal{R}_K$ is the best such factor among identical channels with at most K outputs (Section 2). To study how the alphabet requirement grows, fix $1 < \alpha < 2$ and let L grow. Under m, let B be a fair sign independent of U, where $\Pr(U = 2^j) = c_L 2^{-\alpha j}$ for $0 \le j \le L$, and set $A = \delta U$, $\delta = 2^{-L-1}$. Define $P = m(1 + BA)$ and $Q = m(1 - BA)$. Here $J \asymp \delta^\alpha$ and $V \asymp \delta$. Evidence thus comes in $L + 1$ strengths $\delta, 2\delta, \ldots, 1/2$, each doubling of strength making it 2^α times rarer. Different contamination levels clip at different strengths and so rely on different parts of the law.

Theorem 3 (Alphabet–overhead trade-off)**.** *For every distinct finite pair P, Q and integer $K \ge 2$,*

$$\log(1 + \mathcal{R}_K(P,Q)) \le C\left(1 + \frac{\log(1/J)}{K}\right). \tag{8}$$

On the preceding family (P_L, Q_L),

$$\log(1 + \mathcal{R}_K(P_L,Q_L)) \asymp_\alpha 1 + \frac{L}{K}. \tag{9}$$

A two-output encoder chosen separately for each supplied level has $n_\varepsilon^{T_\varepsilon}(P_L,Q_L) \le C_\alpha n_\varepsilon(P_L,Q_L)$ throughout the feasible range.

The upper bound chooses $\log q = O(1+\log(1/J)/K)$ in Theorem 1. A sign bit handles the smallest budgets (Appendix C.1). For the hard family, the lower bound is $K\log(1 + R_T) \ge c_\alpha L$. Constants may depend on α; the matching order concerns $\log(1 + R_T)$.

At raw clipping threshold $t_\ell = \delta 2^\ell$, the robust information is $\asymp \delta^2 2^{(2-\alpha)\ell}$. A bit that detects an appropriate signed amplitude near this scale preserves the order of the information. The relevant amplitude scale varies with the contamination level. For a fixed channel, let a_y be the mass of output y under the *nominal* m and write $s_y = a_y^{-1/\alpha}$. A tail-integral bound gives the following upper bound on the Hellinger information of the encoded raw least-favorable pair, normalized by raw information:

$$C_\alpha \sum_y (2^\ell/s_y)^\alpha \min\{1, (s_y/2^\ell)^2\}. \tag{10}$$

Each output's contribution decays geometrically away from its scale. Overhead R_T requires information of order $1/R_T$ at every level. Capping the contributions at this requirement leaves each output able to cover only $O_\alpha(\log(1 + R_T))$ scales. Summing over the $L - 1$ interior levels gives the lower bound (Appendix D). The argument allows private randomization and uses admissible input contamination.

A known schedule of encoders. If the encoder may depend on the observation index, cycle through the sign bit and the $L + 1$ positive-amplitude event bits. Given ε, the decoder keeps the matching subsequence. This fixed schedule has $O_\alpha(L)$ sample overhead and sends one bit per observation. Appendix D.3 proves the bound. The decoder uses the known index to tell the encoders apart.

Proposition 4 (Clean information and sign are insufficient)**.** *There exist deterministic encoders with $O(\log(1/J))$ outputs that preserve V exactly and a constant fraction of $h^2(P,Q)$, but whose uniform robust sample overhead diverges along the preceding family. The sign-preserving cutoff refinement has the stated TV and clean-Hellinger guarantees. Both cutoff variants lose uniform robust sample efficiency.*

The failure can be seen by retaining each signed raw symbol above $j_0 = (1 - \alpha/2)L + O(1)$ and merging all lower values separately by sign. This partition refines the posterior-KL cutoff construction because posterior divergence is comparable to A^2 for $A \le 1/2$. Most clean information lies in the retained upper levels, so merging the lower values has little effect on clean performance. At

$\ell = \gamma L$ with $0 < \gamma < 1 - \alpha/2$, the encoded raw least-favorable pair instead has relative information at most

$$C_\alpha \left[2^{-(2-\alpha)\ell} + 2^{-\alpha(j_0 - \ell)} \right] \longrightarrow 0. \tag{11}$$

Since the partition preserves V, it never loses feasibility. It instead fails when its information ratio vanishes, so along the family its sample cost grows without bound relative to the raw test.

5 Learning a reusable representation

Everything so far took P and Q as given. In practice they are estimated, estimation error can change which evidence the encoder keeps, and there is no chance to repair the encoder once the contamination bound arrives. Our learner sees N clean iid observations per class and must commit to one channel before the bound is supplied. To isolate what it costs to learn the encoder, the lower bound below even hands the decoder the true nominal laws. We call a training run successful if its realized encoder keeps testing feasible at every raw-feasible contamination level (every $\varepsilon < \varepsilon_*$), uniformly over admissible contaminants.

Proposition 5 (Geometry of actual input contamination). *For $\Delta = P - Q$, define*

$$\kappa_T = \min_{Tv = T\Delta} \frac{\|v\|_1}{2} = \max_{f \in \mathrm{row}(T) \cap [0,1]^d} \langle f, \Delta \rangle. \tag{12}$$

The encoded input classes overlap exactly when $\varepsilon/(1-\varepsilon) \geq \kappa_T$. If every Δ_i is nonzero, then $\kappa_T = V$ if and only if $\mathbf{1}\{\Delta > 0\} \in \mathrm{row}(T)$.

κ_T is the separation that survives encoding, the smallest total variation of an input perturbation that the encoder cannot tell apart from $P - Q$. It replaces V in (1). The row-space condition comes from the cube: when no Δ_i vanishes, the sign indicator $\mathbf{1}\{\Delta > 0\}$ is the unique maximizer of $\langle f, \Delta \rangle$ there. It also means that output total variation does not decide feasibility under input contamination. An invertible noisy channel can shrink output TV and still have $\kappa_T = V > \mathrm{TV}(TP, TQ)$, in which case the input overlap boundary does not move at all.

Theorem 6 (A training obstruction at constant clean difficulty). *For every $r \geq 1$ and $0 < \eta \leq 1/4$, there is a family on $d = 2r + 2$ symbols with $J = 1/8$. If a learner with at most K outputs preserves testing feasibility at every raw-feasible common Huber bound with probability at least $1 - \beta$, uniformly over this family, then*

$$N\eta^2 \geq [(1-\beta)r - K - 1] \log 2. \tag{13}$$

In particular, a uniformly bounded output alphabet admits no finite uniform pilot budget depending only on clean J for this objective.

The hard family is simple. Two strong anchor symbols fix the clean difficulty, and r weak pairs carry unknown signs $\theta \in \{-1, 1\}^r$ at amplitude η. By Proposition 5, a map that keeps every raw-feasible level feasible must have the full true sign indicator in its row space. A K-dimensional row space has room for at most 2^K of the 2^r sign patterns, yet the pilot data reveal at most $N\eta^2$ nats about which pattern is the true one. A list-Fano argument converts this mismatch into (13).

Fitting a test above the estimation-error threshold. We certify a fitted test when its estimated robust separation is large compared with a high-probability bound on the nominal estimation error. Suppose a pilot event with probability at least $1 - \beta$ ensures $h(P, \widehat{P}) + h(Q, \widehat{Q}) \leq \rho$. We construct $T = T_2(\widehat{P}, \widehat{Q})$ once and define $\widehat{d}_\varepsilon = \sqrt{\mathcal{I}_\varepsilon(T\widehat{P}, T\widehat{Q})}$.

Theorem 7 (Finite-sample fitted test). *When $\widehat{d}_\varepsilon > 0$, fit the full-output Huber least-favorable likelihood-ratio test and use $n = \lceil 8/\widehat{d}_\varepsilon^2 \rceil$ fresh encoded observations. Certify it only if $\widehat{d}_\varepsilon \geq 14\rho$. On the pilot event, both worst-case errors of every certified test are below $1/4$ under input contamination. Every level satisfying*

$$\sqrt{\mathcal{I}_\varepsilon(P, Q)} \geq 81\rho \tag{14}$$

is certified and has $n = O(1/\mathcal{I}_\varepsilon(P, Q))$. The same pilot event works for all such levels, with the encoder fixed.

The proof uses the same contaminants under the true and estimated laws, applies Hellinger contraction, and transfers the fitted product test's risk at the original ε. For empirical laws on d symbols, $\rho = \sqrt{4(d-1)/(N\beta)}$ is a valid choice. The contamination level may be chosen after the pilot but is fixed before deployment. The procedure abstains when it cannot certify a test.

Independent calibration. The elementary radius above is conservative. It grows with d because it controls the whole nominal law, and it relies on Markov's inequality (Appendix G.3). Another option is to hold out a second clean sample and use it to check the fitted test itself. Condition on the training data, which fixes T and the order of the nominal output likelihood ratios. Let P_v, Q_v be the empirical laws of M fresh clean observations per class, pushed through T with fresh private randomness, and set $\delta_M = \sqrt{\log(4/\beta)/(2M)}$. DKW, applied once per class, gives a single event of probability at least $1-\beta$ on which both output CDFs, read in this fixed order, lie within δ_M of their empirical versions.

Corollary 8 (Independent monotone calibration)**.** *Let $(\widehat{p}^*_\varepsilon, \widehat{q}^*_\varepsilon)$ be the full-output Huber least-favorable pair for $T\widehat{P}, T\widehat{Q}$, and define $g_\varepsilon(y) = \log(\widehat{p}^*_\varepsilon(y)/\widehat{q}^*_\varepsilon(y))$, with score zero when both probabilities vanish. For a finite score and $\lambda > 0$, put $f_P = e^{-\lambda g_\varepsilon}$ and $f_Q = e^{\lambda g_\varepsilon}$. For $R = P, Q$, define*

$$C_R = \min\{\max_y f_R(y),\ \mathbb{E}_{R_v} f_R + \delta_M(\max_y f_R(y) - \min_y f_R(y))\},$$
$$B_R(\varepsilon,\lambda) = (1-\varepsilon)C_R + \varepsilon \max_x \sum_y T(y \mid x) f_R(y). \tag{15}$$

Minimize B_P and B_Q over evaluated positive tilts, allowing the trivial bound one, and let b be the larger minimum. The tilts may be chosen after calibration. If $b < 1$, decide P when $\sum_{i=1}^n g_\varepsilon(Y_i) \geq 0$ and Q otherwise, with $n = \lceil \log 4/(-\log b) \rceil$. On the single DKW event, each worst-input error is at most $1/4$, simultaneously for all supplied ε for which the procedure certifies. Abstain when the estimated classes overlap, the score is infinite, or $b \geq 1$.

The corollary works in three steps. Training fixes T and one order of the outputs by fitted likelihood ratio; Huber clipping never reverses this order, so f_P and f_Q are monotone in it for every ε and λ. On the DKW event, summation by parts gives $\mathbb{E}_{TR} f \leq \mathbb{E}_{R_v} f + \delta_M(\max f - \min f)$ for every monotone f, so C_R bounds the clean moment for all levels and tilts at once. B_R then adds the exact worst *input*-contaminant moment, and Markov's inequality turns B_R^n into the error bound (Appendix G.4). Under either hypothesis, the probability of certifying and then erring is at most $\beta + 1/4$, averaged over training, calibration, and deployment; conditional error also depends on the certification rate. The level and test size may use the calibration data but must be fixed before deployment.

6 Experimental evaluation

We evaluate the representation in synthetic finite experiments where the true information can be computed. We sample clean training and calibration data; deployment error bounds are computed from the specified laws, with numerical binomial-tail evaluation for binary pairs. All law and information calculations use 80 decimal digits.

Output budget and information loss. We vary the output budget on 16 geometric pairs with $L \in \{8, 16, 32, 64\}$, giving 104 pair–budget settings. At each budget, we fit the adjustable-growth encoder and a sign-preserving clean-MI partition. The growth factor is chosen by output count alone. We report $D_T = \max_{\varepsilon \in \mathcal{E}} \mathcal{I}_\varepsilon(P,Q)/\mathcal{I}_\varepsilon(TP,TQ)$ on a grid $\mathcal{E}$ of clipping scales and boundary fractions. The denominator uses full output Huber classes. D_T measures information loss on this grid.

Figure 2 shows how the loss changes with L/K. At $L = 64$ and $\alpha = 1.5$, the adjustable encoder has D_T of 1,384, 9.06, and 2.06 at budgets 8, 16, and 32. Clean MI has losses of $1.25 \cdot 10^8$, $3.11 \cdot 10^7$, and $4.63 \cdot 10^6$ at the same budgets. At $L = 8$, clean MI does better at budgets 4 and 8. On smaller laws, exhaustive search finds partitions that retain more than ours. Appendix I.15 reports every budget setting, the restricted partition search, and the one-bit schedule (Appendix D.3).

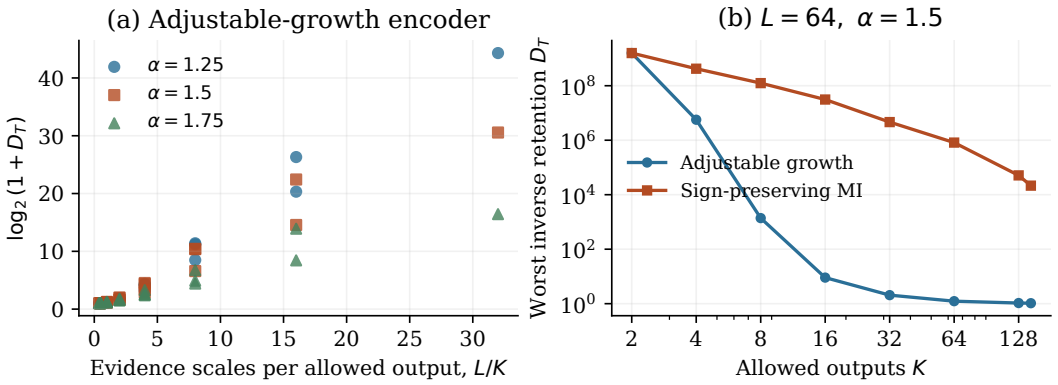


Figure 2: Output budget and worst inverse information retention. (a) Adjustable encoders on the 12 pairs with four amplitudes per dyadic interval. (b) Comparison at a fixed law and equal allowed alphabet size. Lower D_T means better retention. These finite-grid information losses are distinct from the uniform sample overhead R_T in Theorem 3.

Comparison at $q = 2$. In this known-law comparison the growth factor is fixed at $q = 2$, and every reusable map is fitted once and evaluated unchanged at all supplied levels. We use 15 pairs and 667 contamination levels in total: twelve pairs from the geometric family with $\alpha \in \{1.25, 1.5, 1.75\}$ and $L \in \{16, 32, 48, 64\}$, and three dense pairs with $\alpha = 1.5$, $L \in \{16, 32, 64\}$, and 32 distinct amplitude levels per dyadic interval, which makes the likelihood ratios within each sign distinct. The grid covers clean data, five fractions of the overlap boundary, and every interior integer clipping scale. The large-L geometric pairs are asymptotic stress tests. The dense $L = 64$ pair, for example, has $J \approx 9.36 \times 10^{-30}$ and $\varepsilon_* \approx 8.04 \times 10^{-20}$, so even uncompressed clean testing requires a very large sample. We compare against the posterior-KL geometric construction, a sign-preserving variant of it held to our output budget, the sign bit, a threshold bit selected for clean information, and an oracle that picks the best threshold bit separately at each ε. Appendix I gives the constructions, grids, seeds, and the identifier of each run.

The savings can be large. On the dense $L = 64$ pair, the $q = 2$ encoder maps 4,098 input symbols to 146 outputs (13 bits down to 8 under fixed-length coding, an alphabet about 28 times smaller) and still keeps at least 96.36% of raw robust information at every tested level (Table 1; Table 4 has all three dense pairs).

Table 1: Dense $L = 64$ pair ($d = 4{,}098$), as fractions of the raw values. "Worst" is the minimum over 69 levels.

Map	K	Clean MI	Worst
Ours, $q = 2$	146	.9642	.9636
Clean-MI optimal	146	.9984	$1.7 \cdot 10^{-5}$
Signed KL	142	.9816	.0174
Design at $.5\varepsilon_*$	146	.9961	$6.2 \cdot 10^{-6}$

Would maximizing clean mutual information with the same output budget do as well? Following Kurkoski & Yagi (2014), we optimize contiguous likelihood-ratio partitions within each sign and choose how to split the K outputs between signs. At $K = 146$ this map keeps the most clean MI in Table 1, but its robust information collapses at intermediate clipping scales (Figure 3). It preserves TV exactly, so testing stays feasible. The cost is a larger sample size. Appendix I.3 checks all three optimized partitions with an independent solver.

To test how far a level-specific design can be reused, we quantize the raw least-favorable pair at $.1, .5$, or $.9\varepsilon_*$ with our output budget, then freeze the map. Spare symbols refine tied least-favorable ratios to preserve clean MI. At $L = 64$ all three keep at least 99.98% of robust information at their design levels, yet the best minimum over the grid is only $1.707 \cdot 10^{-5}$ (Appendix I.10 gives each design before and after refinement).

All twelve coarse cases trigger identity fallback. Their support has few atoms across many evidence scales. A density sweep at $\alpha = 1.5$ and $L = 16, 32, 64$ yields strict compression with two distinct amplitudes per scale: $66, 130, 258$ inputs become $38, 74, 146$ outputs. Increasing the density to 32 values per scale leaves those output counts unchanged (Appendix I.11).

Heterogeneous discrete pairs. We also sample 384 pairs from sparse Dirichlet, uniform Dirichlet, lognormal, and correlated ensembles, with $d \in \{8, 32, 128, 512\}$. The correlated ensemble includes nearly identical laws. Each pair is evaluated at nine fixed fractions of its overlap boundary. The deployed alphabet is strictly smaller than the input alphabet in 377 of these 384 pairs, and the smallest retained-information ratio is .78385. Only two pairs trigger identity fallback, both correlated pairs with $d = 8$. At $d = 512$, median output counts are 4, 6, 4, and 13 in the four ensembles, corresponding to 2, 3, 2, and 4 bits instead of 9. Appendix I.4 gives all results.

Learning and certification. How much clean data does the elementary certificate (Theorem 7 with the radius of Appendix G.3) actually need? We try the binary pair $(.8, .2), (.2, .8)$ and two constant-J families from Theorem 6 with $r = 16, 64$ and $\eta = .03$, fixing every weak sign at $+1$ so that the experiment measures how training noise affects the fitted encoder and test. For each $N \in \{10^3, 10^4, 10^5, 10^6, 10^7, 10^8\}$ per class we draw 32 independent pilot pairs as multinomial counts and evaluate the levels $0, .1, .5, .9, .99$ of the true overlap boundary with $\beta = .05$ and unsmoothed empirical laws. Table 7 lists the smallest tested pilot sizes at which all 32 fitted tests are certified.

The elementary certificate can require far more training data than the fitted test needs to perform well. For the $d = 34$ family at $N = 10^5$, all 32 tests at $.9\varepsilon_*$ have worst-input Chernoff bounds below $1/4$ when evaluated with the true laws, although none is certified. Fitted tests can also fail near overlap. For the binary pair at $N = 10^4$ and $.99\varepsilon_*$, one proposed test has worst-world error above $1 - 10^{-11}$, and the certificate abstains. If the fitted score has the wrong sign in expectation under an admissible law, increasing the deployment sample size drives the test toward the wrong decision. Appendix I.5 gives the full results.

Calibration at equal total clean budget. Is it worth setting aside half of the clean data for calibration? We compare independent calibration (Corollary 8) with the elementary certificate on the same three families at the same total clean budget B per class. The calibrated method fits on $B/2$ observations and calibrates on an independent $B/2$. The baseline fits its own encoder on all B and certifies with the conservative ambient-alphabet Hellinger radius obtained from Markov's inequality (Appendix G.3). There are 32 paired replicates at each of the same six budgets and five levels, with the seeds specified in Appendix I.6 and $\lambda \in \{1/8, 1/4, 1/2, 3/4, 1\}$. Both methods use $\beta = .05$ and an error target of $1/4$ on their confidence events, and each selects its own deployment sample size.

For the clean $d = 130$ pair, the smallest tested budget at which all 32 runs certify is $B = 10^4$ with calibration and $B = 10^8$ with the elementary certificate (Table 9). The calibrated tests at $B = 10^4$ use $n = 34$–57 deployment observations. Appendix I.6 reports certification counts, confidence-event checks, and deployment costs.

Where the deployment cost comes from. Near overlap, both the underlying testing problem and calibration uncertainty contribute to deployment cost. For $d = 130$ at $.99\varepsilon_*$, the raw testing bounds require at least 35,722 observations and give a sufficient count of 396,160. At $B = 10^7$, calibration increases the true-moment sufficient count of the fitted score by a median factor of 3.62. Wider tilt searches substantially reduce the largest counts in this group but leave the median nearly unchanged (Appendices I.7 and I.12).

7 Scope and Limitations

The logarithmic alphabet guarantee uses a common contamination bound. For separate bounds under P and Q, exact feasibility preservation can require one output per distinct nominal likelihood ratio, even at constant clean difficulty (Appendix H). Whether allowing a margin from the asymmetric overlap boundary reduces this cost remains open.

Appendix I.14 extends the construction to known laws with an atomless contrast distribution. Applying it to measured data requires estimates of the nominal contrast and its mixture quantiles, together with a contamination model appropriate to deployment.

With learned laws, calibration may abstain at raw-feasible levels or require more observations than the uncompressed optimum. Near overlap, the moment margin shrinks and CDF uncertainty can substantially increase the certified sample size. Determining sharp calibration costs, including the benefit of variance-sensitive confidence bounds, remains open (Appendix I.13).

AI USE STATEMENT

We used generative AI tools to assist with theoretical development, proof writing, code editing, and manuscript revision. We reviewed all AI-assisted material, checked the mathematical arguments and proofs, tested the code, and verified literature claims against the original sources. We take responsibility for the final manuscript and accompanying artifacts.

REPRODUCIBILITY STATEMENT

We provide complete proofs in Appendices A–H and experimental specifications in Appendix I. The accompanying source package includes the experimental protocols, experiment drivers, numerical results, and source hashes. It also includes the encoder implementation and the independent full-recurrence solver used to check the MI baseline.

REFERENCES

Alankrita Bhatt, Bobak Nazer, Or Ordentlich, and Yury Polyanskiy. Information-Distilling Quantizers. arXiv:1812.03031, version 2, 2019. URL https://arxiv.org/abs/1812.03031v2.

Fabrizio Carpi, Siddharth Garg, and Elza Erkip. Single-Shot Compression for Hypothesis Testing. arXiv:2107.09778, 2021. URL https://arxiv.org/abs/2107.09778.

Patrik Róbert Gerber and Yury Polyanskiy. Likelihood-free hypothesis testing. arXiv:2211.01126, version 5, 2024. URL https://arxiv.org/abs/2211.01126v5.

Gökhan Gül. Minimax Robust Decentralized Detection in Parallel Sensor Networks. arXiv:1703.07670, 2017. URL https://arxiv.org/abs/1703.07670.

Peter J. Huber. A robust version of the probability ratio test. *Annals of Mathematical Statistics*, 36 (6):1753–1758, 1965. URL https://doi.org/10.1214/aoms/1177699803.

Yadong Jiao, Xiaoyan Cheng, Yuansheng Tang, and Ming Xu. On the degradations of Binary-Input Discrete Memoryless Channels. arXiv:2604.25552, 2026. URL https://arxiv.org/abs/2604.25552.

Hadi Kazemi, Ankit Pensia, and Varun Jog. The Sample Complexity of Distributed Simple Binary Hypothesis Testing under Information Constraints. arXiv:2506.13686, 2025. URL https://arxiv.org/abs/2506.13686.

Brian M. Kurkoski and Hideki Yagi. Quantization of Binary-Input Discrete Memoryless Channels. *IEEE Transactions on Information Theory*, 60(8):4544–4552, 2014. URL https://doi.org/10.1109/TIT.2014.2327016.

Zihao Li. The Order of Binary Experiments under Endogenous Stopping. arXiv:2608.19897, 2026. URL https://arxiv.org/abs/2608.19897.

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

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009. URL https://arxiv.org/abs/0907.3740.

XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. On surrogate loss functions and f-divergences. *Annals of Statistics*, 37(2):876–904, 2009. URL https://doi.org/10.1214/08-AOS595.

Ankit Pensia, Varun Jog, and Po-Ling Loh. Communication-constrained hypothesis testing: Optimality, robustness, and reverse data processing inequalities. arXiv:2206.02765, version 2, 2023. URL https://arxiv.org/abs/2206.02765v2.

H. Vincent Poor and John B. Thomas. Asymptotically robust quantization for detection. *IEEE Transactions on Information Theory*, 24(2):222–229, 1978. URL https://doi.org/10.1109/TIT.1978.1055864.

Aytijhya Saha and Aaditya Ramdas. Huber-robust likelihood ratio tests for composite nulls and alternatives. arXiv:2408.14015, version 8, 2025. URL https://arxiv.org/abs/2408.14015v8.

Kaining Shi, Qiaosen Wang, and Cong Ma. When Are Trade-Off Functions Testable from Finite Samples? arXiv:2605.10774, 2026. URL https://arxiv.org/abs/2605.10774.

Shankar Vallinayagam, Ankit Pensia, and Varun Jog. On the Sample Complexity of Robust Binary Hypothesis Testing. arXiv:2605.24741, 2026. URL https://arxiv.org/abs/2605.24741.

A HUBER GEOMETRY AND THE TESTING-INFORMATION RELATION

We give the finite-alphabet details underlying the classical robust test. For $U \in \mathcal{H}_\varepsilon(P)$, the coordinate characterization is $U_i \geq (1-\varepsilon)P_i$ and $\sum_i U_i = 1$. A common law for the two classes exists exactly when

$$(1-\varepsilon)\sum_i \max(P_i, Q_i) = (1-\varepsilon)(1+V) \leq 1.$$

This proves (1), including coordinates with zero mass. When $P = Q$, the classes overlap at $\varepsilon = 0$.

For $0 < \varepsilon < \varepsilon_*$, set $k = \varepsilon/(1-\varepsilon)$. For sign $s \in \{-1, 1\}$, the left side of (6) is continuous and nonincreasing in t, equal to $V > k$ at $t = 0$ and to zero at $t = 1$. Subtracting $k(1+t)$ makes it strictly decreasing, so there is a unique root $t_s \in (0,1)$. Put $c_s = (1-t_s)/(1+t_s)$. Define, coordinatewise,

$$P_i^* = (1-\varepsilon)\max(P_i, c_- Q_i), \qquad Q_i^* = (1-\varepsilon)\max(Q_i, c_+ P_i). \tag{16}$$

On the negative-sign side,

$$\sum_i (c_- Q_i - P_i)_+ = \frac{2}{1+t_-}\mathbb{E}_m[(|z| - t_-)_+ \mathbf{1}\{z < 0\}] = k.$$

Thus P^* normalizes; the same calculation holds for Q^*. Both belong to their Huber classes.

Why these laws minimize Hellinger distance. Let $r_i = \sqrt{Q_i^*/P_i^*}$ on positive nominal support, and assign $r_i = 1$ where both probabilities vanish. The ratio is clipped between $\sqrt{c_+}$ and $1/\sqrt{c_-}$. The excess $P^* - (1-\varepsilon)P$ has mass ε entirely at the maximum of r. Therefore P^* maximizes $\sum_i U_i r_i$ over $U \in \mathcal{H}_\varepsilon(P)$. Similarly, Q^* maximizes $\sum_i W_i/r_i$ over $W \in \mathcal{H}_\varepsilon(Q)$. The elementary inequality

$$\sqrt{uw} \leq \tfrac{1}{2}(r_i u + w/r_i)$$

is an equality at (P_i^*, Q_i^*). Summing proves

$$\sum_i \sqrt{U_i W_i} \leq \sum_i \sqrt{P_i^* Q_i^*}.$$

The same inequality handles zero nominal support using $r_i = 1$. Consequently $\mathcal{I}_\varepsilon(P, Q) = h^2(P^*, Q^*)$. At $\varepsilon = 0$ the classes are singletons, so the conclusion is immediate without clipping.

Finite-sample upper and lower bounds. Test P when $\prod_{j=1}^n P^*(X_j)/Q^*(X_j) \geq 1$ and Q otherwise. Because P^* maximizes the expectation of r, Markov's inequality and independence give, for every $U \in \mathcal{H}_\varepsilon(P)$,

$$\Pr_{U^{\otimes n}}(\text{decide } Q) \leq \left(\sum_i P_i^* r_i\right)^n = (1 - \mathcal{I}_\varepsilon(P,Q)/2)^n \leq e^{-n\mathcal{I}_\varepsilon(P,Q)/2}.$$

The same bound holds in the Q world using r^{-1}. For $\varepsilon = 0$, the ordinary likelihood-ratio test gives this bound, with any convention on sequences where both product laws vanish. Hence $n_\varepsilon \leq \lceil 2\log 4/\mathcal{I}_\varepsilon \rceil \leq 5/\mathcal{I}_\varepsilon$. Conversely, any test whose two errors are at most $1/4$ for the pair (P^*, Q^*) requires product TV at least $1/2$. Since

$$\mathrm{TV}((P^*)^{\otimes n}, (Q^*)^{\otimes n}) \leq h((P^*)^{\otimes n}, (Q^*)^{\otimes n}) \leq \sqrt{n\mathcal{I}_\varepsilon(P,Q)},$$

we obtain $n_\varepsilon \geq 1/(4\mathcal{I}_\varepsilon)$. The same product bound holds for the image of any single input contamination pair; it does not need the encoded pair to be least favorable.

Clipped coordinates. On sign s, the contrast of (16) is $s\min(|z|, t_s)$. Its mean $m_i^* = (P_i^* + Q_i^*)/2$ relative to m_i equals $1-\varepsilon$ below the clipping threshold and $(1-\varepsilon)(1+|z_i|)/(1+t_s)$ above it. Keeping the common factor gives

$$(1-\varepsilon)m_i \leq m_i^* \leq 2(1-\varepsilon)m_i.$$

For any contrast $a \in [-1, 1]$, $a^2 \le 2(1 - \sqrt{1 - a^2}) \le 2a^2$. Let

$$G_{\mathrm{raw}} = \sum_{s \in \{-1,1\}} \mathbb{E}_m[\min(|z|, t_s)^2 \mathbf{1}\{\operatorname{sign} z = s\}].$$

It follows that

$$(1 - \varepsilon) G_{\mathrm{raw}} \le \mathcal{I}_\varepsilon(P, Q) \le 4(1 - \varepsilon) G_{\mathrm{raw}}. \tag{17}$$

At zero contamination choose thresholds one for this comparison.

B Proof of the clipped-moment lemma

The law of A is conditional on one nonzero sign and normalized to one, with finite support in $(0, 1]$. Write $j = \mathbb{E}A^2$, $a_0 = \sqrt{j/2}$, and $M_0 = \Pr(A \ge a_0)$. Since $\mathbb{E}[A^2 \mathbf{1}\{A < a_0\}] \le j/2$ and $A^2 \le 1$, we have $M_0 \ge j/2$. Counting bins is then straightforward. There are at most $\lceil \log_q(1/a_0) \rceil$ high bins. Empty high intervals are omitted. Each complete lower bin multiplies the mass assigned so far by q, starting from M_0, so at most $\lceil \log_q(1/M_0) \rceil$ lower bins are needed, counting the last partial one.

When an atom is split, fractions of its mass go to adjacent bins, and conditional on the atom these fractions are the channel probabilities. By construction, every value in a later bin is at most every value in an earlier bin, apart from shared endpoints. Let $\bar{A}$ be the mean of the bin to which the mass is assigned.

First clipped moments. The upper inequality $F(t) \le \bar{F}(t)$ follows from concavity. For the lower one, high bins can be handled one at a time. A high bin is contained in some $[a, qa]$, and its contribution to $\bar{F}(t/q)$ never exceeds its contribution to $F(t)$. Per unit mass, the raw contribution is t when $t \le a$ and the bin mean when $t \ge qa$, and when $a < t < qa$ every raw value is at least t/q.

Lower bins need not lie in a factor-q range. At most one lower bin strictly straddles t. Let m_b be its mass and $M_{<b}$ the mass before it, all of whose values are at least t. Then

$$\text{excess of the straddling bin} \le m_b\, t/q \le (q-1) M_{<b}\, t/q = \underbrace{M_{<b}\, t(1 - 1/q)}_{\text{deficit of earlier bins}},$$

using the capacity rule $m_b \le (q - 1)M_{<b}$. Bins entirely below t contribute no positive excess, and thresholds in gaps or at endpoints obey the same inequality. Thus $\bar{F}(t/q) \le F(t)$ for every $t \ge 0$.

Second clipped moments. In a high bin $A \le q\bar{A}$, so $\min(A, t)^2 \le q^2 \min(\bar{A}, t/q)^2$. Let lower bin k have mass m_k and mean $\bar{A}_k$, and let a_1 be the largest value in the first lower bin, so that $m_1 \le (q - 1)M_0$. High values and high-bin means are at least a_1. Writing $\bar{G}_{\mathrm{high}}(t/q)$ for the high part's contribution to $\bar{G}(t/q)$, the first lower bin contributes to $G(t)$ at most

$$m_1 \min(a_1, t)^2 \le (q-1) q^2 M_0 \min(a_1, t/q)^2 \le (q-1) q^2\, \bar{G}_{\mathrm{high}}(t/q),$$

and with the high part's own factor q^2 the total factor is $(q - 1)q^2 + q^2 = q^3$. Lower bin $k + 1$ has $m_{k+1} \le q m_k$ and values at most $\bar{A}_k$, so it contributes to $G(t)$ at most

$$m_{k+1} \min(\bar{A}_k, t)^2 \le q^3\, m_k \min(\bar{A}_k, t/q)^2,$$

which is q^3 times bin k's contribution to $\bar{G}(t/q)$. These targets are disjoint. Summing gives $G(t) \le q^3 \bar{G}(t/q)$.

C Proof of the reusable representation theorem

Let $m_s = \Pr_m(\operatorname{sign} z = s)$ and $j_s = \mathbb{E}_m[z^2 \mid \operatorname{sign} z = s]$. Both signs have positive mass and absolute first moment $V/2$. Hence

$$j_s \ge \left(\frac{V}{2m_s} \right)^2 \ge J^2/4.$$

For either sign, the bin count is at most

$$2 + \frac{\log(1/a_0) + \log(1/M_0)}{\log q} \le 2 + \frac{\frac{3}{2}\log(2/j_s)}{\log q}.$$

Summing over signs and allowing one neutral output gives $K_q \le 5 + [3\log 8 + 6\log(1/J)]/\log q$.

Within a signed bin, the encoded mean mass is its allocated nominal mass, and its contrast is the sign times the conditional mean of $|z|$. Both signed first moments are therefore preserved, giving $\mathrm{TV}(T_qP, T_qQ) = V$.

For $0 < \varepsilon < \varepsilon_*$, put $k = \varepsilon/(1-\varepsilon)$. Raw and compressed clipping thresholds solve (7) with the same right side. At t_s/q, the compressed left side satisfies

$$\bar{F}_s(t_s/q) + \frac{kt_s}{2q} \le F_s(t_s) + \frac{kt_s}{2} = \frac{V-k}{2}.$$

At t_s, it is at least $(V-k)/2$. Strict monotonicity gives $t_s/q \le \bar{t}_s \le t_s$. Thus

$$\mathcal{I}_\varepsilon(T_qP, T_qQ) \ge (1-\varepsilon)\sum_s \bar{G}_s(\bar{t}_s) \ge \frac{1-\varepsilon}{q^3}\sum_s G_s(t_s) \ge \frac{\mathcal{I}_\varepsilon(P,Q)}{4q^3}.$$

At $\varepsilon = 0$, $G(1) \le q^3\bar{G}(1/q) \le q^3\bar{G}(1)$ and $J \le h^2 \le 2J$ give the stronger factor $1/(2q^3)$. The full-output Huber test is valid under input contamination. Appendix A then gives

$$n_\varepsilon^{T_q} \le \frac{5}{\mathcal{I}_\varepsilon(T_qP, T_qQ)} \le \frac{20q^3}{\mathcal{I}_\varepsilon(P,Q)} \le 80q^3 n_\varepsilon.$$

If the construction exceeds d outputs, identity satisfies both bounds. An observer of the raw data can simulate the encoder and its decoder, so $R_T \ge 1$ for every channel.

C.1 A PRESCRIBED OUTPUT BUDGET

For $K \ge 10$, choose

$$q = \max\left\{2, \exp\left(\frac{3\log 8 + 6\log(1/J)}{K-5}\right)\right\}.$$

Theorem 1 uses at most K outputs. Since $K - 5 \ge K/2$, its overhead satisfies $\log(1 + R_{T_q}) \le C(1 + \log(1/J)/K)$.

For smaller budgets, use the true sign bit S, assigning neutral inputs to either side. It preserves nominal TV. Set $\Delta = V - k$ and $D = \Delta/(1+k)$ for $k = \varepsilon/(1-\varepsilon) < V$. Every pair in the binary output Huber classes has TV at least D, so $\mathcal{I}_\varepsilon(SP, SQ) \ge D^2$. If $k < V/2$, then $D \ge V/4$ and

$$\frac{\mathcal{I}_\varepsilon(P,Q)}{\mathcal{I}_\varepsilon(SP,SQ)} \le \frac{32J}{V^2} \le \frac{32}{J}.$$

If $k \ge V/2$, equation (7) gives $t_s \le \Delta/k$ and $F_s(t_s) \le \Delta/2$. Therefore $\sum_s G_s(t_s) \le \Delta^2/k$, and (17) gives

$$\frac{\mathcal{I}_\varepsilon(P,Q)}{\mathcal{I}_\varepsilon(SP,SQ)} \le \frac{4(1+k)}{k} \le \frac{16}{V} \le \frac{16}{J}.$$

The sample-complexity bounds imply $R_S \le 640/J$. This proves (8) for $2 \le K \le 9$ after adjusting the universal constant.

D PROOF OF THE ALPHABET LOWER BOUND

All comparison constants in this section may depend on the fixed $\alpha \in (1,2)$, but not on L, the channel, or its outputs. Write $w_j = c_L 2^{-\alpha j}$, with c_L bounded above and below by positive constants. The geometric sums imply

$$V = \delta\sum_{j=0}^{L} w_j 2^j \asymp \delta, \qquad J = \delta^2\sum_{j=0}^{L} w_j 2^{2j} \asymp \delta^\alpha.$$

The two signed conditional laws are identical. For a raw threshold $t \in (0, 1/2]$, its common contamination parameter is determined by

$$k(t) = \frac{\mathbb{E}(A-t)_+}{1+t}, \qquad \varepsilon(t) = \frac{k(t)}{1+k(t)}. \tag{18}$$

This is continuous, decreases from V to zero as t goes from zero to $1/2$, and parameterizes the positive feasible range. At the clean endpoint we may choose $t = 1/2$.

Let $G(t) = \mathbb{E}\min(A,t)^2$. For $t = \delta 2^\ell$, $1 \le \ell \le L-1$, summing separately below and above ℓ gives

$$G(t) \asymp \delta^2 2^{(2-\alpha)\ell}, \qquad \mathcal{I}_{\varepsilon(t)}(P,Q) \asymp G(t). \tag{19}$$

More generally, $G(t) \asymp \delta^\alpha t^{2-\alpha}$ for $\delta \le t \le 1/2$, and $G(t) = t^2$ for $t \le \delta$. These estimates include the endpoints by the same geometric sums.

D.1 A separately selected bit at every level

The conditional clipping equation can also be written as $\mathbb{E}\min(A,t)+kt = V-k$. The nominal sign bit has probabilities $(1+V)/2$ and $(1-V)/2$ under the two hypotheses. Its closest contaminated Bernoulli probabilities differ by

$$D = \frac{V-k}{1+k} = \frac{\mathbb{E}\min(A,t)+kt}{1+k} > 0. \tag{20}$$

When $t \le \delta$, $D = t$, so its squared Hellinger information is comparable to $G(t) = t^2$. Fix an integer h depending only on α. For $\delta < t < \delta 2^h$, $D \ge \delta/2$ while $G(t) \le \delta^2 2^{2h}$. The sign bit remains within a constant factor on this entire range.

For larger t, put $\ell = \lfloor \log_2(t/\delta) \rfloor$ and $j = \ell - h \ge 0$. Use the event bit $\mathbf{1}\{B = +1, U = 2^j\}$. The geometric tail gives

$$k(t) \le C_\alpha \delta 2^{(1-\alpha)\ell}, \qquad w_j A_j = c_L \delta 2^{(1-\alpha)\ell} 2^{(\alpha-1)h}.$$

Choose h large enough that $w_j A_j \ge 2k(t)$ for all relevant ℓ, L. The event's nominal gap is $w_j A_j$. Its closest Huber Bernoulli probabilities $\bar{p}, \bar{q}$ satisfy

$$\bar{p} - \bar{q} \ge \frac{w_j A_j - k}{1+k} \ge \frac{w_j A_j}{4}, \qquad \bar{p} + \bar{q} \le w_j + \varepsilon \le \tfrac{5}{4} w_j,$$

using $\varepsilon \le k \le w_j A_j/2$ and $A_j \le 1/2$. Keeping only the event coordinate in the rationalized Hellinger formula, their squared Hellinger distance is at least a constant times $(\bar{p}-\bar{q})^2/(\bar{p}+\bar{q})$, hence at least a constant times $w_j A_j^2$. Since

$$w_j A_j^2 \asymp \delta^2 2^{(2-\alpha)(\ell-h)} \asymp G(t),$$

this bit keeps the raw information up to a constant factor, and by data processing it keeps at most the raw information. For the finitely many small $L < h$ the sign-bit argument already gives a uniform constant. At $\varepsilon = 0$ the event at $j = L$ is another clean-optimal choice. This proves the supplied-level claim on the whole feasible range.

D.2 One channel cannot allocate a bit to every scale

Fix an arbitrary memoryless Markov channel T. For an output y, let $a_y = \mathbb{E}_m T(y \mid X)$. Omit outputs with $a_y = 0$. We first bound $\mathbb{E}_m[T(y \mid X)U]$. Since $0 \le T(y \mid X) \le 1$ and the law of U satisfies $\Pr(U \ge u) \le C_\alpha u^{-\alpha}$ for $u \ge 1$, layer cake yields

$$\mathbb{E}_m[T(y \mid X)U] \le a_y + \int_1^\infty \min\{a_y, C_\alpha u^{-\alpha}\}\, du \tag{21}$$

$$\le C_\alpha a_y^{1-1/\alpha}. \tag{22}$$

For a raw least-favorable pair at scale t_ℓ, write its average density as $m^* = c_\ell(A) m$, with $1/2 \le c_\ell \le 2$ and contrast $B\min(A, t_\ell)$. Let $\bar{m}_y$ and d_y be the mean and half-difference of its two encoded output probabilities. Then

$$\bar{m}_y \ge a_y/2, \qquad |d_y| \le 2\min\{t_\ell a_y, \delta C_\alpha a_y^{1-1/\alpha}\}.$$

Rationalizing Hellinger on each output gives an upper bound $2d_y^2/\bar{m}_y$. Thus, using $s_y = a_y^{-1/\alpha}$ and (19),

$$\frac{h^2(TP_\ell^*, TQ_\ell^*)}{\mathcal{I}_{\varepsilon(t_\ell)}(P,Q)} \le C_\alpha \sum_y (2^\ell/s_y)^\alpha \min\{1, (s_y/2^\ell)^2\}. \tag{23}$$

Write the summand in (23) as $f_y(\ell)$. The image of each raw least-favorable pair is an admissible input attack. Appendix A therefore gives

$$\frac{1}{4h^2(TP_\ell^*, TQ_\ell^*)} \le n^T_{\varepsilon(t_\ell)} \le \frac{5R_T}{\mathcal{I}_{\varepsilon(t_\ell)}(P,Q)}.$$

For finite R_T, the information ratio is at least $1/(20R_T)$, so

$$1 \le A_\alpha R_T \sum_y f_y(\ell) \quad \Longrightarrow \quad 1 \le \sum_y \min\{1, A_\alpha R_T f_y(\ell)\},$$

where we take $A_\alpha \ge 1$. The implication holds for any nonnegative summands where either one term reaches one, or capping leaves the sum unchanged. Put $b_y = \log_2 s_y$. Then

$$f_y(\ell) = \begin{cases} 2^{\alpha(\ell - b_y)}, & \ell \le b_y, \\ 2^{-(2-\alpha)(\ell - b_y)}, & \ell \ge b_y. \end{cases}$$

The interval on which $A_\alpha R_T f_y(\ell) \ge 1$ has length

$$\left(\frac{1}{\alpha} + \frac{1}{2-\alpha}\right) \log_2(A_\alpha R_T).$$

Outside it, two geometric tails have bounded sum, uniformly in b_y. Consequently

$$\sum_{\ell \in \mathbb{Z}} \min\{1, A_\alpha R_T f_y(\ell)\} \le B_\alpha(1 + \log R_T).$$

Summing first over the $L-1$ interior levels and then over outputs gives $L - 1 \le B_\alpha K(1 + \log R_T)$. Since $R_T \ge 1$, this proves $K \log(1 + R_T) \ge c_\alpha L$ for $L \ge 2$. The remaining bounded values of L follow after reducing the constant. Infinite overhead satisfies the bound as well. Combining this result with Appendix C.1 and $\log(1/J_L) \asymp_\alpha L$ proves Theorem 3.

D.3 A FIXED ONE-BIT SCHEDULE

Let S be the sign bit and $E_j = \mathbf{1}\{B = +1, U = 2^j\}$ for $0 \le j \le L$. Appendix D shows that for every feasible ε, at least one member B_ε of this $M = L + 2$ element library has $n_\varepsilon^{B_\varepsilon} \le C_\alpha n_\varepsilon$. Fix the round-robin sequence $S, E_0, \ldots, E_L, S, E_0, \ldots$ before ε is supplied. The decoder knows the index of each observation. Among the first $M n_\varepsilon^{B_\varepsilon}$ observations, exactly $n_\varepsilon^{B_\varepsilon}$ use B_ε. Their encoded values are iid under each allowed raw law, so its robust test has the required error bound. Hence

$$n_\varepsilon^{\text{schedule}} \le (L+2) C_\alpha n_\varepsilon.$$

Each observation transmits one bit. The index identifies the mapping without a per-symbol encoder label. This protocol uses different channels at different indices and is outside the infimum defining $\mathcal{R}_K$.

E WHY THE CLEAN CUTOFF LOSES ROBUST INFORMATION

Let $q = 1 - \alpha/2 > 0$ and choose $j_0 = \lfloor qL \rfloor + O(1)$. Consider the deterministic partition R that retains every signed raw symbol with $j \ge j_0$ and merges all $j < j_0$ into one output for each sign. Because no signs are mixed, $\mathrm{TV}(RP, RQ) = V$. The upper geometric levels contain a constant fraction of J and are unchanged, so $h^2(RP, RQ) \ge c_\alpha h^2(P,Q)$.

Fix $\ell = \lfloor \gamma L \rfloor$ for $0 < \gamma < q$ and use the actual raw least-favorable pair at $t_\ell = \delta 2^\ell$. Each low-sign output has mean mass bounded below by a constant. Its absolute probability difference is $O(\delta)$, since the uncompressed signed absolute first moment is $O(\delta)$ and $m^*/m \le 2$. Its Hellinger contribution is $O(\delta^2)$. The retained high outputs have total nominal mass $O(2^{-\alpha j_0})$ and clipped

contrasts bounded by t_ℓ, so their contribution is $O(\delta^2 2^{2\ell-\alpha j_0})$. Dividing by (19) proves (11). The product lower bound in Appendix A then makes the robust sample overhead diverge, even when the decoder knows the exact images of the input classes rather than the larger output neighborhoods.

For equal class priors, the posterior divergence of a signed amplitude is

$$D(A) = \mathrm{KL}\left(\mathrm{Bern}\left(\frac{1+A}{2}\right) \middle\| \mathrm{Bern}(1/2)\right).$$

On $0 \le A \le 1/2$, $D(A) \asymp A^2$, and the clean mutual information $\mathbb{E}_m D(A)$ is comparable to J. A cutoff at a fixed positive multiple of that information lies at $j_0 = qL + O(1)$. The partition R refines the cutoff construction by preserving every upper raw value and the lower signs. By data processing, the coarser cutoff construction has at least the sample cost of R, proving Proposition 4.

F Geometry of encoded input contamination

Write $\Delta = P - Q$ and $k = \varepsilon/(1-\varepsilon)$. If the encoded input classes intersect, then $T\Delta = kT(R_Q - R_P)$ for some input laws R_P, R_Q. The vector $v = k(R_Q - R_P)$ is feasible in (12) and has $\|v\|_1/2 \le k$. Conversely, if such a feasible v has norm at most $2k$, column-stochasticity implies $\sum_i v_i = 0$. Its positive and negative parts have the same mass $a \le k$. For $k > 0$, choose any input law S and put

$$R_Q = v_+/k + (1 - a/k)S, \qquad R_P = v_-/k + (1 - a/k)S.$$

Their images create an intersection. At $k = 0$ both statements reduce directly to $T\Delta = 0$. Disjoint encoded classes are compact and convex. A strict linear separator and concentration of its bounded iid scores give a finite uniform test. In this finite model, then, a test exists exactly when the classes do not intersect.

The dual of the finite norm minimization is

$$\max_u \langle u, T\Delta\rangle \quad \text{subject to} \quad \|T^\top u\|_\infty \le 1/2.$$

Feasibility and boundedness give equality of primal and dual values. The all-ones vector belongs to $\mathrm{row}(T)$ because the columns sum to one. Translate $T^\top u$ by $\mathbf{1}/2$. Since $\sum_i \Delta_i = 0$, this translation does not change the objective and yields the cube dual in (12). It also proves $\mathrm{TV}(TP, TQ) \le \kappa_T \le V$. The left inequality follows directly from $\|Tv\|_1 \le \|v\|_1$ for a stochastic matrix.

If each $\Delta_i \ne 0$, the unique maximizer of $\langle f, \Delta\rangle$ over the full cube is $s = \mathbf{1}\{\Delta > 0\}$, with value V. The intersection of the cube with the row space is compact, so it achieves value V precisely when it contains s. If it does not, $\kappa_T < V$, and every contamination parameter strictly between $\kappa_T/(1+\kappa_T)$ and $V/(1+V)$ is raw-feasible but encoded-infeasible.

For an invertible noisy two-by-two stochastic matrix, the row space is all of $\mathbb{R}^2$. Hence $\kappa_T = V$ even when output TV shrinks.

G Training obstruction and fitted-test guarantee

G.1 The constant-information sign family

For $r \ge 1$ and $0 < \eta \le 1/4$, let $a_\eta = \sqrt{1/4 - \eta^2}$. Under m, give each anchor mass $1/4$, and each of the $2r$ weak symbols mass $1/(4r)$. For $\theta \in \{-1, 1\}^r$ let

$$z_\theta = (a_\eta, -a_\eta, \theta_1\eta, -\theta_1\eta, \ldots, \theta_r\eta, -\theta_r\eta), \qquad P_\theta = m(1 + z_\theta), \quad Q_\theta = m(1 - z_\theta).$$

The laws are strictly positive and normalize, with $J = \frac{1}{2}a_\eta^2 + \frac{1}{2}\eta^2 = 1/8$ and $V = (a_\eta + \eta)/2$. Their clean sample complexity is bounded above and below by constants independent of η, r.

For a fixed channel T, let $\mathcal{L}(T)$ be the set of sign patterns whose full sign indicators belong to $\mathrm{row}(T)$. A linear or affine subspace of dimension at most K contains at most 2^K vertices of a Boolean cube: some set of at most K coordinate projections is injective on its direction space, and each cube vertex projects to a distinct binary string. The map from θ to the full sign indicator is injective, so $|\mathcal{L}(T)| \le \min(2^r, 2^K)$. This dimension bound applies to channels with arbitrary real entries.

Put a uniform prior on θ, and let S contain the N observations from each class. A reference pair keeps the anchors fixed and sets all weak contrasts to zero. The chi-squared upper bound on KL gives at most $\eta^2/2$ per observation in each class. Independence and the reference-distribution bound on mutual information therefore imply

$$I(\theta; S) \le N\eta^2.$$

Learner randomization is independent of θ, so $I(\theta; T) \le I(\theta; S)$. If all-level feasibility succeeds with probability at least $1 - \beta$, then by Proposition 5, $\theta \in \mathcal{L}(T)$ with at least that probability. Let E indicate this event. Conditional entropy satisfies

$$H(\theta \mid T) \le H(E \mid T) + K \log 2 + \beta r \log 2 \le \log 2 + K \log 2 + \beta r \log 2.$$

Subtract from $H(\theta) = r \log 2$ to prove Theorem 6. The uniform prior is only a device: since

$$\sup_\theta \Pr_\theta(\text{failure}) \ge \mathbb{E}_\theta \Pr_\theta(\text{failure}),$$

a guarantee that holds for every pair in the family also holds on average over θ, and the argument uses nothing more. Taking r sufficiently larger than K and letting $\eta \to 0$ then yields the absence, claimed in Theorem 6, of a finite pilot bound depending only on J. The identity map with $K = d$ escapes this obstruction.

G.2 Continuity at the same contamination bound

Let $d_\varepsilon(P, Q) = \sqrt{\mathcal{I}_\varepsilon(P, Q)}$. Joint convexity of squared Hellinger gives, for any shared contaminant R,

$$h((1-\varepsilon)P + \varepsilon R, (1-\varepsilon)\widehat{P} + \varepsilon R) \le \sqrt{1-\varepsilon}\, h(P, \widehat{P}).$$

Match a contaminant under each hypothesis and apply the triangle inequality. Taking infima in both directions yields

$$|d_\varepsilon(P, Q) - d_\varepsilon(\widehat{P}, \widehat{Q})| \le \sqrt{1-\varepsilon}\{h(P, \widehat{P}) + h(Q, \widehat{Q})\}. \tag{24}$$

On the pilot event, Theorem 1 with $q = 2$ and (24) give

$$\widehat{d}_\varepsilon \ge \frac{d_\varepsilon(\widehat{P}, \widehat{Q})}{\sqrt{32}} \ge \frac{d_\varepsilon(P, Q) - \rho}{\sqrt{32}}. \tag{25}$$

If the estimated laws coincide, the right side is nonpositive by (24). In that case we use a constant map and issue no certificate. If estimated classes overlap, no test is certified. Otherwise use their least-favorable likelihood-ratio test with $n = \lceil 8/\widehat{d}_\varepsilon^2 \rceil$. Under either estimated class, its error is at most $\exp(-n\widehat{d}_\varepsilon^2/2) \le e^{-4}$.

For any true contaminated input law, form the corresponding estimated law with the same contaminant. Hellinger contracts under T, so the one-observation distance between their encoded laws is at most ρ. Their product TV is at most $\sqrt{n}\rho$. The actual error of this specific fitted decoder is consequently bounded by

$$\min\{1, \exp(-n\widehat{d}_\varepsilon^2/2) + \sqrt{n}\rho\}. \tag{26}$$

Any tie-breaking rule works when both estimated product likelihoods are zero. Ours picks P. Since $\widehat{d}_\varepsilon^2 \le 2$, $n \le 10/\widehat{d}_\varepsilon^2$. If $\widehat{d}_\varepsilon \ge 14\rho$, the bound is at most $e^{-4} + \sqrt{10}/14 < 1/4$. Equation (25) proves eligibility when $d_\varepsilon(P, Q) \ge (1 + 14\sqrt{32})\rho$, and the stated threshold 81ρ is this constant rounded up. On that range $\widehat{d}_\varepsilon \ge (14/81) d_\varepsilon(P, Q)$, which gives the claimed sample order, and the nominal triangle inequality keeps $h(\widehat{P}, \widehat{Q})$ comparable to $h(P, Q)$, so the fitted map still has an $O(1 + \log(1/J))$ alphabet bound.

All of the above conditions on one pilot event, independent of the fresh test sample, so no union bound over ε is needed. The unconditional probability of an incorrect *certified* decision is at most $\beta + 1/4$. For deterministic ρ, any eligible level fixed in advance is certified with probability at least $1 - \beta$. Outside the eligible range, the bound and sample size must be chosen before the deployment observations are seen.

G.3 An elementary pilot confidence radius

For the empirical law $\widehat{P}$ from N iid observations,

$$\mathbb{E}h^2(\widehat{P}, P) \le \mathbb{E}\chi^2(\widehat{P}\|P) = \frac{d_P - 1}{N} \le \frac{d-1}{N},$$

where d_P is the number of positive-mass symbols. Zero-mass symbols are never observed and may be omitted. The same holds for Q. By $(a+b)^2 \le 2(a^2+b^2)$ and Markov's inequality,

$$\Pr\left\{ h(P,\widehat{P}) + h(Q,\widehat{Q}) > \sqrt{\frac{4(d-1)}{N\beta}} \right\} \le \beta.$$

The learning study uses this ambient-alphabet radius. Its dependence on d and the Markov bound make it conservative, motivating the direct calibration of fitted scores below.

G.4 Independent monotone calibration

Fix arbitrary training data, and with it T. Order $[K]$ by $\widehat{p}_y/\widehat{q}_y$, with ratio one for jointly zero probabilities and ties broken by symbol index. Given training, the M encoded calibration observations in each class are iid, because their inputs and private coins are independent, so DKW with failure probability $\beta/2$ per class (Massart, 1990; Shi et al., 2026) gives the CDF event of the main text.

Lemma 9 (Monotone deviation)**.** *Let p, v be laws on $[K]$, ordered as above, with $D_j = \sum_{i\le j}(p_i - v_i)$ and $\max_j |D_j| \le \delta_M$. Then for every bounded f that is monotone in this order,*

$$\mathbb{E}_p f - \mathbb{E}_v f = \sum_{j<K} D_j(f_j - f_{j+1}) \le \delta_M(\max_j f_j - \min_j f_j). \tag{27}$$

The identity is summation by parts with $D_K = 0$, and the increments of a monotone sequence telescope to its range. On the CDF event the hypothesis holds with $p = TP$ or TQ and v the corresponding calibration law. It does not involve f, so (27) holds simultaneously for all such f, including f chosen after seeing the calibration data. The cap in (15) is $\mathbb{E}_p f \le \max f$.

For a fitted nonoverlapping Huber pair at $\varepsilon > 0$,

$$\frac{\widehat{p}^*_\varepsilon(y)}{\widehat{q}^*_\varepsilon(y)} = \min\Big\{c_+, \max\Big\{c_-, \frac{\widehat{p}_y}{\widehat{q}_y}\Big\}\Big\}, \qquad 0 < c_- < 1 < c_+,$$

so g_ε is finite and nondecreasing in the fixed order for every supplied ε, and $g_\varepsilon = 0$ on jointly zero nominal symbols. At $\varepsilon = 0$ a one-sided nominal zero can make g_ε infinite. The method then abstains, as it does when the estimated classes overlap. For every finite-score candidate and $\lambda > 0$, $f = e^{\pm\lambda g_\varepsilon}$ is monotone in the fixed order and Lemma 9 applies. The argument does not require the true likelihood ratio to follow this order, and it may not.

Under an actual P-world contaminant W,

$$\begin{aligned}\mathbb{E}_{T((1-\varepsilon)P+\varepsilon W)} e^{-\lambda g_\varepsilon} &= (1-\varepsilon)\mathbb{E}_{TP} e^{-\lambda g_\varepsilon} + \varepsilon \sum_x W_x \sum_y T(y \mid x) e^{-\lambda g_\varepsilon(y)} \\ &\le (1-\varepsilon)C_P + \varepsilon \max_x \sum_y T(y \mid x) e^{-\lambda g_\varepsilon(y)} = B_P,\end{aligned}$$

and the Q-world bound is the same with $+\lambda$. For any vector u on the inputs, the maximum of $\sum_x W_x u_x$ over the input simplex is $\max_x u_x$, so the second term is exactly the worst-input contaminant moment, not merely a bound on it. For n independent deployment observations, Markov's inequality gives

$$\Pr_P\Big(\sum_{i=1}^n g_\varepsilon(Y_i) < 0\Big) \le B_P^n, \qquad \Pr_Q\Big(\sum_{i=1}^n g_\varepsilon(Y_i) \ge 0\Big) \le B_Q^n,$$

with λ allowed to differ between the two, and by the definition of b and n both are at most $1/4$. Conditional on the training data, the CDF event has probability at least $1 - \beta$, and on it these

inequalities hold for all ε, λ and the prescribed n simultaneously. Hence ε, λ and n may depend on the training and calibration data, with no further union bound, provided they are fixed before any deployment observation arrives. Integrating over training and calibration, under either hypothesis,

$$\Pr(\text{certify and decide incorrectly}) \le \beta + 1/4.$$

The bound is on the joint event: controlling $\Pr(\text{incorrect} \mid \text{certify})$ would also need a lower bound on $\Pr(\text{certify})$. The proof assumes a fixed deployment sample size and iid Huber observations. Unlike the radius of Appendix G.3, δ_M has no explicit alphabet factor, but through the moment bound the certification and deployment cost still depend on the range of the fitted score, the quality of training and the separation. The method can abstain at raw-feasible levels and can need more than the raw-optimal order of samples.

H Why two separate future bounds change the alphabet cost

With separate contamination bounds under the two hypotheses, preserving feasibility can require one output per distinct likelihood ratio.

Proposition 10 (Separate bounds)**.** *Let P, Q be strictly positive with L distinct values of P_i/Q_i. If a channel T with K outputs preserves feasibility for every pair $(\varepsilon_P, \varepsilon_Q)$ for which the raw Huber classes are disjoint, then $K \ge L$. Merging symbols with equal P_i/Q_i gives such a channel with $K = L$.*

Necessity uses the row-space characterization of Proposition 5. Choose τ strictly between two adjacent likelihood ratios and set

$$M(\tau) = \sum_i \max(P_i, \tau Q_i), \quad a = 1/M(\tau), \quad b = \tau/M(\tau), \quad w = aP - bQ.$$

Because τ is interior, $M(\tau) > \max(1, \tau)$, so $a, b \in (0, 1)$. The bounds $\varepsilon_P = 1 - a$, $\varepsilon_Q = 1 - b$ put the raw classes at overlap:

$$\tfrac{1}{2}\|w\|_1 = 1 - (a + b)/2 = (\varepsilon_P + \varepsilon_Q)/2.$$

The encoded input classes intersect exactly when $\min_{Tv=Tw} \|v\|_1/2 \le (\varepsilon_P + \varepsilon_Q)/2$. Indeed, $\sum_i v_i = a - b = \varepsilon_Q - \varepsilon_P$, so this inequality is equivalent to the positive and negative parts having masses at most $\varepsilon_Q, \varepsilon_P$ respectively. Add a common remaining subprobability mass to construct the two contaminants, as in Appendix F.

The cube-duality argument shows that the minimum norm is strictly less than $\|w\|_1/2$ unless $\mathbf{1}\{P_i > \tau Q_i\}$ belongs to $\mathrm{row}(T)$. To verify the translation when $\sum_i w_i \ne 0$, the dual value is $\max_{f \in \mathrm{row}(T) \cap [0,1]^d} \langle f, w \rangle - \frac{1}{2}\sum_i w_i$. The constant does not affect its unique cube maximizer. Now suppose the threshold indicator is absent, and increase both a and b by a factor $1 + u$ for sufficiently small $u > 0$. The raw norm then exceeds the remaining contamination budget, while, by continuity and the strict gap just established, the encoded norm still falls below it. The raw classes are disjoint but the encoded ones overlap, a contradiction.

So each of the $L - 1$ nested likelihood-ratio threshold indicators lies in the row space. Together with the all-ones vector they are linearly independent, which gives $\mathrm{rank}(T) \ge L$ and hence $K \ge L$.

For sufficiency, let T merge each group g of equal ratios, write $P(g) = \sum_{i \in g} P_i$ and $Q(g) = \sum_{i \in g} Q_i$, and let $\pi_g(i) = P_i/P(g) = Q_i/Q(g)$ for $i \in g$. If an output law r satisfies $r \ge (1 - \varepsilon_P)TP$ and $r \ge (1 - \varepsilon_Q)TQ$, then $\tilde{r}_i = r_g \pi_g(i)$ is an input law with $\tilde{r} \ge (1 - \varepsilon_P)P$ and $\tilde{r} \ge (1 - \varepsilon_Q)Q$. So the encoded classes intersect only if the raw classes do. Since a finite pair can have constant J and arbitrarily large L, this stronger kind of uniformity behaves differently from Theorem 1 where now under separate bounds, J alone no longer controls the alphabet.

I Experimental specifications and complete summary tables

I.1 Sparse construction and numerical evaluation

The implementation stores only nonzero channel entries. For $q = 2$ and one sign s it runs as follows. The other sign is processed separately in the same way. Masses need not be normalized within a sign, since every bin capacity scales with them.

1. Compute the sign mass m_s, second moment $j_s = m_s^{-1} \sum_{i:\,\mathrm{sign}\, z_i = s} m_i z_i^2$, and $a_0 = \sqrt{j_s/2}$.
2. Assign each $|z_i| \geq a_0$ to its dyadic interval: take the interval index from the binary exponent of $|z_i|/a_0$, then check the two endpoints. A value exactly on an interior boundary goes to the interval on its right, and the final interval includes 1. Keep the occupied intervals, in increasing order.
3. Set $H = \sum_{|z_i| \geq a_0} m_i$. Sort the remaining atoms by decreasing $|z_i|$, ties by input index, and open a lower bin y with capacity H.
4. Move $\min(u_i, H - b_y)$ from the current atom i, with remaining mass u_i (initially m_i), to bin y, where b_y is the mass already in y. If $u_i = 0$, go to the next atom. If $b_y = H$, append y, set $H \leftarrow H + b_y$, and open bin $y + 1$ with capacity H. The last bin is appended only if $b_y > 0$.
5. With b_{yi} the mass moved from atom i to bin y, set $T(y \mid i) = b_{yi}/m_i$. All i with $z_i = 0$, including those with $m_i = 0$, share one additional output. If more than d outputs have been constructed, use the identity channel instead.

If $P = Q$ the implementation returns a constant map. Otherwise both nonzero signs have positive mass, so every initial capacity H is positive. An atom i may be split over several successive mass bins, and column i of T then holds the shares b_{yi}/m_i. Equal-amplitude atoms may be visited in any fixed order. The moment argument does not depend on it. With K_c the number of constructed outputs before fallback, each move in step 4 exhausts an atom or a bin, so

$$\#\{(y, i) : T(y \mid i) > 0\} \leq d + K_c.$$

At fixed arithmetic precision, sorting the lower atoms and the occupied upper indices costs $O(d \log d)$ comparisons, allocation $O(d + K_c)$ arithmetic operations, and storage $O(d + K_c)$. The supplied sampler draws from a column with r_i nonzero entries by scanning its cumulative probabilities, in $O(r_i)$ time. Precomputed cumulative sums would give $O(\log r_i)$ per lookup.

Construction and information calculations use 80 decimal digits. Hellinger terms are computed as $(p - q)^2/(\sqrt{p} + \sqrt{q})^2$, omitting jointly zero coordinates, which avoids subtracting nearly equal square roots, and each clipping root is found on a sorted amplitude interval, where its equation is linear. We check nonnegativity, normalization, signed first moments, and channel signatures with no tolerance used to merge weak signs. The indexed and scan-based reference implementations agree to within 10^{-65} per channel entry on all 384 heterogeneous pairs and all 15 geometric pairs, before and after fallback. The endpoint and zero-support checks include a large low-amplitude atom spread over more than two outputs, and the worst information-ratio case of each heterogeneous ensemble is recomputed at 120 digits. Table 2 gives construction times and sparse entry counts.

Table 2: Encoder construction on an AMD Ryzen 9 6900HS, in seconds. Timings include channel validation in the 80-digit Python implementation. Dense rows give medians of three repetitions. Lognormal rows give the median across three inputs of those per-input medians. Each method is warmed up once and their measurement order alternates. Entries count nonzero channel coefficients.

Family	d	K	Entries	Scan	Indexed
Lognormal	128	4	128	0.010	0.011
Lognormal	512	4	512	0.039	0.041
Lognormal	2,048	4	2,048	0.163	0.176
Lognormal	8,192	4	8,192	0.699	0.712
Dense	1,026	38	1,038	0.089	0.081
Dense	2,050	74	2,074	0.229	0.188
Dense	4,098	146	4,146	0.734	0.324

Indexing helps on the dense pairs, which have many upper intervals. On the largest, with $K \times d = 146 \times 4{,}098$, the indexed version takes a median of .324 seconds and stores 4,146 entries. The four-output lognormal examples have few upper intervals. For these, indexing does not improve the measured construction time.

Proposition 11 (Alphabet reduction). *For the $q = 2$ construction and sign s, let $a_{0,s} = \sqrt{j_s/2}$ and let $M_{0,s}$ be the conditional mass in the upper intervals. Then*

$$K_c \le 1 + \sum_{s\in\{-,+\}} \left[\left\lceil \log_2 \frac{1}{a_{0,s}} \right\rceil + \left\lceil \log_2 \frac{1}{M_{0,s}} \right\rceil \right],$$

and d larger than the right side guarantees $K_c < d$.

At most $\lceil \log_2(1/a_{0,s}) \rceil$ upper bins are occupied, and after r full lower bins the allocated mass is $2^r M_{0,s}$, so $\lceil \log_2(1/M_{0,s}) \rceil$ lower bins suffice. The extra output is needed only if some $z_i = 0$, and empty upper intervals make the count smaller. Reduction occurs exactly when $d > K_c$, which depends on the occupied evidence scales and their masses. A bound in terms of J alone need not predict fallback.

I.2 Known-law grid and baselines

For a family with length L and refinement s,

$$A_j = 2^{-L-1+j/s}, \qquad w_j \propto 2^{-\alpha j/s}, \qquad j = 0, \dots, sL,$$

with an independent fair sign. Coarse cases have $s = 1$ and dense cases $s = 32$. Each case is evaluated at $\varepsilon = 0$, at $c\,\varepsilon_*$ for $c \in \{.5, .9, .99, .9999, .99999999\}$, and at the levels of (18) with $t = 2^{-L-1+\ell}$, $\ell = 1, \dots, L-1$. Levels that agree to 70 digits are counted once, which leaves 667 points in total.

The posterior-KL construction uses $I = \mathbb{E}_m D(|z|)$, a cutoff $I/4$, and $\kappa = \log 2$, with

$$b = \left\lceil \log_2 \frac{\kappa}{I/4} \right\rceil$$

geometric levels between them. Neighboring levels differ by at most a factor of two. Values above the cutoff keep their sign, and values at or below it share one output. The signed variant for an output budget K reserves three low outputs (two signs and ties), takes $b = \lfloor (K-3)/2 \rfloor$ geometric high levels on the same interval, and falls back to the sign bit when $b < 1$. Empty bins are removed. This adaptation of the posterior-KL construction never uses more outputs than our encoder.

The sign bit sends zero-contrast symbols to its positive side. The threshold bit sorts the nominal likelihood ratios and tries every strict boundary and ties are never split. It keeps the boundary whose Bernoulli pair has the largest robust squared Hellinger distance at the supplied ε. The clean bit is fitted once, at $\varepsilon = 0$. The oracle repeats the search at every supplied level. Normalization, unchanged channel signatures, data processing, TV preservation for sign-preserving maps, and the $1/32$ information inequality are checked with relative tolerance 10^{-30}. At this tolerance the information inequality holds at all 667 levels.

For a deterministic partition every output contaminant lifts to the input support, so overlap of the output classes is an attack that can be carried out on the inputs. For a randomized map the input feasibility criterion is κ_T of Proposition 5, and disjoint full-output classes remain sufficient for feasibility.

I.3 Optimal clean-MI baseline at the same output budget

The clean-MI baseline spends the same outputs on clean information. For each of the three dense symmetric pairs, we compute the clean mutual-information optimum among sign-preserving maps with exactly K outputs, K being our encoder's output count. Within a sign, write a_j for the amplitudes in increasing order and w_j for their conditional weights. For a contiguous block $[l, r)$ define

$$W_{lr} = \sum_{j=l}^{r-1} w_j, \qquad A_{lr} = \sum_{j=l}^{r-1} w_j a_j, \qquad S(l,r) = W_{lr}\,\phi(A_{lr}/W_{lr}),$$

where $\phi(x) = x\,\mathrm{atanh}(x) + \frac{1}{2}\log(1-x^2)$ is the posterior divergence to the equal-prior Bernoulli law. We evaluate $\log(1-x^2)$ as `log1p(-x*x)` to retain accuracy when x^2 is small. By the

Table 3: Coarse known-law results. Here our construction would be larger than the input ("Constructed K" is its size before fallback), so the deployed map is the identity. Minima are robust information ratios over the tested grid, and "Level-bit" is the threshold-bit oracle, re-selected at each level. Ordinary KL and the clean-selected bit reach minimum ratio zero in every row.

α	L	Input/ours K	Constructed K	Signed KL K	Signed KL min.	Level-bit min.
1.25	16	34	38	24	.38624	.28594
1.25	32	66	74	44	.03019	.28590
1.25	48	98	108	64	.00160	.28590
1.25	64	130	144	84	.0000749	.28590
1.50	16	34	40	28	.66591	.23807
1.50	32	66	76	52	.22341	.23743
1.50	48	98	112	76	.05855	.23743
1.50	64	130	148	100	.01481	.23743
1.75	16	34	40	30	.79426	.16267
1.75	32	66	74	58	.55568	.15474
1.75	48	98	110	86	.37592	.15427
1.75	64	130	144	114	.24198	.15424

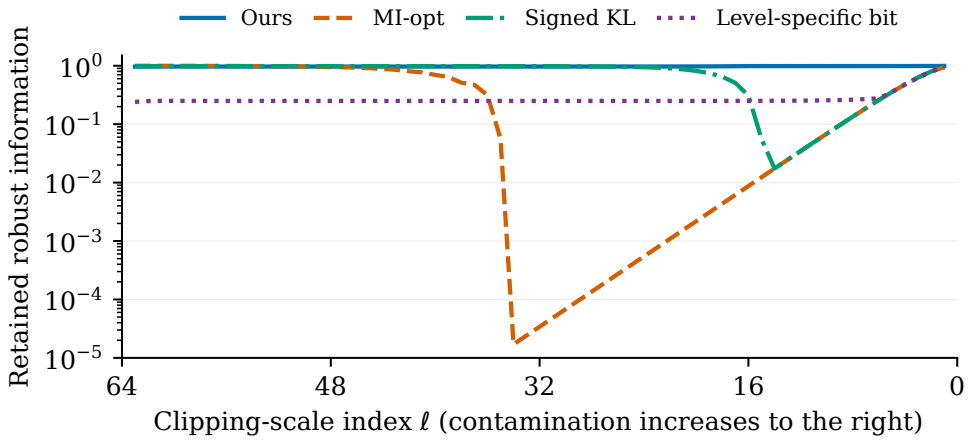


Figure 3: Robust information retained on the dense $L = 64$ pair at the 63 interior clipping scales $t_\ell = 2^{-65}2^\ell$. The grid for Table 4 also includes the clean and fixed-fraction points.

contiguous-partition characterization of Kurkoski & Yagi (2014), the optimum within a sign is obtained by

$$D_k(r) = \max_{k-1 \le l < r} \{D_{k-1}(l) + S(l, r)\}, \qquad D_1(r) = S(0, r).$$

We compute $D_k(n)$ for every $1 \le k < K$, with n the number of positive amplitudes, pick the k that maximizes $(D_k(n)+D_{K-k}(n))/2$, and backtrack both partitions. This fixes how many outputs each sign gets. The optimization chose equal splits: 19/19, 37/37 and 73/73. Restricting to deterministic partitions is without loss. Mutual information is convex in the channel. For a fixed assignment of outputs to signs, the sign-restricted channels form a polytope with deterministic extreme points, and a convex function on a polytope is maximized at an extreme point.

The exact recurrence is accelerated with monotone maximizing boundaries. Consecutive blocks obey

$$S(a, c) + S(b, d) \ge S(a, d) + S(b, c), \qquad a \le b \le c \le d.$$

Table 4: Dense known-law pairs ($\alpha = 1.5$, $d = 2(32L + 1)$). Each information value is a fraction of its raw counterpart. Clean MI and clean h^2 are at $\varepsilon = 0$. "Min. robust" is the smallest full-output $\mathcal{I}_\varepsilon$ ratio on the tested grid.

L	Encoder	Outputs	Clean MI	Clean h^2	Min. robust
16	Ours	38	0.96450	0.96392	0.96392
16	MI-opt	38	0.98517	0.98503	0.86741
16	Signed KL	34	0.97331	0.97300	0.64817
32	Ours	74	0.96421	0.96363	0.96363
32	MI-opt	74	0.99419	0.99415	0.13480
32	Signed KL	70	0.98013	0.97981	0.23123
64	Ours	146	0.96420	0.96362	0.96362
64	MI-opt	146	0.99842	0.99841	$1.707 \cdot 10^{-5}$
64	Signed KL	142	0.98156	0.98125	0.01740

Let L, M, R be the three consecutive blocks. The perspective $g(W, A) = W\phi(A/W)$ has mixed directional derivative, when the left and right blocks are added fractionally to the middle block,

$$\frac{W_L W_R \phi''(\mu)}{W}(\mu_L - \mu)(\mu_R - \mu) \le 0,$$

since the current mean μ lies between μ_L and μ_R. Integrating over the unit square gives the quadrangle inequality. Empty blocks follow by continuity. By the usual crossing comparison the previous-row terms cancel, so the smallest maximizing boundary can be taken nondecreasing, and evaluating each DP row by divide and conquer finds the exact optimum of the full recurrence.

At 80-digit precision the accelerated solver matches a separate full recurrence at every feasible prefix for geometric and nonuniform linear amplitudes of lengths 9, 17 and 33 with up to eight bins, and exhaustive enumeration on the six length-9 cases with two, three and four bins. These checks use fixed numerical tolerances and adjacent-quadrangle tests. Table 5 compares the full benchmark instances with an independently written C++ solver in IEEE binary128 arithmetic (113 significand bits), which evaluates every feasible predecessor without monotone-boundary pruning. For the geometric weights it computes block masses and moments from geometric sums rather than differences of nearly equal prefix sums, and it evaluates $\phi(x) = \sum_{j \ge 1} x^{2j}/[(2j)(2j-1)]$ by adding positive terms until the sum stops changing at this precision. The Python solver keeps its 80-digit prefix sums. Reevaluating the independently verified partitions reproduces the three robust-information minima reported below.

Table 5: Independent check of the clean-MI optimum: transitions inspected and feasible prefix states compared by the full solvers, and the largest relative difference in prefix objective. All differences are below the prespecified 10^{-25} tolerance, and every maximizing boundary and selected sign allocation agrees.

L	Transitions	States	Largest rel. difference
16	4,412,901	18,315	$3.64 \cdot 10^{-32}$
32	35,228,921	72,197	$7.71 \cdot 10^{-32}$
64	281,540,529	286,665	$1.57 \cdot 10^{-31}$

All maps are fitted before any contamination level is supplied and are tested on complete grids of 21, 37 and 69 points. The MI-optimal maps keep at least as much clean MI as both comparison maps, and they preserve TV. Yet their minimum robust-information ratios fall to .86740973, .13479847 and $1.70734385 \cdot 10^{-5}$, at integer clipping scales 2, 8 and 34, and no full-output classes overlap. Because the partitions are deterministic, every output contaminant lifts to an input contaminant, and the same ratios hold under input contamination even though the feasibility boundary is preserved. Among the other maps on the $L = 64$ pair, signed KL retains 98.12% of clean Hellinger information but only 1.74% at its weakest robust level. Ordinary KL uses 195 symbols and retains 99.04% of

clean Hellinger information, yet loses feasibility at 19 of 69 levels. The clean-selected bit loses feasibility at 64 levels, while the bit selected separately for each supplied level retains about 23.00% at its weakest point. The sign bit stays feasible across the grid but retains only $6.82 \cdot 10^{-10}$ of clean Hellinger information.

I.4 Heterogeneous finite pairs and fallback

Unlike the geometric families, where positive and negative evidence mirror each other, the following four ensembles allow asymmetric positive and negative likelihood evidence:

- sparse: P, Q independent symmetric Dirichlet with concentration .2 per coordinate;
- uniform: the same with concentration 1;
- lognormal: $P_i \propto e^{Z_i}$ and $Q_i \propto e^{Z'_i}$ with all $Z_i, Z'_i \sim N(0, 4)$ independent;
- correlated: $P, R \sim \mathrm{Dirichlet}(1)$ independent and $Q = (1 - \tau)P + \tau R$, with τ cycling through $10^{-1}, 10^{-3}, 10^{-6}$ so that each value has eight replicates per dimension.

For each ensemble and $d \in \{8, 32, 128, 512\}$ there are 24 pairs, and each encoder is fitted once and evaluated at the fractions $0, .1, .25, .5, .75, .9, .99, .9999, .99999999$ of the pair's raw overlap boundary. The sign map and budget-constrained signed KL are as in the geometric study.

Of the 384 pairs in Table 6, 377 use fewer than d outputs and 331 save at least one fixed-length bit. The minimum retained-information ratio over all 3,456 evaluations is .78384675. Identity fallback occurs only for two correlated pairs with $d = 8$, each requiring nine constructed outputs. The density sweep in Appendix I.11 looks at why the coarse geometric family falls back more often.

Table 6: Heterogeneous pairs. Each row summarizes 24 draws, each evaluated at nine fixed boundary fractions. Median K is the median deployed output count and Fallback counts identity fallbacks out of 24. The three "min." columns give the worst full-output robust-information ratio over all draws and levels in the row. Signed KL is held to our output budget, and the sign map has two outputs.

Ensemble	d	Median K	Fallback	Ours min.	KL min.	Sign min.
Dirichlet(.2)	8	4	0	0.8571	0.5493	0.4324
Dirichlet(.2)	32	4	0	0.8109	0.5611	0.4626
Dirichlet(.2)	128	4	0	0.7838	0.5633	0.5633
Dirichlet(.2)	512	4	0	0.8174	0.6119	0.6119
Dirichlet(1)	8	5	0	0.8796	0.7207	0.4548
Dirichlet(1)	32	6	0	0.8713	0.5968	0.5597
Dirichlet(1)	128	6	0	0.9075	0.7723	0.5490
Dirichlet(1)	512	6	0	0.9274	0.7820	0.6024
Lognormal	8	4	0	0.8423	0.7098	0.4630
Lognormal	32	4	0	0.8653	0.6112	0.6112
Lognormal	128	4	0	0.8373	0.6491	0.6491
Lognormal	512	4	0	0.8626	0.6863	0.6863
Correlated	8	6	2	0.8868	0.3297	0.0715
Correlated	32	9	0	0.8442	0.4318	0.0339
Correlated	128	10	0	0.9129	0.4689	0.0048
Correlated	512	13	0	0.9124	0.5154	0.0355

Pairs are generated with PCG64 and `SeedSequence([20260911, ensemble, d, replicate])`, with ensembles numbered in the order above and replicates from zero. Gamma and normal draws use ordinary floating point, and normalization and every later law calculation use 80 decimal digits. No pair is discarded. We record the constructed and deployed output counts, split atoms, nonzero channel entries, fixed-length bits, and robust information at every level. For each ensemble, the pair with the smallest retained-information ratio under the proposed encoder was reevaluated at 120 digits. Across its three methods and nine levels the largest relative differ-

ence is below $7.1 \cdot 10^{-68}$. In every ensemble each encoder preserved TV, passed the information data-processing check, and kept one channel signature across all supplied levels.

I.5 Pilot generation and fitted-test evaluation

In the constant-J pilot families we set $\theta_j = +1$ for every j, so that the experiment measures how training noise affects the fitted encoder and test. Under the true laws P_i/Q_i takes four values, so the true-law oracle encoder T^{or} has four outputs, one per likelihood-ratio value. It is sufficient, and $\mathcal{I}_\varepsilon(T^{\mathrm{or}}P, T^{\mathrm{or}}Q) = \mathcal{I}_\varepsilon(P, Q)$ at every ε. Estimated laws can have more distinct ratios, and the fitted encoder may then mix true weak signs. From each pilot the procedure forms unsmoothed empirical laws, computes the specified ρ, and fits one encoder that it uses for all five supplied levels. Over all 2,880 candidates the elementary certificate certifies 672.

Table 7: Elementary certification. Each entry is the smallest tested pilot size per class at which all 32 fitted tests are certified. "none" means that no tested size up to 10^8 achieves this.

Nominal family	Clean	$.1\varepsilon_*$	$.5\varepsilon_*$	$.9\varepsilon_*$	$.99\varepsilon_*$
Binary, $d = 2$	10^5	10^5	10^6	10^7	none
Constant J, $d = 34$	10^7	10^7	10^8	none	none
Constant J, $d = 130$	10^8	10^8	10^8	none	none

Each proposed test is judged against the true laws by a Chernoff bound. Let g_y be the estimated least-favorable log-likelihood ratio of a proposed fitted test. For fixed $\lambda > 0$, its worst-input P-world moment factor is

$$M_P(\lambda) = (1-\varepsilon)\sum_y (TP)_y e^{-\lambda g_y} + \varepsilon \max_x \sum_y T(y \mid x) e^{-\lambda g_y}, \tag{28}$$

and M_Q uses TQ and a positive exponent. By independence and Markov's inequality the two errors are at most $M_P(\lambda)^n$ and $M_Q(\lambda)^n$, and the maximum over input symbols makes the contaminant term the exact worst case. We minimize over $\lambda \in \{0, 1/8, 1/4, 1/2, 3/4, 1\}$ (at zero the bound is one), treat infinite scores conservatively, and drop zero-probability terms.

For the binary pair the realized encoder is a bijection, possibly reversing the output order, and binary search through the implemented decision function locates its raw-count threshold. The worst probability of the positive-evidence symbol is $(1-\varepsilon)P_0$ under P and $(1-\varepsilon)Q_0 + \varepsilon$ under Q, and SciPy's binomial CDF and survival function give the exact-model errors numerically. These are checked against (28), and against (26) whenever the confidence event holds.

Table 8: Counts out of 32 at $.99\varepsilon_*$: certified fitted tests and tests with an oracle worst-input Chernoff bound below $1/4$. A pilot proposing no finite test meets neither criterion.

	Binary		$d = 34$		$d = 130$	
N	cert.	oracle bound	cert.	oracle bound	cert.	oracle bound
10^3	0	4	0	0	0	0
10^4	0	10	0	0	0	0
10^5	0	32	0	2	0	0
10^6	0	32	0	28	0	31
10^7	0	32	0	32	0	32
10^8	0	32	0	32	0	32

For the binary pair at $.99\varepsilon_*$, 8 pilots at $N = 10^3$ and 3 at $N = 10^4$ estimate overlapping classes. Among the remaining proposed tests, the largest exact worst-world errors are .9999997609 and greater than .99999999999, respectively. At $N = 10^5, 10^6, 10^7, 10^8$ no estimated pair overlaps, and the largest exact errors are .03869, .01401, .00353 and .00279. These are the worst cases among the 32 pilots we drew.

Runs are seeded with NumPy's PCG64 via `SeedSequence([20260909, i, N, r])`, with i the family index, N the pilot size and r the replicate. One stream produces both independent class multinomials. The sampler receives probabilities in ordinary floating point, and normalization and every information calculation use 80-digit arithmetic. For every pilot and level we log the true estimation error, whether the confidence event held, the map signature, output count, retained information, proposed test size, certificate, and diagnostic risk values.

I.6 Paired independent-calibration comparison

Table 9: Smallest tested total clean budget per class for which all 32 runs certify. Calibration splits the budget equally between fitting and calibration. The elementary method uses it all for fitting. "None" means no tested budget through 10^8 meets the criterion.

Family	Certificate	Clean	$.1\varepsilon_*$	$.5\varepsilon_*$	$.9\varepsilon_*$	$.99\varepsilon_*$
$d = 2$	Calibration	10^3	10^3	10^3	10^4	10^6
	Elementary	10^5	10^5	10^6	10^7	none
$d = 34$	Calibration	10^4	10^4	10^4	10^6	10^7
	Elementary	10^7	10^7	10^8	none	none
$d = 130$	Calibration	10^4	10^4	10^4	10^6	10^7
	Elementary	10^8	10^8	10^8	none	none

Of the 2,880 candidates, 2,222 certify. Of the rest, 608 have insufficient moment margin, 18 have estimated overlapping neighborhoods, and 32 have infinite clean scores. At the same total clean budgets the elementary baseline certifies only 672. Under independent calibration, $B = 10^3$ in the clean setting certifies 20/32 runs for $d = 34$ and 0/32 for $d = 130$. At $B = 10^5$ and $.9\varepsilon_*$ the counts are 31/32 and 27/32, respectively.

The true calibration CDF event fails in three of 576 pilot pairs, all for $d = 34$: $(B, r) = (10^3, 24), (10^4, 8), (10^8, 28)$, with zero-based replicate indices. We check 28,150 candidate moments on the realized CDF events. Fixed edge cases cover equal laws, one-sided zeros, insufficient calibration, a randomized input channel, and a jointly unobserved nominal output symbol with score zero.

The largest oracle Chernoff bound among certified candidates is .248415973, including candidates whose CDF event fails. For the 832 certified binary candidates, the largest exact-model binomial error is .0651897431. We compute these values using the actual-input moments and implemented decision thresholds described above. Deployment cost varies much more widely: for $d = 130$, $B = 10^7$ and $.99\varepsilon_*$, the selected sample sizes range from 754,837 to 164,895,920. Replicate 20 attains the maximum. Its calibrated upper moment lies much closer to one than its true moment, so certification prescribes far more observations than the oracle Chernoff bound requires.

The comparison uses the binary pair $(.8, .2), (.2, .8)$ and the constant-J families of Appendix I.5, with total budgets $B \in \{10^3, 10^4, \ldots, 10^8\}$ per class and 32 paired replicates per family. Replicate r of family i at budget B uses six PCG64 streams $\xi_1, \ldots, \xi_6$ spawned from `SeedSequence([20260910, family_index, B, replicate])`, with `family_index` $= i$ and `replicate` $= r$: ξ_1, ξ_2 give the P and Q training counts, ξ_3, ξ_4 the P and Q calibration counts, and ξ_5, ξ_6 the P and Q encoder coins. The calibrated method fits its encoder $\widehat{T}$ (the T of Corollary 8) on $B/2$ observations per class and pushes the calibration counts c_P, c_Q ($\|c_P\|_1 = \|c_Q\|_1 = B/2$) through $\widehat{T}$ with fresh multinomial private coins. $\widehat{T}$ and the nominal order are fixed before c_P, c_Q or ε are seen. The baseline fits its own encoder to the summed training and calibration input counts, all B per class. Neither method smooths, and neither restricts the deployment size n. Levels are $\varepsilon/\varepsilon_* \in \{0, .1, .5, .9, .99\}$ for the true raw boundary ε_*, with the moment grid and confidence parameters of the main text. Each run logs c_P, c_Q and their encoded counts, the channel signature, the moment bounds, the selected n, and every abstention and confidence-event outcome.

I.7 Deployment costs and bounded deployment budgets

Four things contribute to the deployment cost: the testing problem itself, information lost by the representation, errors in the fitted decision rule, and slack from calibration. We evaluate deployment costs for every candidate in the paired study and try to separate these contributions in four stages.

1. Raw pair. Appendix A gives
$$\left\lceil \frac{1}{4\mathcal{I}_\varepsilon(P,Q)} \right\rceil \le n_\varepsilon(P,Q) \le U_{\mathrm{raw}} := \left\lceil \frac{2\log 4}{\mathcal{I}_\varepsilon(P,Q)} \right\rceil .$$
2. Representation. For the known-law encoder T° and each realized fitted encoder $\widehat{T}$, the analogous Hellinger sufficient counts
$$U_{T^\circ} = \left\lceil \frac{2\log 4}{\mathcal{I}_\varepsilon(T^\circ P, T^\circ Q)} \right\rceil, \qquad U_{\widehat{T}} = \left\lceil \frac{2\log 4}{\mathcal{I}_\varepsilon(\widehat{T}P, \widehat{T}Q)} \right\rceil$$
use the true pushed-forward laws and full output Huber classes, so that representation loss is measured apart from the decision rule.
3. Fitted score. Keep the fitted score, replace the empirical upper moment in (15) by the true clean moment, keep the exact worst-input contaminant term, and minimize over the same positive λ grid separately in each world, with the trivial bound one allowed as a candidate. If the larger of the two minima, b_{true}, is below one,
$$n_{\mathrm{score}} = \left\lceil \frac{\log 4}{-\log b_{\mathrm{true}}} \right\rceil ;$$
otherwise the score has no finite bound on the tested grid. Estimated overlap and infinite scores are likewise recorded without a finite score bound.
4. Certified. $n_{\mathrm{certified}}$ is the sample size selected by Corollary 8. It uses the same fitted score and tilt grid as n_{score}, so $n_{\mathrm{certified}}/n_{\mathrm{score}}$ measures how much extra sufficient sample size calibration costs.

The Hellinger-based counts are reported separately.

Table 10 traces these stages for the $d = 130$ family at $B = 10^7$. On all 32 fitted maps the true pushed-forward pair keeps the raw robust information at the five evaluated levels. So $U_{\mathrm{raw}} = U_{T^\circ} = U_{\widehat{T}}$, and the representation is not the bottleneck here. At $.99\varepsilon_*$ the raw lower bound is 35,722 observations and $U_{\mathrm{raw}} = 396{,}160$. Fitting the score gives $n_{\mathrm{score}} \in [399{,}461, 764{,}571]$, and calibration gives $n_{\mathrm{certified}} \in [754{,}837, 164{,}895{,}920]$. The top of that range is replicate 20, with $n_{\mathrm{score}} = 649{,}061$ and ratio 254.05, while the median of the 32 calibration-to-score ratios is 3.62. All 32 calibration confidence events hold in this group. Ten times more clean data narrows the calibration gap. At $B = 10^8$ the calibrated counts at this level fall to 493,677–702,317 (median 575,928.5), while the raw information at this level does not change.

Table 10: Deployment costs for $d = 130$ and $B = 10^7$ (32 runs per level). U, the Hellinger sufficient count, is the same for the raw pair, the known-law encoder, and every fitted encoder. The score and certified columns are medians, and the last column is the median over runs of the certified-to-score ratio.

$\varepsilon/\varepsilon_*$	U	n_{score}	$n_{\mathrm{certified}}$	Median ratio
0	21	21	21	1.00
.1	27	27	27	1.00
.5	106	106	108	1.02
.9	3,961	4,019.5	4,405.5	1.10
.99	396,160	464,145	1,582,679	3.62

In practice the deployment sample size may be capped. Under a cap $n_{\max}$, the certified test is run only if its prescribed sample size is at most $n_{\max}$. Otherwise the method abstains. Extra abstentions can only shrink the event of certifying and then erring, so the joint-error guarantee of

Table 11: Accepted runs out of 32 after imposing a deployment cap, for $d = 130$ and $B = 10^7$. Without a cap all 32 certify at every listed level. The original selected sample count is kept whenever a run is accepted.

$\varepsilon/\varepsilon_*$	10^2	10^3	10^4	10^5	10^6
0	32	32	32	32	32
.1	32	32	32	32	32
.5	0	32	32	32	32
.9	0	0	32	32	32
.99	0	0	0	0	2

Corollary 8 still holds. How many candidates survive depends on the cap. Out of all 2,880, caps of $10^2, 10^3, 10^4, 10^5, 10^6$ keep 1,204, 1,674, 1,938, 2,021, and 2,162, against 2,222 with no cap. (The 2,880 include candidates that had already abstained before any cap was imposed.)

I.8 Reproducible runs

All numerical studies use synthetic finite laws, and each configuration was fixed before it was run. Pilot and calibration data are drawn as multinomial counts, and no deployment observations are sampled. Deployment errors are computed from the specified laws, by oracle Chernoff bounds and, for binary pairs, numerical binomial tails. The learned procedures see only their samples while the evaluators use the true laws. Reported counts include all generated pilots, abstentions, and confidence-event failures. The accompanying package contains the experimental protocols, source hashes, individual records, and summaries for the heterogeneous-pair, clean-MI, learning, deployment-cost, and information-bound evaluations. For software we use Python 3.12, NumPy 2.5.1, SciPy 1.18.0, mpmath 1.3.0 with gmpy2 2.3.1, and GCC 13.2.0 for the independent binary128 solver. Timings are on an AMD Ryzen 9 6900HS.

I.9 Implementation and geometry validation

The sparse encoder implementation was compared with an independently written reference on 119 finite pairs and 441 contamination levels, including asymmetric laws, rare support changes, near-identical pairs, neutral mass, and split atoms. Besides the channels, the comparison covers channel interval midpoints, count conservation under private coins, unchanged map signatures, and identity fallback. Sixteen further checks compare the implemented binary fitted-test decisions with exact-model binomial errors under known nominal perturbations. Every check passed. The finite linear-programming diagnostic for Proposition 5 has 2,484 instances, with 1,884 explicit witnesses of actual input contamination and 312 cases in which input feasibility survives although output TV shrinks. The largest primal residual was $2.71 \cdot 10^{-15}$ and the largest witness discrepancy $6.67 \cdot 10^{-16}$, both below the specified tolerance of $2 \cdot 10^{-8}$.

I.10 Designing for a fixed contamination level

One could instead design the quantizer for a single contamination level, using the least-favorable pair at that level (Huber, 1965; Pensia et al., 2023). For each dense pair in Table 4, we take the raw least-favorable pair (p^*, q^*) at a design level $\varepsilon_0/\varepsilon_* \in \{.1, .5, .9\}$ and quantize it with a sign-preserving contiguous partition $\mathcal{B}$ with no more outputs than our encoder, maximizing

$$\sum_{B \in \mathcal{B}} \frac{(p^*_B - q^*_B)^2}{(\sqrt{p^*_B} + \sqrt{q^*_B})^2},$$

the Hellinger information kept from the selected raw least-favorable pair. A dynamic program maximizes this sum and also chooses how many outputs each sign gets. Equal least-favorable likelihood ratios are grouped first. We read the ties off the clipping threshold instead of detecting them with a numerical equality tolerance.

Clipping can leave fewer distinct likelihood ratios than outputs. We spend the spare outputs on the clipped tails, with the contiguous partition that maximizes their clean MI. The design objective does not change, and the whole budget is used. Table 12 reports the robust-information minimum before and after this step, so the comparison includes the effect of this tie resolution. Once fitted, each map is frozen and evaluated on the original nominal laws under full-output Huber contamination at every supplied level. At 80-digit precision, exhaustive predecessor enumeration verifies every feasible prefix state of the accelerated recurrence (relative discrepancies below 10^{-55}). The refined map's design-level robust information agrees with the grouped map's to relative tolerance 10^{-50}, and we checked the tail recurrence against exhaustive predecessor enumeration on instances with 7, 15, and 31 tail levels.

Table 12: Designs for a single contamination level, on the dense pairs. Both "At design" and "Min." are robust information ratios relative to the raw experiment, the latter minimized over the same grid as Table 4. "Grouped" is that minimum before the spare outputs are spent on clean-MI refinement, with its output count in parentheses.

L	$\varepsilon_0/\varepsilon_*$	Outputs	At design	Min.	Grouped (outputs)
16	.1	38	0.99782	0.08719	0.08719 (38)
16	.5	38	0.99997	0.01630	0.01630 (38)
16	.9	38	1.00000	0.86741	0.01139 (2)
32	.1	74	0.99943	0.00037	0.00037 (74)
32	.5	74	1.00000	0.00634	$6.490 \cdot 10^{-5}$ (56)
32	.9	74	1.00000	0.13480	$4.467 \cdot 10^{-5}$ (2)
64	.1	146	0.99987	$5.755 \cdot 10^{-9}$	$5.755 \cdot 10^{-9}$ (146)
64	.5	146	1.00000	$6.201 \cdot 10^{-6}$	$9.903 \cdot 10^{-10}$ (56)
64	.9	146	1.00000	$1.707 \cdot 10^{-5}$	$6.816 \cdot 10^{-10}$ (2)

For $L = 64$, the design at $.1\varepsilon_*$ loses the most information on clean data. The designs at $.5\varepsilon_*$ and $.9\varepsilon_*$ retain .99613 and .99842 of clean MI, yet their weakest robust levels occur at clipping scales 38 and 34. At $.9\varepsilon_*$ the raw least-favorable pair has only two distinct ratios, and grouping ties alone would use two of the 146 outputs. With refinement the map is the clean-MI-optimal partition, and the smallest robust-information ratio rises from $6.816 \cdot 10^{-10}$ to $1.707 \cdot 10^{-5}$ (Table 12). On the same grid, our encoder retains at least .96362.

I.11 Support density and alphabet reduction

All twelve coarse geometric pairs fall back to the identity, while the dense pairs compress. We sweep the number s of distinct amplitudes per dyadic interval, $s \in \{1, 2, 4, 8, 16, 32\}$, for each $L \in \{16, 32, 64\}$, keeping $\alpha = 1.5$ and otherwise following the dense family of the main comparison. That is 18 pairs, with $d = 2(sL + 1)$ symbols each. We record K_c, the deployed count K, clean J, and the whole robust information curve; Table 13 lists the alphabet counts.

Table 13: Input, constructed, and deployed alphabet counts, written $d/K_c/K$. Identity fallback occurs at $s = 1$. For $s \geq 2$, every constructed map reduces the input alphabet.

s	$L = 16$	$L = 32$	$L = 64$
1	34/40/34	66/76/66	130/148/130
2	66/38/38	130/74/74	258/146/146
4	130/38/38	258/74/74	514/146/146
8	258/38/38	514/74/74	1026/146/146
16	514/38/38	1026/74/74	2050/146/146
32	1026/38/38	2050/74/74	4098/146/146

With $s = 1$ each evidence scale has one amplitude, and the constructed map is larger than the input alphabet. For $s \geq 2$ the added symbols land in scales the encoder already covers. Going from 2 to

32 amplitudes per scale raises the fixed-length saving from one to five bits per observation, and the constructed alphabet does not grow. The smallest information ratio over all pairs and tested levels is 0.96064. This sweep should not be over analyzed: whether extra symbols save outputs also depends on how the mass is spread across scales (compare $M_{0,s}$ in Proposition 11), and J by itself does not determine fallback.

I.12 Sensitivity to calibration tilts

We reran certification with richer tilt searches and held fixed the fitted laws, encoders, scores, calibration observations, and confidence level, across all 576 training/calibration pairs and 2,880 supplied-level candidates of Appendix I.6. The reconstructed channel signatures match the stored ones, and the five-point grid reproduces every recorded abstention and certified sample count. The searches are nested. With $\Lambda_0 = \{1/8, 1/4, 1/2, 3/4, 1\}$, the small-tilt and dense grids are

$$\Lambda_{\mathrm{s}} = \Lambda_0 \cup \{2^{-j} : 4 \le j \le 16\}, \qquad \Lambda_{\mathrm{d}} = \Lambda_{\mathrm{s}} \cup \{j/64 : 1 \le j \le 128\},$$

so $\Lambda_0 \subset \Lambda_{\mathrm{s}} \subset \Lambda_{\mathrm{d}}$, and the last search runs a bounded continuous optimization on $[0, 2]$ for each hypothesis and keeps the best evaluated tilt, dense-grid points included. Candidates are generated in floating point with stable exponential increments, and the selected moment bounds are recomputed at 80-digit precision. No new confidence event is needed, because the DKW event of Corollary 8 holds for all of these choices at once.

We then checked all 18,088 selected hypothesis-specific moments against the true nominal laws. On the realized DKW events, every selected bound dominates its true moment and gives an error bound of at most $1/4$ at the certified sample count. For 118 candidates (replicate zero in each eligible cell, plus all 32 highlighted near-boundary runs), a full 80-digit enumeration of the dense grid reproduces the selected sample counts. The 18 estimated-overlap and 32 infinite-score abstentions are unchanged by the tilt search.

Table 14: Effect of the tilt search. The Certified column counts over all 2,880 candidates. The other columns are for the 32 runs with $d = 130$, $B = 10^7$ and $\varepsilon = .99\varepsilon_*$. Median ratio is the median of $n_{\mathrm{certified}}/n_{\mathrm{score}}$, where each method's true-score moments use the same tilt choices as the method itself.

Tilt search	Certified	Median n	Maximum n	Median ratio
Five-point	2,222	1,582,679	164,895,920	3.619
Small tilts	2,274	1,582,679	22,997,800	3.619
Dense grid	2,274	1,560,802.5	22,997,800	3.609
Continuous search	2,274	1,560,356	22,967,784	3.608

Smaller tilts certify 52 additional candidates and change 90 sample counts. Among the 32 runs in Table 14, the effect is concentrated in the largest counts. In the most costly run, the limiting hypothesis selects $\lambda = 1/16$ instead of $1/8$, which lowers the count from 164,895,920 to 22,997,800, and continuous search trims it only a little further, to 22,967,784. In this run, much of the sample requirement comes from the five-point grid. The median barely moves. Across the 32 runs it goes from 1,582,679 to 1,560,356. Even after continuous search, the median ratio of the certified count to the corresponding true-score count is 3.608, against 3.619 on the five-point grid.

I.13 Score scale and calibration uncertainty

The optimal tilt need not approach zero as robust information vanishes. Consider a symmetric binary least-favorable pair

$$P_t^* = \left(\frac{1+t}{2}, \frac{1-t}{2}\right), \qquad Q_t^* = \left(\frac{1-t}{2}, \frac{1+t}{2}\right), \qquad 0 < t < 1.$$

Its log-likelihood score has values $\pm a_t$, where $a_t = \log((1+t)/(1-t))$. Under P_t^*,

$$M_P(\lambda) = \frac{1+t}{2} e^{-\lambda a_t} + \frac{1-t}{2} e^{\lambda a_t}.$$

M_P is strictly convex in λ, and setting its derivative to zero gives the unique minimizer:

$$(1-t)e^{\lambda a_t} = (1+t)e^{-\lambda a_t} \iff e^{2\lambda a_t} = \frac{1+t}{1-t} = e^{a_t} \iff \lambda = \frac{1}{2},$$

for every $t > 0$, and the same holds under Q_t^*. The minimum $M_P(1/2) = \sqrt{1-t^2}$ is the Hellinger affinity of the pair, so

$$\mathcal{I} = 2\big(1 - \sqrt{1-t^2}\big) \sim t^2.$$

The tilt stays at $1/2$ while the score shrinks as $t \to 0$. Estimation changes the fitted score, and that can move its best tilt (Appendix I.12).

The same example shows how large the DKW slack is. At $\lambda = 1/2$ the exponential score has range $2t/\sqrt{1-t^2}$. If the score is correctly fitted, the empirical clean moment equals its population value, and the cap in C_R is inactive, the calibrated worst-world moment bound is

$$b = \sqrt{1-t^2} + (1-\varepsilon)\delta_M \frac{2t}{\sqrt{1-t^2}} = 1 - \frac{t^2}{2} + O(t^4) + O(\delta_M t).$$

The margin $1-\sqrt{1-t^2}$ has order t^2 and the uncertainty term order $\delta_M t$, so under these assumptions this DKW certificate can certify only if $\delta_M = O(t)$. For a fixed symmetric nominal pair approaching overlap, the corresponding calibration scale is of order $\log(1/\beta)/\mathcal{I}$.

Variance-sensitive confidence bounds are another possible route. For a fixed bounded score, empirical Bernstein bounds have a leading term proportional to $\sqrt{\widehat{\mathrm{Var}}(f)\log(1/\beta)/M}$ and a range-dependent term of order $\log(1/\beta)/M$ (Maurer & Pontil, 2009). To replace the CDF argument, however, such a bound would have to hold uniformly over the pairs (ε, λ) selected after calibration, as the CDF event does for the present certificate.

I.14 Known laws with an atomless contrast distribution

Let P and Q be known laws on a measurable space, and define $m = (P+Q)/2$ and

$$z(x) = \frac{1}{2}\left(\frac{dP}{dm}(x) - \frac{dQ}{dm}(x)\right).$$

Assume that $z(X)$ has an atomless distribution under $X \sim m$, and that z and its mixture quantiles are known. Then

$$|z| \le 1, \qquad \int z\, dm = 0, \qquad V = \int |z|\, dm, \qquad J = \int z^2\, dm.$$

Within each sign, the upper contrast values go into the same dyadic intervals as before. The lower values, in decreasing order, are cut into pieces with target conditional masses $M_0, 2M_0, 4M_0, \ldots$ until the mass is exhausted. The distribution is atomless, so these cuts can be placed at mixture quantiles, and the encoder is a deterministic function of z.

Proposition 12 (Atomless contrast)**.** *Under these assumptions, this encoder T has $O(1+\log(1/J))$ outputs and satisfies $\mathcal{I}_\varepsilon(TP, TQ) \ge \mathcal{I}_\varepsilon(P,Q)/32$ for $0 \le \varepsilon < \varepsilon_*$.*

The proof of Lemma 2 uses ordered bin supports, their masses, and their conditional means, and each step holds with integrals in place of finite sums. On sign s the conditional first moment is $V/(2m_s)$, so

$$j_s \ge \left(\frac{V}{2m_s}\right)^2 \ge \frac{J^2}{4}.$$

The clipping equations and the coordinatewise Hellinger bounds integrate under m in the same way. The argument therefore gives $O(1+\log(1/J))$ outputs and the $1/32$ information guarantee for this class of known laws.

I.15 ALPHABET–INFORMATION TRADE-OFF

The budget study uses $\alpha \in \{1.25, 1.5, 1.75\}$, $L \in \{8, 16, 32, 64\}$, and four distinct amplitudes per dyadic interval,

$$A_j = 2^{-L-1+j/4}, \qquad w_j \propto 2^{-\alpha j/4}, \qquad 0 \le j \le 4L,$$

plus four cases from the exact coarse family with $\alpha = 1.5$ and one amplitude per interval. For each pair the budgets are the elements of $\{2, 4, 8, 16, 32, 64, 128\}$ below the $q = 2$ output count, together with that count, giving 104 pair–budget settings in total.

Given a budget K, we scan $q = 2^{j/4}$ over integers $j \ge 4$ and use the first construction with at most K outputs, so q depends only on the constructed alphabet, never on contamination levels or information values. Budgets too small for separate high and low bins use the sign bit, and the identity is used whenever it fits. At each budget the clean-MI baseline uses the full capacity and optimizes the split between signs. The sign bit is kept as a control and no map changes across evaluation levels.

The evaluation set $\mathcal{E}$ contains clean data, the boundary fractions $.25, .5, .9, .99, .9999, .99999999$, and all positive integer and half-integer clipping scales below L, with repeated levels removed. We report

$$D_T = \max_{\varepsilon \in \mathcal{E}} \frac{\mathcal{I}_\varepsilon(P, Q)}{\mathcal{I}_\varepsilon(TP, TQ)}.$$

The denominator uses full output Huber classes. For the actual input-contamination overhead restricted to this grid, Appendix A gives the bounds

$$\max_{\varepsilon \in \mathcal{E}} \frac{\mathcal{I}_\varepsilon(P, Q)}{20h^2(TP^*_\varepsilon, TQ^*_\varepsilon)} \le \max_{\varepsilon \in \mathcal{E}} \frac{n^T_\varepsilon}{n_\varepsilon} \le 20 D_T.$$

Both bounds are saved with the curves. D_T is only a finite-grid information statistic. The all-level guarantee comes from the theorem.

Table 15: All budget settings. s is the number of amplitudes per dyadic interval. K_q is the adjustable construction's actual output count. Both methods have capacity K. Smaller D_T is better.

α	L	s	K	K_q	D_T, growth	D_T, MI
1.25	8	4	2	2	8.441	8.441
1.25	8	4	4	4	2.58	2.765
1.25	8	4	8	8	1.298	1.475
1.25	8	4	16	14	1.072	1.103
1.25	8	4	20	20	1.035	1.07
1.25	16	4	2	2	358.6	358.6
1.25	16	4	4	4	19.76	72.02
1.25	16	4	8	8	2.495	17.5
1.25	16	4	16	14	1.251	4.799
1.25	16	4	32	32	1.057	1.819
1.25	16	4	38	38	1.035	1.601
1.25	32	4	2	2	$1.31{\cdot}10^{6}$	$1.31{\cdot}10^{6}$
1.25	32	4	4	4	2691	$2.44{\cdot}10^{5}$
1.25	32	4	8	8	18.06	$5.14{\cdot}10^{4}$
1.25	32	4	16	14	2.284	8458
1.25	32	4	32	32	1.232	1416
1.25	32	4	64	60	1.057	190.3
1.25	32	4	74	74	1.035	116.4
1.25	64	4	2	2	$2.18{\cdot}10^{13}$	$2.18{\cdot}10^{13}$
1.25	64	4	4	4	$8.33{\cdot}10^{7}$	$4.04{\cdot}10^{12}$
1.25	64	4	8	8	2373	$8.45{\cdot}10^{11}$
1.25	64	4	16	14	15.61	$1.37{\cdot}10^{11}$
1.25	64	4	32	32	1.991	$2.21{\cdot}10^{10}$
1.25	64	4	64	64	1.197	$2.77{\cdot}10^{9}$

α	L	s	K	K_q	D_T, growth	D_T, MI
1.25	64	4	128	116	1.057	$4.33{\cdot}10^{7}$
1.25	64	4	144	144	1.035	$1.53{\cdot}10^{7}$
1.5	8	1	2	2	7.783	7.783
1.5	8	1	4	4	2.727	2.651
1.5	8	1	8	8	1.341	1.52
1.5	8	1	16	16	1.096	1.089
1.5	8	1	18	18	1	1
1.5	16	1	2	2	119.2	119.2
1.5	16	1	4	4	19.24	34.15
1.5	16	1	8	8	2.964	9.338
1.5	16	1	16	16	1.287	2.66
1.5	16	1	32	28	1.103	1.089
1.5	16	1	34	34	1	1
1.5	32	1	2	2	$3.04{\cdot}10^{4}$	$3.04{\cdot}10^{4}$
1.5	32	1	4	4	1294	8679
1.5	32	1	8	8	21.34	2155
1.5	32	1	16	16	2.023	538.6
1.5	32	1	32	32	1.236	33.9
1.5	32	1	64	62	1.06	1.089
1.5	32	1	66	66	1	1
1.5	64	1	2	2	$2{\cdot}10^{9}$	$2{\cdot}10^{9}$
1.5	64	1	4	4	$5.5{\cdot}10^{6}$	$5.69{\cdot}10^{8}$
1.5	64	1	8	8	1345	$1.41{\cdot}10^{8}$
1.5	64	1	16	16	9.407	$3.53{\cdot}10^{7}$
1.5	64	1	32	32	2.024	$2.21{\cdot}10^{6}$
1.5	64	1	64	62	1.237	8613
1.5	64	1	128	120	1.06	1.089
1.5	64	1	130	130	1	1
1.5	8	4	2	2	6.258	6.258
1.5	8	4	4	4	2.784	2.349
1.5	8	4	8	8	1.389	1.3
1.5	8	4	16	16	1.055	1.056
1.5	8	4	20	20	1.04	1.042
1.5	16	4	2	2	94.67	94.67
1.5	16	4	4	4	19.47	27.9
1.5	16	4	8	8	3.054	9.013
1.5	16	4	16	16	1.22	2.5
1.5	16	4	32	32	1.053	1.244
1.5	16	4	38	38	1.041	1.186
1.5	32	4	2	2	$2.42{\cdot}10^{4}$	$2.42{\cdot}10^{4}$
1.5	32	4	4	4	1317	6428
1.5	32	4	8	8	21.95	1902
1.5	32	4	16	16	2.055	475.4
1.5	32	4	32	30	1.223	77.93
1.5	32	4	64	60	1.058	12.75
1.5	32	4	74	74	1.041	9.024
1.5	64	4	2	2	$1.58{\cdot}10^{9}$	$1.58{\cdot}10^{9}$
1.5	64	4	4	4	$5.65{\cdot}10^{6}$	$4.21{\cdot}10^{8}$
1.5	64	4	8	8	1384	$1.25{\cdot}10^{8}$
1.5	64	4	16	16	9.058	$3.11{\cdot}10^{7}$
1.5	64	4	32	30	2.058	$4.63{\cdot}10^{6}$
1.5	64	4	64	60	1.235	$8.18{\cdot}10^{5}$
1.5	64	4	128	118	1.056	$5.11{\cdot}10^{4}$
1.5	64	4	146	146	1.041	$2.15{\cdot}10^{4}$
1.75	8	4	2	2	4.238	4.238
1.75	8	4	4	4	2.469	1.939

α	L	s	K	K_q	D_T, growth	D_T, MI
1.75	8	4	8	6	1.402	1.194
1.75	8	4	16	16	1.076	1.047
1.75	8	4	22	22	1.033	1.026
1.75	16	4	2	2	20.35	20.35
1.75	16	4	4	4	8.77	9.773
1.75	16	4	8	8	2.112	3.468
1.75	16	4	16	14	1.265	1.409
1.75	16	4	32	32	1.054	1.082
1.75	16	4	40	40	1.036	1.048
1.75	32	4	2	2	344.6	344.6
1.75	32	4	4	4	106.5	175.2
1.75	32	4	8	8	4.768	72.73
1.75	32	4	16	16	1.864	22.72
1.75	32	4	32	30	1.241	4.575
1.75	32	4	64	60	1.058	1.362
1.75	32	4	74	74	1.037	1.195
1.75	64	4	2	2	$8.86{\cdot}10^4$	$8.86{\cdot}10^4$
1.75	64	4	4	4	$1.54{\cdot}10^4$	$4.52{\cdot}10^4$
1.75	64	4	8	8	29.33	$1.9{\cdot}10^4$
1.75	64	4	16	16	5.006	6141
1.75	64	4	32	32	1.887	1407
1.75	64	4	64	60	1.248	270.2
1.75	64	4	128	116	1.059	42.72
1.75	64	4	144	144	1.037	29.84

Small ordered partitions. For the coarse laws with $\alpha = 1.5$ and $L = 4, 6$, we enumerate every deterministic sign-preserving contiguous partition with $K = 2, 4, 6, 8$ outputs, including every allocation between signs leading to 1,613 partitions across the eight settings. The partition chosen by maximizing minimum normalized information on the integer clipping scales and fixed boundary fractions is then evaluated on the added half-integer scales as well. Table 16 compares its evaluated D_T with the adjustable and clean-MI maps. The optimization is over this stated partition class and design grid.

Table 16: Small-instance budget comparison. "Grid" is the partition chosen by exhaustive search on the design levels. Values use the finer evaluation grid.

L	K	Partitions	D_T, grid	D_T, growth	D_T, MI
4	2	1	2.254	2.254	2.254
4	4	28	1.255	1.63	1.255
4	6	70	1.094	1.125	1.105
4	8	28	1.05	1.125	1.088
6	2	1	4.082	4.082	4.082
6	4	66	1.523	2.058	1.523
6	6	495	1.194	1.229	1.256
6	8	924	1.097	1.29	1.107

The smallest ratio between the adjustable map's minimum information retention and the grid-selected partition's is .740. Increasing its allowed budget need not improve the measured ratio monotonically, because changing q shifts the bin boundaries: for $L = 6$, the adjustable map's D_T changes from 1.229 at $K = 6$ to 1.290 at $K = 8$. The budget rule chooses q by output count alone.

Known one-bit schedule. For the coarse $\alpha = 1.5$ family, fix the library from Appendix D.3. At each supplied level, choose the library member with largest robust information and retain its subsequence. If D_{selected} denotes the worst inverse retention of the chosen bit, the information-cost proxy per received observation is $(L + 2)D_{\text{selected}}$, which accounts for discarding the other

subsequences. The schedule and the identical sign bit each send one bit per observation. The schedule also requires a known observation index.

Table 17: Identical sign bit and known round-robin schedule. The selected bit is allowed to depend on ε; the schedule is fixed beforehand.

L	Library	D_T, sign	D_{selected}	$(L+2)D_{\mathrm{selected}}$
4	6	2.254	2.254	13.52
8	10	7.783	6.678	66.78
16	18	119.2	48.29	869.2
32	34	$3.04{\cdot}10^4$	100.6	3419
64	66	$2{\cdot}10^9$	101.2	6678

Numerical checks. All primary calculations use 80 decimal digits. The clipped-moment inequalities were also tested numerically, on 120 conditional laws, among them singleton laws, laws with repeated amplitudes, and laws with small atom masses, each paired with every $q \in \{2, 3, 4, 8, 32, 256\}$. For these 720 law–growth combinations we evaluated the clipped-moment differences at all 30,838 of their breakpoints. Between breakpoints the first-moment differences $F(t) - \bar{F}(t/q)$ and $\bar{F}(t) - F(t)$ are linear and $q^3\bar{G}(t/q) - G(t)$ has the form $a + bt^2$, so the values at the two endpoints decide each interval. These are still finite-precision checks of the tested laws. On all 120 laws the $q = 2$ channel matches the reference implementation up to a relabeling of outputs, and further checks cover 108 prescribed budgets and 144 information values computed by an independent clipping-root bisection. For the clean-MI baseline, a full MI recurrence agrees with the accelerated solver at all 109,959 states of the 16 budget-study laws. Recomputing 232 selected curve points at 120 digits moves retention by less than 10^{-35} in relative terms, and the independent bisection agrees at the selected worst levels too. Every curve, output count, growth choice, sufficient sample count, and source hash is in the accompanying experiment package.