CAPACITY FLOORS AND READOUT LIMITS FOR BOTTLENECKED MEDICAL IMAGE CLASSIFIERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical image classifiers often route every input through a bounded store, a codebook, a concept layer, or a prototype bank, chosen for interpretability or for deployment cost. When such a model falls short, the cause may be a store too small to separate the classes, an encoder that underuses the store it has, or a decision rule that discards information the store holds, and each calls for a different fix. A validation curve reports the total error alone and leaves the binding cause unidentified. This paper decomposes the error into one term for each of these three causes. The first is a capacity floor, the smallest error any sampled positive quadratic readout attains on a store of dimension m, equal to one minus the sum of the m largest prior masses and therefore fixed by the architecture and the prior alone. The second is the representation gap: the error of the best readout in that class on the learned store, obtained as a semidefinite program with a dual certificate, minus the floor. The third is the readout gap: the attained error minus that best error, the excess due to the decision rule. The floor is tight and holds for every decoder on a discrete store and for the positive-readout class on a continuous store, and a two-dimensional counterexample shows that this decoder boundary is sharp. Inverting the floor sizes a store before training: the prior and a target error fix the narrowest bottleneck that could reach it, so a proposed architecture can be ruled out on paper. Across four benchmarks, five bottleneck families, and an unbottlenecked reference, the decomposition separates certified capacity limits from errors carried by the representation or the readout. On discrete stores and our own register, where it is exact, the representation gap dominates the readout gap in all 12 configurations, and in the 12 diagnostic ones continuous heads deployed with argmax beat the best positive readout of their store, as the decoder boundary predicts.

1 INTRODUCTION

Many classifiers pass every input through a store of limited size before they produce a label: a vector-quantised codebook (van den Oord et al., 2017), a concept layer (Koh et al., 2020), a prototype bank (Li et al., 2018; Chen et al., 2019), or a narrow projection (Hinton & Salakhutdinov, 2006), adopted for interpretability (Rudin, 2019), compression (Esser et al., 2021), or retrieval (Khandelwal et al., 2020). Medical imaging amplifies these pressures: clinicians want representations they can inspect, deployment favours compact models, class prevalence is shaped by disease frequency, and enlarging an approved model may require additional validation.

The error inside such an interface has three possible sources: the store may lack the dimensions the classes need, the encoder may leave those dimensions unused, or the decision rule may discard information the store holds. Each admits a different intervention, and a validation curve reports the total alone, so the question is usually settled by trying each in turn.

Existing bounds answer parts of this question and leave the rest open: Fano's inequality (Cover & Thomas, 2006) and the information bottleneck need a mutual information estimate that is unreliable in high dimension, a hard partition bounds only a discrete store (Blackwell, 1953), and the one-shot counting bound (Shah, 2025) is tight but assumes a single copy of the state in place of its full description.

This paper gives a decomposition that answers the question directly. For a prior π over C classes and a store of dimension m, any readout of the store that samples the label from quadratic scores, given

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.

by positive matrices summing to the identity, has error at least $1 - \Sigma_m(\pi)$, the complement of the m largest prior masses, so this capacity floor depends on prior and dimension alone. Against it we place the error of the best readout the learned store admits, a minimum-error discrimination problem (Helstrom, 1969; Watrous, 2018; Eldar et al., 2003), and the attained error; the two gaps in that chain indicate which component carries the excess error and motivate the corresponding interventions.

The contributions of this work are as follows:

1. A sharp decoder boundary for the capacity floor: it holds for every decoder on a discrete store and for the positive-readout class on a continuous one, and a two-dimensional counterexample, with a sampled positive readout on the floor at every class count and an exact argmax over the same probabilities, fixes the decoder class a guarantee must state.
2. A sizing rule that fixes before training the narrowest store whose floor meets a target error, stated as a dimension, a codebook size, a concept count, or a prototype count, and that stays valid when the prior is estimated: the floor moves by at most the total variation distance between two priors, so sizing at a tightened target certifies the original one.
3. An error attribution separating the store dimension, the evidence the encoder placed in it, and the decision rule applied to it, computed with a dual certificate for codebooks, binary bottlenecks, prototype banks, and narrow projections, and joined to the sizing rule in a diagnostic procedure that returns the admissible dimension for a capacity verdict and the optimal positive readout for a readout verdict. For a continuous store we give BornReg, whose positive readout makes the decomposition exact by construction and whose certified evaluation is also its deployed one.
4. An evaluation on three MedMNIST subsets and CIFAR-100 across five bottleneck families and an unbottlenecked reference, in which the representation gap exceeds the readout gap wherever the decomposition is exact and argmax heads beat the best positive readout of their store.

In a binary task the counting floor is zero for any $m \geq 2$, so capacity is ruled out as a source of error. This matters most in medical imaging, where changing a deployed architecture carries a validation cost, and because the floor depends on the evaluation prior it adapts to a clinically imbalanced class distribution without being retuned.

2 Related Work

Width is commonly selected by architecture-specific ablation or validation in place of an explicit task-dependent lower bound. Codebooks report sizes chosen by ablation (van den Oord et al., 2017; Razavi et al., 2019), concept layers fix the number of named variables in advance (Koh et al., 2020), and prototype architectures fix the size of the exemplar bank (Li et al., 2018; Chen et al., 2019). The sizing rule here supplies such a bound from the evaluation prior.

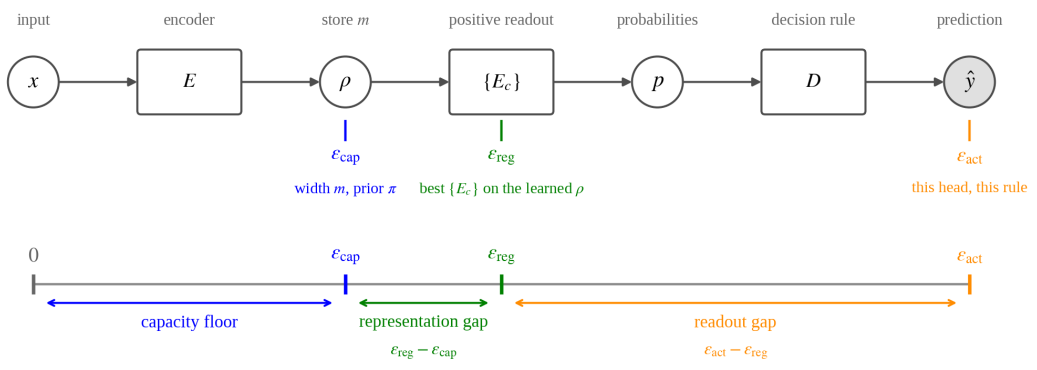


Figure 1: Error attribution in a bottlenecked classifier. The architecture-only capacity floor ε_{cap}, the optimal positive-readout error ε_{reg} on the learned store, and the attained error ε_{act} define the representation and readout gaps.

Algorithm 1 Diagnose a trained bottlenecked classifier

Require: class-conditional states $\{\rho_c\}$ (or codes for a discrete store) and predictions on a held-out split, evaluation prior π, target $\varepsilon^\star$, gap tolerance τ

$q \leftarrow m$ for a continuous store, $q \leftarrow M$ for a discrete store of cardinality M
$\varepsilon_{\mathrm{cap}} \leftarrow 1 - \Sigma_q(\pi)$ ▷ architecture and prior only, available before training
$\varepsilon_{\mathrm{reg}}, \{E_c\} \leftarrow$ minimum-error positive readout on $\{\rho_c\}$ under π, with its dual certificate
$\varepsilon_{\mathrm{act}} \leftarrow$ risk of the model's own predictions under π
$u \leftarrow \chi_U / \log_2 q$ ▷ reported beside the verdict
if $\varepsilon_{\mathrm{act}} \le \varepsilon^\star$ **then**
    return target met
else if $\varepsilon_{\mathrm{cap}} > \varepsilon^\star$ **then**
    return certified dimension limited, widen to $m^\star$ or $M^\star$
else if $\varepsilon_{\mathrm{act}} - \varepsilon_{\mathrm{reg}} > \max(\tau, \varepsilon_{\mathrm{reg}} - \varepsilon_{\mathrm{cap}})$ **then**
    return readout limited, install $\{E_c\}$
else
    return encoder limited, retrain with the utilisation regulariser
end if

Bounds on bounded representations have come from the information side. The information bottleneck trades compression against predictive information (Tishby et al., 2000; Alemi et al., 2017), and mutual information is hard to estimate in high dimension, so such bounds rarely serve as diagnostics (McAllester & Stratos, 2020); lossy-compression estimators make the estimate tractable by compressing the representation first, and have been used to analyse learned networks (Butakov et al., 2024). Concept bottlenecks have been analysed the same way, with an information bottleneck at the concept layer to enforce minimality (Galliamov et al., 2026) or to show that the standard model enforces none (Almudevar et al., 2026); those works constrain the content of a store, while the present one bounds the capacity of any store of a given dimension and attributes the remainder. Logit adjustment and decoupled classifier retraining (Menon et al., 2021; Kang et al., 2020) correct the same readout stage under a skewed prior, by reweighting instead of by solving for the optimum. Our entropic term is the Holevo quantity (Holevo, 1973), reported through a one-sided bootstrap bound.

The counting bound is known in the one-shot communication setting, where a receiver holds a single physical copy and applies one measurement and a classical decision rule (Shah, 2025). We restate this bound for classification and invert it to size a store. The difference is that our decoder holds the full description of the store instead of one copy of it, and our boundary result turns on that difference.

Minimum-error discrimination has a closed form for two hypotheses (Helstrom, 1969), a semidefinite formulation in general (Yuen et al., 1975; Eldar et al., 2003), and a tractable approximation within a factor of two of the optimum (Barnum & Knill, 2002). The register we use belongs to the data re-uploading family (Pérez-Salinas et al., 2020; Schuld et al., 2021), chosen because it makes the best-readout term computable.

3 Bounded stores and their readouts

Let X be the input, $Y \in \{1, \dots, C\}$ the label, and π the evaluation prior, with $\pi_c = \Pr(Y = c)$. Let $\Sigma_q(\pi) = \sum_{i=1}^{\min(q,C)} \pi_{(i)}$, with $\pi_{(i)}$ the i-th largest entry of π, denote the sum of the q largest entries, which equals one whenever $q \ge C$, and let $H(\pi)$ denote the Shannon entropy of π in bits.

A bottlenecked classifier factors as an encoder followed by a readout, drawn with the three errors in Figure 1 and computed from a trained model by Algorithm 1. The encoder maps an input to a state $\rho(x) \in \mathcal{D}(\mathcal{H}_m)$, the positive semidefinite matrices of unit trace on a complex space of dimension m; the store is the pair $(\mathcal{H}_m, \rho)$. Two readout classes are considered.

Classical interpretation. Every quantity below is computed with ordinary linear algebra on classical hardware. A feature vector z becomes the matrix $\rho = zz^\dagger / \|z\|^2$, and a positive readout scores class c as

$$p_c(z) = \mathrm{Tr}(E_c \rho) = \frac{z^\dagger E_c z}{\|z\|^2}, \qquad E_c \succeq 0, \quad \sum_c E_c = I,$$

so the object under study is a normalised quadratic head with positive class matrices summing to the identity I, from which sampling draws a label and an argmax takes the largest entry. The matrix notation states the capacity and best-readout problems compactly.

A *positive readout* is a collection $\{E_c\}_{c=1}^{C}$ with $E_c \succeq 0$ and $\sum_c E_c = I$, which assigns to a state the distribution $p_c(x) = \mathrm{Tr}(E_c \rho(x))$ and emits a class drawn from it. Its risk is

$$R(\{E_c\}) \;=\; 1 - \sum_{c=1}^{C} \pi_c \,\mathrm{Tr}(E_c \rho_c), \qquad \rho_c \;=\; \mathbb{E}[\rho(X) \mid Y = c]\,. \tag{1}$$

The risk is linear in the state, so the class-conditional mean state ρ_c is the only feature of the encoder that a positive readout can respond to, and Eq. 1 holds for input-dependent encoders without modification.

A *decoder* is any measurable map $D : \mathcal{D}(\mathcal{H}_m) \to \{1, \ldots, C\}$, with risk $\Pr(D(\rho(X)) \neq Y)$. Every positive readout induces a decoder by sampling, and the decoder class is strictly larger, a distinction that the following section puts to work.

A store is *discrete* with cardinality M when $\rho(X)$ takes at most M values. A codebook of size K used at T positions is discrete with $M \leq K^T$; a concept layer with n binary concepts is discrete with $M \leq 2^n$; a prototype bank of P prototypes with hard assignment is discrete with $M \leq P$.

4 Capacity floors

4.1 The counting floor

Lemma 1 (Counting floor, Shah, 2025)**.** *Let the store have dimension m and let $\{E_c\}$ be any positive readout. Then*

$$R(\{E_c\}) \;\geq\; 1 - \Sigma_m(\pi). \tag{2}$$

For every prior π and every $m \leq C$ the bound is attained.

Proof sketch. Positivity gives $\mathrm{Tr}(E_c \rho_c) \leq \min(1, \mathrm{Tr}\, E_c)$, and $\sum_c E_c = I$ gives $\sum_c \mathrm{Tr}\, E_c = m$, so the success probability is bounded by a separable concave allocation of total mass m whose greedy optimum places unit mass on the m largest entries of π. Attainment follows by encoding the m heaviest classes on distinct basis vectors and collapsing the rest onto one of them. The argument is the one-shot counting argument (Shah, 2025), restated for a positive readout with one element per class. □

The statement depends on the store only through its dimension, so the floor is available before any training, and it depends on the prior through Σ_m and not through $H(\pi)$, so a skewed evaluation prior admits a much narrower store than a uniform one at the same class count.

Corollary 2 (Discrete stores)**.** *If the store is discrete with cardinality M, then every decoder, deterministic or randomised, has risk at least $1 - \Sigma_M(\pi)$.*

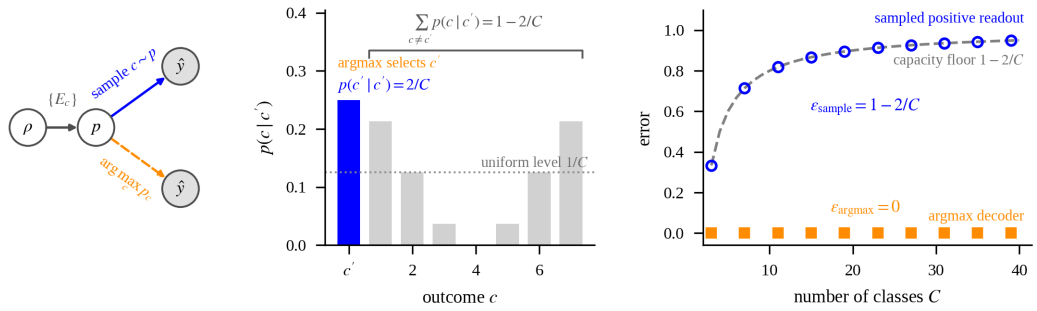
Proof sketch. A deterministic decoder takes at most M values, so it names at most M classes and succeeds with probability at most $\Sigma_M(\pi)$; a randomised one is a mixture of deterministic ones. □

4.2 Decoders outside the positive class

Taking the most probable class instead of sampling replaces the outcome of a positive map by a function of the state, a decoder outside the class the guarantee covers.

Theorem 3 (Decoder boundary)**.** *Fix $C \geq 3$ and a uniform prior. There exist a store of dimension two, an encoder, and a positive readout $\{E_c\}_{c=1}^{C}$ such that the risk of the positive readout equals $1 - 2/C$, meeting the counting floor with equality, while the decoder $D(\rho) = \arg\max_c \mathrm{Tr}(E_c \rho)$ is correct on every input.*

Proof sketch. Encode class c at a pure qubit state whose Bloch vector sits at angle $2\pi c/C$ on a great circle and set $E_c = (2/C)\rho_c$. The Bloch vectors cancel, so the elements sum to I; the success probability is $2/C = \Sigma_2(\pi)$; and $\mathrm{Tr}(E_c \rho_{c'})$ is largest at $c = c'$ alone. □



(a) (b) (c)

Figure 2: The capacity floor depends on the decoder class. (a) Two decision rules applied to the probability vector p of one positive readout. (b) For one state of the two-dimensional witness, the correct class c' carries the largest entry $2/C$ while the wrong classes carry $1 - 2/C$ between them. (c) A sampled positive readout therefore attains the floor $1 - 2/C$ at every class count, whereas argmax over the same probabilities is exact.

Proposition 4 (Argmax margin). *Let every input of class c be encoded at ρ_c, let $\{E_j\}$ be a positive readout, and let $s_{jc} = \mathrm{Tr}(E_j \rho_c)$. The sampled risk $1 - \sum_c \pi_c s_{cc}$ depends on the scores through the weighted diagonal alone, while the argmax decoder is correct on every input under any tie-breaking if and only if $\gamma = \min_c [s_{cc} - \max_{j \neq c} s_{jc}] > 0$.*

Proof sketch. An input of class c is decoded to c under every tie-breaking exactly when s_{cc} exceeds every other entry of column c. □

The floor constrains the diagonal and the margin constrains each column, and Theorem 3 meets both at once: on the witness the diagonal mass equals $\Sigma_2(\pi)$ while $\gamma > 0$ at every class count. Figure 2 traces both decision rules against the floor.

4.3 The entropic floor

Let $\chi = S(\sum_c \pi_c \rho_c) - \sum_c \pi_c S(\rho_c)$ be the Holevo quantity of the label-conditional ensemble, with S the von Neumann entropy in bits. It upper-bounds the label information any positive readout can extract from the store and satisfies $\chi \leq \log_2 m$.

Holevo's bound and Fano's inequality then give a second floor, $g_C^{-1}([H(\pi) - \chi]_+)$, on the risk of every decoder acting on the outcome of a positive readout, where $[a]_+ = \max(a, 0)$, $g_C(\varepsilon) = h(\varepsilon) + \varepsilon \log_2(C - 1)$, and h is the binary entropy. Its dimension-only form, obtained by replacing χ with the ceiling $\log_2 m$, never exceeds the counting floor, so pre-sizing is settled by the counting floor alone. The remaining quantity χ is learned: it is biased in finite samples and enters the floor with a negative sign, so we report a one-sided upper bound χ_U and treat the result as a post-training diagnostic.

4.4 Sizing a store before training

Definition 5 (Minimum admissible dimension). For a target error $\varepsilon^\star$, set $m^\star(\varepsilon^\star) = \min\{m : 1 - \Sigma_m(\pi) \leq \varepsilon^\star\}$.

Definition 5 depends on the evaluation prior alone, which matters whenever a long-tailed training split is evaluated on a balanced test split. That prior is usually estimated, and the error in the estimate bounds the error in the floor.

Proposition 6 (Prior mis-specification). *For any priors π and $\tilde{\pi}$ on C classes and any q, $|\Sigma_q(\pi) - \Sigma_q(\tilde{\pi})| \leq d_{\mathrm{TV}}(\pi, \tilde{\pi}) = \frac{1}{2}\|\pi - \tilde{\pi}\|_1$, their total variation distance, and the inequality is attained.*

Sizing at $\varepsilon^\star - \delta$ on a prior within δ of the deployment prior therefore certifies $\varepsilon^\star$ under the deployment prior, at the cost of a larger store and with no new computation: for a target of 0.10 at $\delta = 0.05$

the robust size is $m^\star(0.05)$ in Table 1, so on long-tailed CIFAR-100 at imbalance factor 100 an estimated prior raises the required dimension from 48 to 61.

5 Error attribution

5.1 The best readout on a learned store

Given the label-conditional states $\{\rho_c\}$ of a trained model, the smallest risk any positive readout attains on that store is

$$\varepsilon_{\mathrm{reg}} \;=\; 1 - \max\left\{ \sum_c \pi_c \operatorname{Tr}(E_c \rho_c) \;:\; E_c \succeq 0,\; \sum_c E_c = I \right\}, \tag{3}$$

which is the Helstrom value for $C = 2$ and a semidefinite program otherwise. We solve Eq. 3 by fixed-point iteration (Ježek et al., 2002) and certify the result against a feasible point of the dual $\min\{\operatorname{Tr} Z : Z \succeq \pi_c \rho_c \; \forall c\}$, so every reported value carries a numerical bracket.

5.2 Representation gap and readout gap

With $\varepsilon_{\mathrm{act}}$ the risk of the decision rule the model deploys, and whenever that rule lies in the class the floor holds for, a positive readout or any decoder of a discrete store, the three quantities satisfy

$$\varepsilon_{\mathrm{cap}} \;\le\; \varepsilon_{\mathrm{reg}} \;\le\; \varepsilon_{\mathrm{act}}, \qquad \varepsilon_{\mathrm{cap}} = 1 - \Sigma_q(\pi), \qquad q = \begin{cases} m & \text{continuous store,} \\ M & \text{discrete store, Corollary 2.} \end{cases} \tag{4}$$

For a continuous store deployed with argmax only $\varepsilon_{\mathrm{cap}} \le \varepsilon_{\mathrm{reg}}$ is guaranteed, and $\varepsilon_{\mathrm{act}} - \varepsilon_{\mathrm{reg}}$ can take either sign. The entropic bound depends on the learned representation through χ_U, so it is reported beside the chain, which keeps the representation gap dependent on the representation at one end.

The two inequalities of Eq. 4 carry the diagnosis. The representation gap $\varepsilon_{\mathrm{reg}} - \varepsilon_{\mathrm{cap}}$ is the excess error of the learned representation over the architecture-only floor. The construction attaining the floor assumes an encoder that separates the classes as needed, so the gap covers encoder inefficiency together with irreducible ambiguity, label noise, and overlap in the input itself. For an out-of-class argmax head the readout gap $\varepsilon_{\mathrm{act}} - \varepsilon_{\mathrm{reg}}$ compares two decoder classes. Within the exact chain it is the excess error of the deployed rule over the best readout and the only gap that can be closed without changing the store, so Algorithm 1 tests it once a capacity limitation has been ruled out.

5.3 Sizing and growth

Sizing is stated separately for the two store kinds, since the counting argument constrains the number of outcomes a store can distinguish and each family realises that number differently. For a continuous store with a positive readout the constrained quantity is the dimension $m^\star$ of Definition 5; for a discrete store it is the cardinality $M^\star$, defined the same way with Σ_M. A target on $M^\star$ translates into a different architectural number in each of the families of Section 3: $K \ge \lceil (M^\star)^{1/T} \rceil$ codebook entries, $n \ge \lceil \log_2 M^\star \rceil$ binary concepts, or $P \ge M^\star$ prototypes under hard assignment.

Each of the three verdicts suggests a corresponding intervention, and two are immediate. A capacity verdict returns $m^\star$ or $M^\star$, and a readout verdict returns the optimal positive readout, which the semidefinite program of the previous section produces alongside its value. An encoder verdict needs a differentiable objective of its own, since the representation gap is a diagnostic quantity. The utilisation $u = \chi_U / \log_2 q$ is maximal exactly when every class-conditional state is pure and their mixture is maximally mixed, both differentiable in the states, so their violation is a regulariser driving the store toward full utilisation,

$$L_{\mathrm{fill}} = \sum_c \pi_c \left(1 - \operatorname{Tr} \rho_c^2\right) + \|\bar{\rho} - I/m\|_F^2 , \tag{5}$$

where $\bar{\rho} = \sum_c \pi_c \rho_c$. The first term vanishes when every class collapses to a single state, the second when those states spread evenly, and Eq. 5 vanishes exactly where $u = 1$. Full utilisation and the floor are separate conditions: the construction attaining the floor in Lemma 1 puts the m heaviest

classes on distinct basis states and collapses the rest onto them, at a mixture far from I/m. The second term therefore asks for more than the floor needs, most of all under a strongly skewed prior. Dropping it leaves the overlap that makes two classes hard to separate,

$$L_{\text{sep}} = \sum_{c \neq c'} \pi_c \pi_{c'} \operatorname{Tr}(\rho_c \rho_{c'}) = \operatorname{Tr} \bar{\rho}^2 - \sum_c \pi_c^2 \operatorname{Tr} \rho_c^2, \tag{6}$$

the cross term of $\|\pi_c \rho_c - \pi_{c'} \rho_{c'}\|_F^2$, whose reduction raises a lower bound on every pairwise trace distance. Two-class discrimination error is one half of one minus that distance, so Eq. 6 targets the overlap ε_{reg} is computed from, while Eq. 5 targets utilisation, which correlates with it.

The capacity verdict of Algorithm 1 is proved by the theory, since $\varepsilon_{\text{cap}} > \varepsilon^\star$ rules the target out whatever the encoder and readout do, and the other two verdicts rest on the measured gaps. The utilisation u is reported beside the verdict, since a store used to its dimension can still have a large representation gap.

5.4 BornReg, a store on which the diagnostic is exact

Every term is available for any store presented as label-conditional states: counts for a discrete store, and for a continuous one the states $zz^\dagger/\|z\|^2$ of Section 3, which keep the direction of z and discard its magnitude. BornReg is the instance whose decision rule lies inside the positive class by construction, so for it the chain is a decomposition of the trained model itself.

Its store is the unit vector $\psi(x) = U_L(x) \cdots U_1(x)\, e_1 \in \mathbb{C}^m$, with e_1 the first basis vector, built by data re-uploading: layer ℓ maps a backbone feature $f(x)$ linearly to r coefficients $w_{\ell k}(x)$ and applies $U_\ell(x) = \exp(\mathrm{i}\, G_\ell(x))$, which is unitary because $G_\ell(x) = G_{\ell 0} + \sum_{k=1}^{r} w_{\ell k}(x) G_{\ell k}$ is Hermitian, the $G_{\ell k}$ being learned Hermitian matrices. The readout is built from unconstrained complex $m \times m$ factors A_c by

$$B_c = A_c A_c^\dagger + \frac{\lambda}{C} I, \qquad B = \sum_c B_c, \qquad E_c = B^{-1/2} B_c B^{-1/2}, \tag{7}$$

so that $E_c \succeq 0$ and $\sum_c E_c = I$ hold exactly, with the regulariser $\lambda > 0$ of Eq. 7 inside each B_c. The register is trained by the negative log-likelihood of its outcome probabilities, so training, evaluation, and deployment use one positive readout, and the floors apply to it exactly as stated.

On a discrete store the floor is certified under every decoder, argmax included; on a continuous store, certification comes from the sampled positive rule or from an architecture whose deployed readout lies in that class, as the register's does.

6 Experiments

6.1 Protocol

Three MedMNIST 2D subsets (Yang et al., 2023) carry the experiments, listed with the others in Table 1. Their evaluation priors are naturally imbalanced, the regime the floor is sensitive to. CIFAR-100 (Krizhevsky, 2009) and its long-tailed variants CIFAR-100-LT at imbalance factors 10, 50, and 100 (Cui et al., 2019) are a controlled check at a class count where the floor stays active across the bottleneck dimensions a practitioner would consider. The stores compared are a codebook, a binary bottleneck of learned latent bits, a prototype bank (Snell et al., 2017), a narrow projection, and the complex register, all on a common ResNet-18 backbone (He et al., 2016); the register is trained at $m = 4$ and $m = 9$ on the medical subsets and at $m = 45$ and $m = 90$ on CIFAR-100. Errors are risks under the evaluation prior, every quantity is a mean over three seeds, and comparisons between methods are descriptive. The diagnostic runs at target $\varepsilon^\star = 0.10$ with gap tolerance $\tau = 0.01$.

6.2 Controlled validation on CIFAR-100

A uniform prior over 100 classes needs dimension 50 to admit an error of 0.5 and 95 to admit 0.05; at imbalance factor 100 those fall to 15 and 61, so the departure from the uniform intuition grows with the skew. Figure 3 shows the curve those dimensions come from, with the exactly decomposed medical models on it: most sit well above their floors, so the dimension has slack there, while those at $q = 4$ on OrganAMNIST sit on an active floor, as does CIFAR-100 below $q = 90$.

Table 1: Datasets and the quantities their class priors alone determine: the prior entropy and the narrowest store that admits each target error, both computed before any training.

Dataset	C	Test	$H(\pi)$	$m^\star(0.10)$	$m^\star(0.05)$
OrganAMNIST	11	17,778	3.322	9	10
DermaMNIST	7	2,005	1.635	4	5
PneumoniaMNIST	2	624	0.954	2	2
CIFAR-100	100	10,000	6.644	90	95
CIFAR-100-LT, imbalance 10	100	10,000	6.357	72	84
CIFAR-100-LT, imbalance 50	100	10,000	5.964	55	68
CIFAR-100-LT, imbalance 100	100	10,000	5.790	48	61

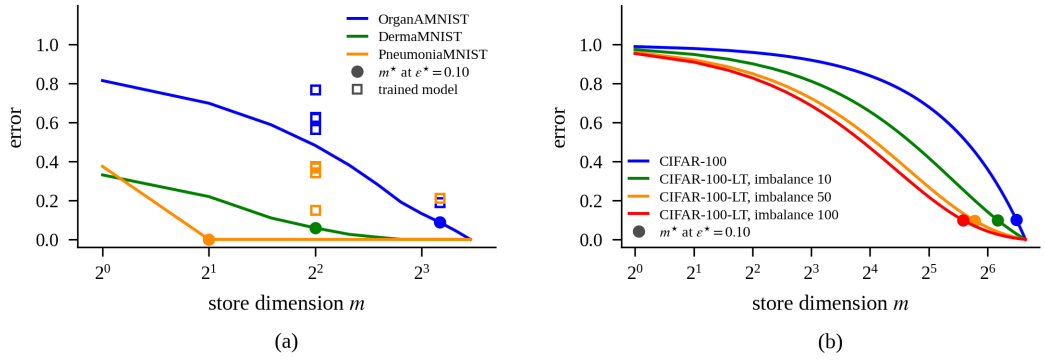


Figure 3: The counting floor against store dimension for each evaluation prior. Filled markers sit at $m^\star$ for $\varepsilon^\star = 0.10$ and open markers at the attained error of trained models with an exact decomposition, above the curve by their two gaps; curves stop at the class count. (a) The medical subsets. (b) CIFAR-100 under the uniform prior and three long-tailed ones.

6.3 Medical classification benchmarks

The MedMNIST subsets have skewed priors and between 2 and 11 classes, so a store of dimension 9 admits the target error on each of them, where CIFAR-100 needs 90. In the 12 configurations of Table 2 with an exact decomposition, the representation gap exceeds the readout gap by at least 0.07 in every one.

6.4 Attribution across bottleneck families

Table 2 reports the chain, both gaps, and the verdict for every head, and Table 5 their reference accuracy. On the exact rows a capacity verdict appears exactly where the floor exceeds the target, for the codebook at $q = 4$ on OrganAMNIST and at $q = 32$ on CIFAR-100; elsewhere the store is admissible and the error lies in its contents. Similar errors can carry different verdicts: on CIFAR-100 the codebook at $q = 32$ and the register at $q = 90$ reach 0.939 and 0.884 and are diagnosed differently, because the codebook sits on an active floor and the register clears it.

The lifted heads compare decoder classes: their argmax risk lies below the best positive readout of the lifted store in all 12 configurations, and at $q = 4$ on OrganAMNIST the prototype bank meets the target on every seed, below a positive-class floor of 0.483, as Theorem 3 allows. Table 3 sets the argmax against the positive readout, the register's own or the best of each lifted store.

Table 4 scores the verdicts as a guide to repair in an exploratory study whose base models were trained for it. The capacity verdict names the winning repair in all 3 decisive OrganAMNIST cases, where the floor is active; over the 23 decisive cases the verdicts name it in 7, against 15 for the best fixed repair of each dataset, chosen post hoc, as the utilisation regulariser behind the encoder verdict raises the error on average. These results scope the claim of this paper to attribution and leave intervention selection open.

Table 2: Error attribution across bottleneck families, averaged over three seeds, with the verdict of Algorithm 1 (the common seed-level one, or mixed) at store size q, a dimension or a cardinality. † marks a continuous head deployed with argmax: its $\varepsilon_{\mathrm{act}}$ is the argmax risk and its $\varepsilon_{\mathrm{cap}}$ and $\varepsilon_{\mathrm{reg}}$ describe positive readouts of its unit-state lift, so $\varepsilon_{\mathrm{act}} - \varepsilon_{\mathrm{reg}}$ can be negative and the row is diagnostic.

Dataset	Method	q	$\varepsilon_{\mathrm{cap}}$	$\varepsilon_{\mathrm{reg}}$	$\varepsilon_{\mathrm{act}}$	$\varepsilon_{\mathrm{reg}} - \varepsilon_{\mathrm{cap}}$	$\varepsilon_{\mathrm{act}} - \varepsilon_{\mathrm{reg}}$	Verdict	Status
CIFAR-100	Codebook	32	0.680	0.927	0.939	0.247	0.012	capacity	exact
	Binary	512	0.000	0.734	0.755	0.734	0.022	encoder	exact
	Prototype†	32	0.680	0.961	0.607	0.281	−0.354	capacity	diagnostic
	Projection†	32	0.680	0.909	0.484	0.229	−0.425	capacity	diagnostic
	Linear†	256	0.000	0.858	0.426	0.858	−0.432	encoder	diagnostic
	Ours (BornReg)	90	0.100	0.848	0.884	0.748	0.036	encoder	exact
DermaMNIST	Codebook	4	0.059	0.308	0.315	0.249	0.007	encoder	exact
	Binary	64	0.000	0.329	0.331	0.329	0.002	encoder	exact
	Prototype†	4	0.059	0.331	0.299	0.272	−0.033	encoder	diagnostic
	Projection†	4	0.059	0.331	0.297	0.272	−0.034	encoder	diagnostic
	Linear†	256	0.000	0.331	0.261	0.331	−0.070	encoder	diagnostic
	Ours (BornReg)	9	0.000	0.315	0.386	0.315	0.071	encoder	exact
OrganAMNIST	Codebook	4	0.483	0.559	0.562	0.076	0.003	capacity	exact
	Binary	64	0.000	0.108	0.134	0.108	0.026	mixed	exact
	Prototype†	4	0.483	0.623	0.073	0.140	−0.550	target met	diagnostic
	Projection†	4	0.483	0.633	0.103	0.150	−0.530	mixed	diagnostic
	Linear†	256	0.000	0.332	0.062	0.332	−0.271	target met	diagnostic
	Ours (BornReg)	9	0.088	0.180	0.190	0.091	0.010	encoder	exact
PneumoniaMNIST	Codebook	4	0.000	0.119	0.152	0.119	0.033	encoder	exact
	Binary	64	0.000	0.130	0.145	0.130	0.015	encoder	exact
	Prototype†	4	0.000	0.375	0.147	0.375	−0.228	encoder	diagnostic
	Projection†	4	0.000	0.359	0.218	0.359	−0.140	encoder	diagnostic
	Linear†	256	0.000	0.366	0.146	0.366	−0.220	encoder	diagnostic
	Ours (BornReg)	9	0.000	0.180	0.213	0.180	0.033	encoder	exact

Table 3: Positive-readout risk minus argmax risk, averaged over configurations, and the Spearman correlation r of the two risks over runs, on the medical subsets.

Dataset	BornReg	Lifted heads	r
OrganAMNIST	0.116	0.454	0.349
DermaMNIST	0.104	0.044	0.632
PneumoniaMNIST	0.028	0.196	−0.024
Overall	0.083	0.231	−0.073

Table 4: Exploratory intervention study: how often the verdict names the repair reducing error most in the decisive cases, against the retrospectively best fixed repair per dataset; the higher count is bold.

		Verdict	Best fixed (post hoc)	
Dataset	Cases	Correct	Repair	Correct
OrganAMNIST	3	**3/3**	capacity	**3/3**
DermaMNIST	4	1/4	capacity	**2/4**
PneumoniaMNIST	8	3/8	capacity	**4/8**
CIFAR-100	8	0/8	readout	**6/8**
Overall	23	7/23 (30%)	per dataset	**15/23 (65%)**

7 Conclusion

In this paper, we treated the bottleneck of a classifier as a store of countable capacity, which lets its width be chosen before training and its error be traced after training to the store, the encoder, or the decision rule. Because the floor needs label counts alone, a model used at several sites can be sized for each site from its own case mix, and Proposition 6 bounds how far the floor moves when that mix drifts, so a deployed store can be re-certified as prevalence changes. Replacing the readout of a continuous head by the register's construction, positive matrices normalised jointly to sum to the identity, places the deployed rule in the positive class and makes the attribution exact. The measured gaps place most of the excess error on the representation side of the decomposition, and how much of that term encoder training can remove is a separate question. The intervention study motivates testing stronger encoder-side objectives, including the separation loss L_{sep}, before verdict-driven repair can be evaluated cleanly. Proposition 4 isolates the margin an argmax exploits, and a floor stated under an assumption linking that margin to the diagonal mass the counting argument constrains would extend the certified chain to the decision rule most deployed classifiers use.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript drafting and editing, methodological discussion, mathematical derivations, and code development. All mathematical claims, code, experiments, references, and results were independently verified by the authors.

REPRODUCIBILITY STATEMENT

Proof sketches appear in the main text, and Appendix A gives the complete proofs, including the entropic floor and its domination by the counting floor. Appendix B describes the solver for the best readout and its dual certificate, Appendix C the hyperparameters and store sizes, Appendix D how to interpret the chain for an argmax head, Appendix E the medical subsets and their priors, and Appendix F the sampled and argmax risks of every continuous head. Appendix G reports the intervention study and Appendix H the utilisation analysis and the numerical checks of the entropic floor and the counterexample. Both constructions are given in closed form and checked numerically by the accompanying test suite. Training saves one archive per model, dataset, and seed, and every number in the main text is computed from those archives with numpy alone, so each floor, each discrimination value, and each verdict can be recomputed without a graphics processor and without repeating the training. The floor and sizing tables need only the label files. The intervention study is the exception, since its stage trains its own models; its per-case records are released so that the tables built from them can be checked. Main comparisons use three seeds.

REFERENCES

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017.

Antonio Almudevar, David Bonet, and Alfonso Ortega. There was never a bottleneck in concept bottleneck models. In *International Conference on Learning Representations*, 2026.

Howard Barnum and Emanuel Knill. Reversing quantum dynamics with near-optimal quantum and classical fidelity. *Journal of Mathematical Physics*, 43(5):2097–2106, 2002.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953.

Ivan Butakov, Alexander Tolmachev, Sofia Malanchuk, Anna Neopryatnaya, Alexey Frolov, and Kirill Andreev. Information bottleneck analysis of deep neural networks via lossy compression. In *International Conference on Learning Representations*, 2024.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K. Su. This looks like that: Deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems*, 2019.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2nd edition, 2006.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Yonina C. Eldar, Alexandre Megretski, and George C. Verghese. Designing optimal quantum detectors via semidefinite programming. *IEEE Transactions on Information Theory*, 49(4):1007–1012, 2003.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Ruslan Galliamov et al. Concepts information bottleneck models. In *International Conference on Learning Representations*, 2026.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Carl W. Helstrom. Quantum detection and estimation theory. *Journal of Statistical Physics*, 1(2): 231–252, 1969.

Geoffrey E. Hinton and Ruslan R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

Alexander S. Holevo. Bounds for the quantity of information transmitted by a quantum communication channel. *Problems of Information Transmission*, 9(3):177–183, 1973.

Miroslav Ježek, Jaroslav Řeháček, and Jaromír Fiurášek. Finding optimal strategies for minimum-error quantum-state discrimination. *Physical Review A*, 65(6):060301, 2002.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*, 2020.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, 2020.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. Deep learning for case-based reasoning through prototypes. In *AAAI Conference on Artificial Intelligence*, 2018.

David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021.

Adrián Pérez-Salinas, Alba Cervera-Lierta, Elies Gil-Fuster, and José I. Latorre. Data re-uploading for a universal quantum classifier. *Quantum*, 4:226, 2020.

Ali Razavi, Aäron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. In *Advances in Neural Information Processing Systems*, 2019.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Maria Schuld, Ryan Sweke, and Johannes Jakob Meyer. Effect of data encoding on the expressive power of variational quantum-machine-learning models. *Physical Review A*, 103(3):032430, 2021.

Ronit Amit Shah. Qudits offer no advantages over dits for sending random messages. Master's thesis, University of Texas at Austin, 2025. Turing Scholars thesis, arXiv:2512.02477.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, 2017.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, 2017.

John Watrous. *The Theory of Quantum Information*. Cambridge University Press, 2018.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. MedMNIST v2: A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Horace P. Yuen, Robert S. Kennedy, and Melvin Lax. Optimum testing of multiple hypotheses in quantum detection theory. *IEEE Transactions on Information Theory*, 21(2):125–134, 1975.

A Complete proofs

This appendix gives in full the two results the main text states as sketches, and restates the other two with their proofs for completeness.

A.1 Lemma 1, the counting floor

Let the store have dimension m and let $\{E_c\}_{c=1}^{C}$ be any positive readout with one element per class, so $E_c \succeq 0$ and $\sum_c E_c = I$. Let ρ_c be the state assigned to class c and π the evaluation prior. The success probability is $\sum_c \pi_c \operatorname{Tr}(E_c \rho_c)$.

Positivity and $\operatorname{Tr} \rho_c = 1$ give $\operatorname{Tr}(E_c\rho_c) \le \|E_c\|_\infty \le \operatorname{Tr} E_c$, and also $\operatorname{Tr}(E_c\rho_c) \le 1$, so $\operatorname{Tr}(E_c\rho_c) \le \min(1, \operatorname{Tr} E_c)$. With $t_c = \operatorname{Tr} E_c \ge 0$, the identity $\sum_c E_c = I$ gives $\sum_c t_c = \operatorname{Tr} I = m$. The success probability is therefore at most

$$\max \Big\{ \textstyle\sum_c \pi_c \min(1, t_c) \;:\; t_c \ge 0,\ \sum_c t_c = m \Big\}.$$

The objective is separable and concave in each t_c, and increasing until $t_c = 1$, so an optimum places unit mass on as many classes as the budget allows, in decreasing order of π_c: the greedy allocation $t_c = 1$ for the m largest entries of π and $t_c = 0$ elsewhere. Its value is $\Sigma_m(\pi)$, so the risk is at least $1 - \Sigma_m(\pi)$.

For attainment, index the classes so that $\pi_1 \ge \cdots \ge \pi_C$, take an orthonormal basis $\{|i\rangle\}_{i=1}^{m}$, set $\rho_c = |c\rangle\langle c|$ for $c \le m$ and $\rho_c = |m\rangle\langle m|$ for $c > m$, and let $E_c = |c\rangle\langle c|$ for $c < m$, $E_m = |m\rangle\langle m|$, and $E_c = 0$ for $c > m$. These are positive and sum to the identity on the span, and the success probability is exactly $\Sigma_m(\pi)$. The construction follows the one-shot setting (Shah, 2025), with one readout element per class.

A.2 Proposition 6, prior mis-specification

For a set of classes J let $\pi(J) = \sum_{c \in J} \pi_c$, so that $\Sigma_q(\pi) = \max_{|J|=q} \pi(J)$ is a maximum of finitely many linear functionals of π. Let J attain that maximum for π, so that $\Sigma_q(\pi) = \pi(J)$ while $\Sigma_q(\tilde\pi) \ge \tilde\pi(J)$. Then

$$\Sigma_q(\pi) - \Sigma_q(\tilde\pi) \;\le\; \pi(J) - \tilde\pi(J) \;\le\; \max_{J' \subseteq \{1,\dots,C\}} |\pi(J') - \tilde\pi(J')| \;=\; d_{\mathrm{TV}}(\pi, \tilde\pi),$$

and exchanging the two priors gives the other direction. The bound is attained: at $C = 3$ and $q = 2$ with $\pi = (0.5, 0.3, 0.2)$ and $\tilde\pi = (0.4, 0.3, 0.3)$ both sides equal 0.1. Since $\varepsilon_{\mathrm{cap}} = 1 - \Sigma_q$, the same bound holds for the floor.

For sizing, let $\hat\pi$ be the estimate the store is sized on and π the deployment prior, and suppose $d_{\mathrm{TV}}(\hat\pi, \pi) \le \delta$ and the store satisfies $1 - \Sigma_q(\hat\pi) \le \varepsilon^\star - \delta$. Then $1 - \Sigma_q(\pi) \le 1 - \Sigma_q(\hat\pi) + \delta \le \varepsilon^\star$, so the floor under the deployment prior meets the target. The rule applies for $\delta < \varepsilon^\star$, and it bounds the floor, the part of the risk the store size and the prior fix.

A.3 Corollary 2, discrete stores

A discrete store of cardinality M emits one of M symbols, and every decoder factors through that symbol. A deterministic decoder assigns one class to each symbol, so it names at most M distinct classes, and its success probability is the total prior mass of the classes it names, which is at most $\Sigma_M(\pi)$ by the greedy argument above. A randomised decoder is a mixture of deterministic ones, and a mixture cannot exceed the best element of its support, so the same bound holds. The risk is therefore at least $1 - \Sigma_M(\pi)$ with no restriction on the decoder class.

A.4 THEOREM 3, THE DECODER BOUNDARY

Place the Bloch vectors $b_c = (\sin\theta_c, 0, \cos\theta_c)$ at angles $\theta_c = 2\pi c/C$ and let v_c be the corresponding unit vectors in $\mathbb{C}^2$, so that $v_c v_c^\dagger = (I + b_c \cdot \sigma)/2$, where $\sigma = (\sigma_x, \sigma_y, \sigma_z)$ are the Pauli matrices. Encode class c at $\rho_c = v_c v_c^\dagger$ and set $E_c = (2/C)\, v_c v_c^\dagger$. The elements are positive by construction, and they sum to the identity because $\sum_c b_c = 0$ for Bloch vectors at equal angles, so $\sum_c E_c = (2/C)\sum_c (I + b_c \cdot \sigma)/2 = I$. Under the uniform prior the success probability is $\sum_c (1/C)\,\mathrm{Tr}(E_c\rho_c) = 2/C = \Sigma_2(\pi)$, so the sampled risk meets the counting floor.

For the decoder, $\mathrm{Tr}(E_c\rho_{c'}) = (1 + \cos(\theta_c - \theta_{c'}))/C$ is strictly maximised over c at $c = c'$ for $C \geq 3$, since cos is strictly decreasing on $[0, \pi]$ and the angles are distinct modulo 2π, so every input is classified correctly. At $C = 2$ the two states are antipodal, which is why the statement fixes $C \geq 3$.

A.5 PROPOSITION 4, THE ARGMAX MARGIN

With every input of class c encoded at ρ_c, the positive readout returns j on such an input with probability s_{jc}, so the sampled risk is $1 - \sum_c \pi_c s_{cc}$ and involves the diagonal alone. The argmax decoder returns an index maximising s_{jc} over j; it returns c under every tie-breaking exactly when $s_{cc} > s_{jc}$ for every $j \neq c$, and this holds for every class exactly when $\gamma > 0$. On the witness of Theorem 3, $s_{cc} = 2/C$ and $s_{jc} = (1 + \cos(\theta_j - \theta_c))/C$, so $\gamma = (1 - \cos(2\pi/C))/C > 0$ while $\sum_c \pi_c s_{cc} = \Sigma_2(\pi)$.

A.6 THE ENTROPIC FLOOR AND ITS DOMINATION

Theorem 7 (Entropic floor)**.** *Let $g_C(\varepsilon) = h(\varepsilon) + \varepsilon \log_2(C-1)$, with h the binary entropy, which is increasing on $[0, 1 - 1/C]$. Every decoder acting on the outcome of a positive readout has risk at least $g_C^{-1}([H(\pi) - \chi]_+)$, where $[a]_+ = \max(a, 0)$ and $g_C^{-1}(0) = 0$.*

Proposition 8 (The dimensional entropic floor is dominated by the counting floor)**.** *For every prior and every dimension, $g_C^{-1}([H(\pi) - \log_2 m]_+) \leq 1 - \Sigma_m(\pi)$.*

For Theorem 7, let $\hat{Y}$ be the class a decoder outputs from the outcome of a positive readout and $\varepsilon = \Pr(\hat{Y} \neq Y)$ its risk. The Holevo bound gives $I(Y;\hat{Y}) \leq \chi$ for the mutual information between label and output, since the outcome is a classical random variable obtained from the ensemble. Fano's inequality bounds the conditional entropy, $H(Y \mid \hat{Y}) \leq g_C(\varepsilon)$ with $g_C(\varepsilon) = h(\varepsilon) + \varepsilon \log_2(C-1)$, so $H(\pi) - \chi \leq H(\pi) - I(Y;\hat{Y}) = H(Y \mid \hat{Y}) \leq g_C(\varepsilon)$. Since g_C is increasing on $[0, 1 - 1/C]$ and $g_C(0) = 0$, inverting gives $\varepsilon \geq g_C^{-1}([H(\pi) - \chi]_+)$, where the positive part handles the case $\chi > H(\pi)$, in which the bound is zero.

For the domination, the readout that reports the m largest classes and a fixed class otherwise satisfies $I(Y;\hat{Y}) \leq \log_2 m$ and has error exactly $1 - \Sigma_m(\pi)$. Fano applied to it gives $g_C(1 - \Sigma_m(\pi)) \geq H(\pi) - \log_2 m$, and monotonicity of g_C gives $g_C^{-1}([H(\pi) - \log_2 m]_+) \leq 1 - \Sigma_m(\pi)$.

B COMPUTING THE BEST READOUT

The term $\varepsilon_{\mathrm{reg}}$ is the value of the minimum-error discrimination problem

$$\min_{\{E_c\}} \; 1 - \sum_c \pi_c \,\mathrm{Tr}(E_c\rho_c) \quad \text{subject to} \quad E_c \succeq 0,\ \textstyle\sum_c E_c = I,$$

a semidefinite program in the E_c. For $C = 2$ the Helstrom measurement gives it in closed form. For $C > 2$ we run the fixed-point iteration for minimum-error discrimination (Ježek et al., 2002): starting from the pretty-good measurement $E_c = \bar{\rho}^{-1/2}\pi_c\rho_c\bar{\rho}^{-1/2}$ with $\bar{\rho} = \sum_c \pi_c\rho_c$, the update $E_c \leftarrow \Lambda^{-1/2}\pi_c\rho_c E_c \pi_c\rho_c\Lambda^{-1/2}$ with $\Lambda = \sum_c \pi_c\rho_c E_c \pi_c\rho_c$, which keeps $\sum_c E_c = I$, is applied until 50 consecutive updates improve the value by less than 10^{-11} or 2000 iterations pass, whichever comes first.

Every reported value is certified independently of how the iteration converged. At the fixed point we take the Hermitian part Z of $\sum_c \pi_c\rho_c E_c$ and the smallest $s \geq 0$ with $Z + sI \succeq \pi_c\rho_c$ for every c, so

that $Z + sI$ is a feasible point of the dual of the problem above whatever the iteration did. Its value gives the lower bound $1 - \operatorname{Tr} Z - sm$ on $\varepsilon_{\mathrm{reg}}$, and the gap between it and the primal value bounds the error in $\varepsilon_{\mathrm{reg}}$. The experiments report that gap; over the 72 configurations for which a certificate is defined its largest value is 1.4×10^{-6}, so every reported $\varepsilon_{\mathrm{reg}}$ carries a numerical bracket.

Cost is governed by C and m through the C matrix square roots of size m taken per iteration. On one core and for ensembles of the stated shape, a store of dimension $m = 9$ at $C = 11$ converges in 63 iterations and 0.07 seconds and $m = 64$ at $C = 100$ in 69 iterations and 10 seconds. The cost climbs steeply with the dimension at a fixed class count: $m = 90$ at $C = 100$ takes 615 iterations and under three minutes, and $m = 256$ needs more than a thousand iterations, the regime where a general-purpose interior-point solver becomes the better choice. The widest stores reported here are the linear head at $m = 256$ and the register at $m = 90$, which cost minutes per trained model; every other configuration costs seconds, against hours to train any of them. The two-class case is closed form.

Continuous stores are lifted to states by $z \mapsto zz^{\dagger}/\|z\|^2$. A row with $\|z\|$ below 10^{-12}, which a rectified projection can produce, is mapped to the uniform state $(1, \dots, 1)/\sqrt{m}$, which carries no class information and keeps the analysis running through a dead unit. Each ρ_c is the mean state of the held-out inputs of class c, so a rare class gives a noisier estimate of ρ_c and hence of $\varepsilon_{\mathrm{reg}}$; shrinkage toward the mixture $\bar{\rho}$ is a natural extension for classes with few samples.

C Hyperparameters

The diagnostic uses $\varepsilon^{\star} = 0.10$ and gap tolerance $\tau = 0.01$, and the intervention is scored at margin $\tau_{\mathrm{int}} = 0.005$, so two arms separated by less than that count as a tie. The Holevo bound χ_U is the upper 95% quantile of a stratified bootstrap over 400 resamples of the test split, stratified by class so that each resample preserves the evaluation prior.

Training uses AdamW at learning rate 10^{-3}, weight decay 5×10^{-4}, batch size 128, and a ResNet-18 backbone adapted to small images by replacing the stem with a 3×3 convolution at stride one and removing the initial pooling. The register of Section 5.4 uses $L = 4$ layers, coefficients of rank $r = 8$, and readout regulariser $\lambda = 10^{-4}$. Main runs use 15 epochs and three seeds. The interventions each receive 5 further epochs from the same weights, so every arm sees the same number of optimiser updates over the same examples. The utilisation regulariser enters the encoder arm at weight 0.01: larger weights raised the task loss by more than the regulariser recovered, and the weight was selected on a validation split.

The attribution runs use the method-specific store-size settings listed in the second column of Table 5, chosen to straddle the minimum admissible dimension of each prior: the register is trained at $m = 4$ and $m = 9$ on the medical subsets and at $m = 45$ and $m = 90$ on CIFAR-100, and each comparison head at the dimension its own family makes comparable. The linear head has no bottleneck of its own and serves as the unbottlenecked reference. The intervention study of Appendix G uses $m = 4$ on the medical subsets and $m = 45$ on CIFAR-100 throughout.

Table 5: Reference argmax accuracy as mean and standard deviation over three seeds, the highest mean per dataset in bold. Each head is evaluated at the store-size setting it was trained with, given in the second column as q and defined as in Table 2, with the medical and CIFAR-100 settings separated by a slash.

Method	q	OrganAMNIST	DermaMNIST	PneumoniaMNIST	CIFAR-100
Codebook	4/32	0.438 ± 0.031	0.685 ± 0.010	0.848 ± 0.014	0.065 ± 0.005
Binary	64/512	0.866 ± 0.075	0.669 ± 0.000	$\mathbf{0.855 \pm 0.005}$	0.227 ± 0.028
Prototype	4/32	0.927 ± 0.003	0.701 ± 0.010	0.853 ± 0.002	0.393 ± 0.009
Projection	4/32	0.897 ± 0.057	0.703 ± 0.017	0.782 ± 0.136	0.516 ± 0.001
Linear	256	$\mathbf{0.938 \pm 0.002}$	$\mathbf{0.739 \pm 0.003}$	0.854 ± 0.004	$\mathbf{0.574 \pm 0.001}$
Ours (BornReg)	9/90	0.885 ± 0.030	0.702 ± 0.032	0.817 ± 0.016	0.273 ± 0.087

D THE CHAIN FOR AN ARGMAX HEAD

Most deployed classifiers take an argmax, and on a discrete store the floor is a bound whatever the decision rule, so a codebook or a binary bottleneck needs no adjustment. For a continuous store the two computed terms are the floor and the optimum of the positive class, and an argmax head lies outside that class: $\varepsilon_{\mathrm{cap}}$ and $\varepsilon_{\mathrm{reg}}$ then serve as diagnostics, $\varepsilon_{\mathrm{act}} - \varepsilon_{\mathrm{reg}}$ can take either sign, and a capacity verdict describes the positive class. The experiments mark those heads with a dagger in Table 2 and report their argmax risk as $\varepsilon_{\mathrm{act}}$. A certified result on a continuous store needs the head evaluated under the sampled rule, at the cost of one extra pass over the test split; the register is evaluated that way by construction.

E MEDICAL BENCHMARK DETAILS

Table 6: The three medical subsets, their splits and priors, and the regime of their verdicts. The prior ratio is the heaviest class mass divided by the lightest.

Dataset	Modality	C	Train	Test	Ratio	$H(\pi)$	$m^\star(0.10)$	Regime
OrganAMNIST	abdominal CT	11	34,581	17,778	4.2	3.322	9	capacity at q=4
DermaMNIST	dermatoscopy	7	7,007	2,005	58.3	1.635	4	encoder
PneumoniaMNIST	chest radiography	2	4,708	624	1.7	0.954	2	encoder

The three subsets are used as distributed in MedMNIST (Yang et al., 2023) at 28×28, with the official train, validation, and test splits and no resizing or augmentation. Grayscale subsets are repeated to three channels and scaled to $[-1, 1]$; the evaluation prior of Table 6 is the empirical class distribution of the test split.

The three were chosen to occupy distinct regimes of the theory, and the last column of Table 6 gives the regime their verdicts fall into, which turns with the store size on OrganAMNIST and stays fixed on the other two. OrganAMNIST classifies 11 organs in abdominal computed tomography (CT) and has a mildly uneven prior, so $m^\star(0.10) = 9$ and a store narrower than that is ruled out before training: for the models at $m = 4$ the floor is 0.483, and the codebook and the register trained at that size are certified dimension limited, while at $m = 9$ it falls to 0.088 and the verdict of the register turns to the encoder. It is also the only subset where an intervention moves the error substantially, with widening winning all 3 decisive cases. DermaMNIST is the skewed case, one class carrying 0.669 of the mass and the lightest 0.011, a ratio of 58; the skew makes $m^\star(0.10)$ only 4 despite 7 classes, so every store in the study is admissible and every run returns an encoder verdict; over its 18 exact runs the mean representation gap is 0.288 against a readout gap of 0.035. PneumoniaMNIST is the binary limiting case the introduction names: with $C = 2$ the floor is zero at every $m \geq 2$, so the framework rules widening out on paper and every run returns an encoder verdict.

Together, the three show the sizing rule doing the work the prior gives it. The subset with the most classes needs the widest store; the subset with the most skew needs a narrower store than its class count suggests; and the binary subset needs nothing beyond $m = 2$.

These are benchmark classification datasets at low resolution. The verdicts describe the trained models reported above, and clinical deployment and decision-making on patients are outside the scope of this study.

F SAMPLED VERSUS ARGMAX READOUTS

Certifying the chain needs a positive readout, and most deployed classifiers take an argmax, so the two are compared here on the trained continuous stores. Table 7 gives both risks and their difference Δ_{dec} for every configuration, and Table 3 summarises them by dataset.

The 15 configurations carry 72 runs between them: the register was trained at two store dimensions with three seeds each, and every other continuous head ignores the requested dimension and so contributes 6 runs at one architecture, of which Table 2 reports the 3 requested at $m = 9$. The positive readout errs more than the argmax in 70 of the 72 runs, and the two others are PneumoniaMNIST projections at the majority-class error of 0.375. For the register, whose own rule is the sampled

one, the difference reaches 0.157, so evaluating it under that rule is conservative, and for a lifted head it reaches 0.551 on OrganAMNIST, where a prototype bank at $q = 4$ attains 0.071 by argmax against 0.623 for the best positive readout of its lifted store. The weak rank agreement in Table 3 is Theorem 3 appearing on trained models and the reason the lifted heads carry a dagger in Table 2. The register is built for this case: on a store whose decision rule already lies in the class the floor holds for, the certified evaluation and the deployed one coincide.

Table 7: Risk of the positive readout and of the argmax on the continuous stores, averaged over seeds. The sampled column is the register's own positive readout and, for a head deployed with argmax (†), the best positive readout of its lifted store, $\varepsilon_{\mathrm{reg}}$.

Dataset	Method	q	Sampled	Argmax	Δ_{dec}
OrganAMNIST	Ours (BornReg)	4	0.768	0.612	0.157
	Ours (BornReg)	9	0.190	0.115	0.075
	Linear†	256	0.332	0.062	0.269
	Projection†	4	0.628	0.086	0.542
	Prototype†	4	0.623	0.071	0.551
DermaMNIST	Ours (BornReg)	4	0.449	0.328	0.121
	Ours (BornReg)	9	0.386	0.298	0.088
	Linear†	256	0.331	0.261	0.070
	Projection†	4	0.331	0.297	0.034
	Prototype†	4	0.331	0.303	0.028
PneumoniaMNIST	Ours (BornReg)	4	0.344	0.317	0.027
	Ours (BornReg)	9	0.213	0.183	0.029
	Linear†	256	0.366	0.146	0.220
	Projection†	4	0.362	0.220	0.142
	Prototype†	4	0.375	0.149	0.226

G THE INTERVENTION STUDY

Each verdict suggests a corresponding repair, so we tested whether the diagnosed component predicts the intervention that reduces error most. The test runs on the continuous-store models, which receive all three interventions from matched starting weights and over the same number of updates. The capacity arm widens the store and continues training backbone and head; the encoder arm continues training both at the same dimension with L_{fill} added to the task loss; the readout arm freezes the backbone and refits the head on the store it holds. The codebook and the binary bottleneck appear in the main tables, and L_{sep} was added after this study was run. Table 8 gives every case behind Table 4; against its 7 of 23, chance gives 1 in 3 and widening everywhere 9 of 23.

The arm-level gains explain these counts: averaged over the 39 runs, widening reduces error by 0.018 and refitting the head by 0.004, while the utilisation regulariser raises it by 0.010 and improves 15 of the 39. The encoder verdict is returned on 27 and is scored against that regulariser, so it is correct on 4 of its 15 decisive cases, against 3 of 8 for the capacity verdict. The one configuration with a floor far from zero, OrganAMNIST at $m = 4$ with $\varepsilon_{\mathrm{cap}} = 0.483$, is where widening wins on every seed, as the capacity verdict predicts.

A clean test of intervention selection needs two changes to this stage. The first is an encoder objective that lowers the error by itself, as widening and refitting do. The second is base models trained like the rest of the paper: this stage trained its own, without the cosine schedule used for the archives behind Tables 2 and 5, and at OrganAMNIST $m = 4$ the register's base error over three seeds is 0.284, 0.327 and 0.611, a spread wider than the differences the arms are scored on.

The released implementation resolves the training-schedule mismatch, adds L_{sep} as a second encoder objective, and admits the discrete stores, translating a diagnosed cardinality into the architectural parameter of Section 5.3. Tables 4 and 8 report the study as run with the earlier code, and their per-case records are released for verification. Rerunning the study with the released implementation tests intervention selection on base models trained like the rest of the paper, and with it whether L_{sep} provides an effective encoder intervention.

Table 8: Per-seed intervention results: the verdict and the error reduction each intervention produced, the winner in bold. A case whose best two arms differ by less than $\tau_{\text{int}} = 0.005$ is a tie, with every arm within τ_{int} of the best in bold, and is left out of the score.

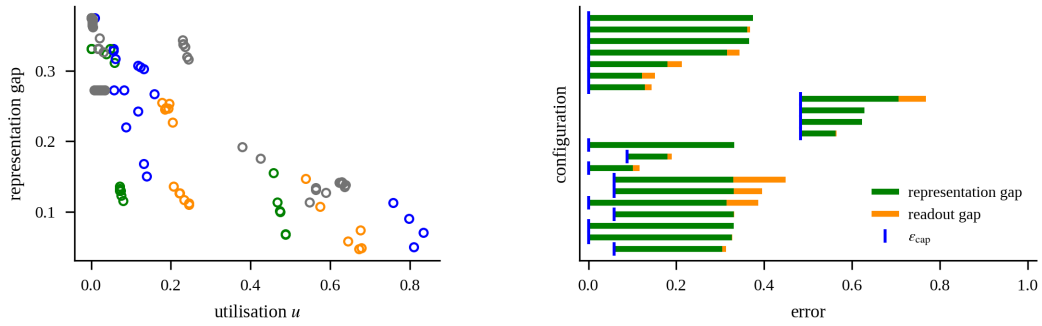
Dataset	Method	Seed	Verdict	Capacity	Encoder	Readout
OrganAMNIST	Ours (BornReg)	0	capacity	**+0.204**	−0.140	−0.037
		1	capacity	**+0.206**	−0.016	−0.008
		2	capacity	**+0.208**	−0.224	+0.013
DermaMNIST	Ours (BornReg)	0	encoder	−0.022	**−0.004**	**−0.004**
		1	encoder	−0.005	**+0.003**	−0.002
		2	encoder	**+0.003**	**+0.004**	**+0.003**
	Linear	0	encoder	−0.021	**+0.013**	**+0.014**
		1	encoder	−0.017	**−0.001**	**+0.001**
		2	encoder	**+0.002**	−0.007	**+0.005**
	Projection	0	encoder	**−0.001**	−0.023	**−0.000**
		1	encoder	**+0.033**	+0.018	+0.008
		2	encoder	**+0.004**	−0.010	**+0.000**
	Prototype	0	encoder	−0.051	−0.029	**−0.004**
		1	encoder	**−0.004**	**−0.004**	**−0.005**
		2	encoder	**+0.010**	+0.004	−0.002
PneumoniaMNIST	Ours (BornReg)	0	encoder	+0.019	**+0.046**	+0.006
		1	encoder	**+0.000**	**+0.000**	**+0.000**
		2	encoder	+0.010	+0.018	**+0.067**
	Linear	0	encoder	**−0.021**	−0.038	**−0.022**
		1	encoder	−0.067	**−0.050**	**−0.050**
		2	encoder	**+0.045**	+0.011	+0.014
	Projection	0	encoder	**+0.162**	+0.000	+0.000
		1	encoder	**+0.048**	+0.021	+0.022
		2	encoder	−0.026	**+0.006**	+0.000
	Prototype	0	encoder	**+0.006**	**+0.003**	−0.022
		1	encoder	−0.088	**−0.005**	−0.016
		2	encoder	**−0.006**	−0.040	−0.030
CIFAR-100	Ours (BornReg)	0	capacity	**−0.002**	**−0.001**	**−0.001**
		1	capacity	**−0.001**	**−0.001**	**−0.000**
		2	capacity	**−0.000**	**−0.000**	**−0.001**
	Linear	0	encoder	+0.001	−0.014	**+0.029**
		1	encoder	+0.023	+0.017	**+0.042**
		2	encoder	−0.003	+0.006	**+0.032**
	Projection	0	capacity	−0.000	−0.002	**+0.022**
		1	capacity	+0.009	−0.005	**+0.021**
		2	capacity	−0.011	−0.006	**+0.015**
	Prototype	0	capacity	+0.014	**+0.021**	+0.012
		1	capacity	+0.013	**+0.022**	+0.015
		2	capacity	**+0.015**	−0.001	**+0.012**

H ADDITIONAL RESULTS

Figure 4 bears on the status of the regulariser. Across the 108 runs on the three medical subsets the utilisation correlates with the representation gap at -0.79, and within OrganAMNIST, DermaMNIST, and PneumoniaMNIST at -0.87, -0.61, and -0.84, so a store used closer to its dimension does tend to have a smaller representation gap. Over the 54 runs with an exact decomposition, its correlation with the readout gap is -0.27 and changes sign between datasets, which is the contrast the regulariser needs: it targets the first gap and leaves the second to the readout. Since the attainment construction of Lemma 1 reaches the floor at a mixture far from I/m, the regulariser remains an intervention to be tested against the gap it targets.

Subtracting the entropic floor at the architectural ceiling from the counting floor gives a non-negative difference at every prior and dimension examined, which confirms Proposition 8 numerically. The

counterexample of Theorem 3 is checked for every class count from 3 to 40: the sampled risk equals $1 - 2/C$ to machine precision and the argmax decoder is exact at each one.



(a) (b)

Figure 4: (a) The representation gap against the utilisation u of the store, one point per run. This is the gap the utilisation tracks, and the side of the decomposition the regulariser is aimed at. (b) The chain for every configuration with an exact decomposition, averaged over seeds: the tick is $\varepsilon_{\mathrm{cap}}$, the first segment the representation gap, the second the readout gap, so the attained error is the right-hand end.