OUTPUT CODING IN MULTICLASS CLASSIFICATION UNDER SHARED BOTTLENECKS

Anonymous authors
Paper under double-blind review

ABSTRACT

In output coding for multiclass classification, each class is assigned a binary codeword, and the model learns the corresponding bits jointly from a shared feature representation. When the shared representation is constrained to be low-dimensional, a representation bottleneck, the binary prediction tasks compete for limited representational capacity, so the choice of code can affect which distinctions between classes are preserved. We study this effect under squared-loss training with a shared low-rank representation. Our goal is not to show that output coding universally improves classification, but to understand how the choice of code shapes the representation learned under this bottleneck. Even when two codes can represent exactly the same multiclass scoring rules, they can weight the training objective differently and therefore select different feature subspaces. We characterize this selected subspace and show that the resulting increase in classification error is proportional to the loss of Gaussian width, a geometric measure of the class structure removed by the projection. We also show that a linear decoder achieves the lowest classification error possible using only the retained representation. To show that conventional code properties do not determine which class information survives the bottleneck, for each bottleneck rank we construct pairs of binary codes with matched conventional code properties and binary-task difficulty, yet whose learned representations retain disjoint feature directions and have provably different multiclass risks. Moreover, the risk gap is at least $3/35$ of the improvement the Bayes classifier achieves over the best classifier that ignores the features, and this separation persists under small perturbations. Frozen-feature experiments further show that sensitivity to code choice can be substantial and task-dependent: coding improves over one-hot regression on Fashion-MNIST, but not consistently on CIFAR-100. These results show that conventional code properties alone do not determine which class information survives a shared representation bottleneck.

1 INTRODUCTION

In multiclass classification, a model that predicts the class from a shared low-dimensional representation, a representation bottleneck, cannot preserve all information about the classes equally well. We are particularly interested in *class distinctions*: information that changes the relative scores of classes and can therefore affect which class is predicted. One way to formulate multiclass prediction is through output coding, where each class is assigned a binary codeword and the model learns the corresponding binary targets jointly from the shared representation. Output codes are commonly characterized using properties such as separation between codewords, the geometry of the code matrix, and the difficulty of the induced binary prediction tasks (Dietterich & Bakiri, 1995; Allwein et al., 2000). We refer to such properties collectively as *code summaries*.

Although output coding is classical, its role changes when all coded outputs must rely on the same low-dimensional representation: the binary tasks become coupled before decoding and must share a limited set of feature directions. We model this constraint with a hard rank limit. Changing the output code changes the binary targets and therefore how squared loss weights different class distinctions, which can change the selected subspace. This leads to our central question: *with the input–label distribution and representation dimension fixed, can the choice of output code change which class distinctions are preserved by training, even when the compared codes can represent exactly the same mappings from features to class scores?*

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 answer this question using a rank-constrained map from features to coded outputs, trained by expected squared loss. We characterize the feature subspace selected by each code, relate the information lost under projection to multiclass classification error, and construct matched codes that select different subspaces and attain different risks.

In summary, our main contributions are the following:

How code choice determines the retained feature directions and classification error. For squared-loss training under a rank-r bottleneck, we derive the optimal predictor and show that the choice of code determines which r shared feature directions are retained by changing how different class distinctions are weighted during training. We then consider a tractable model in which the features are uniformly distributed on a sphere and the class probabilities depend affinely on those features, and show that the increase in classification error caused by the bottleneck is exactly proportional to the Gaussian width lost by projection onto the retained subspace. We further prove that the standard linear score decoder achieves the lowest classification error possible from the retained representation (Theorem 2).

A separation between conventional code properties and retained class information. For every bottleneck rank, we construct pairs of binary codes that match standard measures of code geometry and have the same optimal error for each corresponding binary prediction task. Despite these matched properties, squared-loss training with the two codes retains subspaces with no nonzero direction in common and yields provably different multiclass risks. The risk gap is at least $3/35$ of the improvement the Bayes classifier achieves over the best classifier that ignores the features, and the separation persists under sufficiently small perturbations that preserve the matching conditions (Theorem 4).

Empirical evidence of substantial and task-dependent sensitivity to code choice. Using frozen image features, we compare coded and one-hot training using the same training examples, bottleneck rank, and downstream decoder. On Fashion-MNIST, the selected code lowers test error by 11.39 percentage points relative to the matched one-hot low-rank baseline with 3,000 labeled examples, whereas on CIFAR-100 it performs worse with 2,000 labeled examples and shows no clear difference with 6,000 labeled examples. These experiments test whether code choice can materially affect the learned representation, rather than whether output coding is universally better than one-hot training. We also compare complete methods when each is given the same total number of labeled examples, including a low-rank softmax baseline, to separate the effect of code choice from differences in the overall training procedure.

1.1 Related work

Output coding for multiclass classification. Error-correcting output codes (ECOCs) reduce multiclass prediction to a collection of binary prediction tasks by assigning each class a codeword. Classical work established the connection between codeword separation, binary losses, decoding rules, and multiclass performance (Dietterich & Bakiri, 1995; Allwein et al., 2000), while subsequent approaches have developed alternative class encodings and data-dependent code designs that account for relationships or confusability among classes and component tasks (Crammer & Singer, 2002; Bautista et al., 2018; Liu et al., 2016). Calibration theory provides a complementary perspective by characterizing the prediction dimensions required for consistent multiclass surrogate losses (Ramaswamy & Agarwal, 2016). These lines of work primarily study the code, the component prediction problems, or the final decoding stage; our focus is instead on how the code changes the shared feature representation selected during training.

Output coding with neural and shared representations. More recent work has incorporated output coding into deep neural classifiers for probability estimation, adversarial robustness, ensemble learning, and robust code design (Verma & Swami, 2019; Zhang et al., 2020a; Song et al., 2021; Gupta & Amin, 2021; Yu et al., 2024). In particular, parameter-sharing architectures have been proposed to reduce the cost of learning multiple coded tasks (Zhang et al., 2020a), and end-to-end ECOC training has been used to couple the component classifiers rather than training them independently (Song et al., 2021). Recent theoretical work has also revisited code design for neural networks in the presence of weight perturbations (Yu et al., 2024). These studies show that output coding continues to be used with neural classifiers, including settings in which the coded prediction tasks share parameters or are trained jointly end to end. However, they mainly study predictive accuracy, robustness, ensemble

construction, or code design. We study a different question: when all coded outputs must rely on the same low-rank feature representation, how does the choice of code affect which class-relevant feature directions are retained by training?

Regression and dimension reduction. Reduced-rank regression and optimal scoring connect low-rank prediction with multiclass discrimination by forcing multiple outputs to depend on a common lower-dimensional representation (Izenman, 1975; Hastie et al., 1994). Rank-constrained estimation has also been used to learn low-dimensional label embeddings (Mineiro & Karampatziakis, 2015), and recent work studies regression-based multiclass discriminants with excess-risk guarantees (Bing et al., 2026). More broadly, sufficient dimension reduction seeks low-dimensional projections that preserve information relevant to the response (Cook & Yin, 2001; Zhang et al., 2020b), with recent work considering how such directions can be ordered by predictive relevance (Boonstra et al., 2025). Our setting connects these ideas to output coding: rather than choosing the reduced representation directly, we ask how changing the coded targets changes the subspace selected by squared-loss training, and how much classification information that selected subspace retains.

2 Problem Formulation

Inputs, labels, and the code matrix Let $U \in \mathbb{R}^d$ be the input, $Y \in \{1, \dots, C\}$ its class label, and u, y their realizations. For example, in vision applications, u can be a fixed encoder embedding. Then, we fix a code length n and a code matrix $M \in \{-1, +1\}^{C \times n}$. Row c is the codeword for class c; column j defines the binary target M_{Yj}. All n tasks share the input u: M changes the label targets, not the features. With $e_y \in \{0, 1\}^C$ denoting the one-hot label, we encode y as

$$z_y = M^\top e_y = (M_{y1}, \dots, M_{yn})^\top \in \{-1, +1\}^n.$$

To illustrate, we consider a four-class problem ($C = 4$) with classes *cat*, *dog*, *car*, and *truck*, ordered by rows. For code length $n = 3$, consider $M = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \end{pmatrix}$. The first column separates animals from vehicles; the others define alternative balanced partitions. A *dog* example has target $z_2 = (1, -1, -1)^\top$, while prediction still chooses among the four original classes. We restrict attention to codes whose columns are balanced and span the full centered class-score space:

$$M^\top \mathbf{1}_C = 0, \qquad \operatorname{rank}(M) = C - 1.$$

For a binary code, the first condition means that each column contains equally many $+1$ and -1 entries, and therefore requires even C. It ensures that every coded task represents a contrast between classes rather than the common score direction, which does not affect the $\arg\max$ prediction. Together with balance, the rank condition implies that the columns of M span the entire centered subspace $\mathbf{1}_C^\perp$. Consequently, all codes considered in our comparison can represent the same centered class-score functions; they differ in how squared-loss training weights those class distinctions, not in which class-score functions are expressible. The rank condition also requires $n \geq C - 1$. Define the centering matrix $H_C = I_C - \mathbf{1}_C \mathbf{1}_C^\top / C$, the orthogonal projector onto $\mathbf{1}_C^\perp$. The example code satisfies both conditions.

Shared low-rank bottleneck. We constrain all coded outputs to depend on the same r linear features ($1 \leq r \leq d$) via a projection matrix $V \in \mathbb{R}^{d \times r}$. A weight matrix $W \in \mathbb{R}^{r \times n}$ then maps these shared features to coded scores. This factorization restricts the coefficient matrix $\Theta = VW \in \mathbb{R}^{d \times n}$ to rank at most r. (For example, if $n = 3$ and $r = 1$, all three coded outputs depend on the same one-dimensional projection). We minimize the population squared loss:

$$\Theta_{M,r} \in \operatorname*{arg\,min}_{\Theta : \operatorname{rank}(\Theta) \leq r} \mathbb{E}\left[\left\| z_Y - \Theta^\top U \right\|_2^2 \right]. \tag{1}$$

For an input u, the fitted coded scores are $f_{M,r}(u) = \Theta_{M,r}^\top u \in \mathbb{R}^n$. This is a reduced-rank regression (RRR) problem, and we use these real-valued coded scores directly, without thresholding. Here, r is the bottleneck rank and n is the code length.

Decoding to the original label space. With a full centered span, we have $MM^\dagger = H_C$, where $M^\dagger \in \mathbb{R}^{n \times C}$ is the Moore–Penrose pseudoinverse. Given code scores $f \in \mathbb{R}^n$, we reconstruct the centered scores $g = (M^\dagger)^\top f \in \mathbb{R}^C$ and predict the class via:

$$D_M(f) = \operatorname{argmax}_c \left[(M^\dagger)^\top f \right]_c. \tag{2}$$

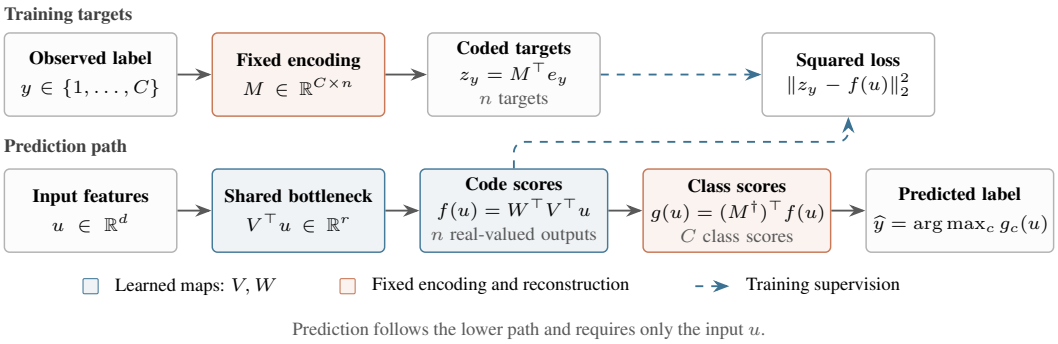

Figure 1: We encode the observed label with row y of M and fit all n coded targets through the same r learned features. At prediction time, we reconstruct C class scores from the real-valued code scores and choose one class. Blue boxes use learned coefficients; orange boxes use the fixed code. Here $f = f_{M,r}$ and g denotes the reconstructed class scores. The code acts on label targets and reconstruction, while V acts on input features.

Our final prediction is therefore $\widehat{y}(u) = D_M(f_{M,r}(u))$, breaking ties by choosing the smallest class index. Note that the reconstructed scores g need not be valid probabilities (see Figure 1 for the complete training and inference pipeline).

Consequently, all codes considered in our comparison can represent the same centered class-score functions; they differ in how squared-loss training weights those class distinctions, not in which class-score functions are expressible (see Appendix J for the formal rank-preserving equivalence).

A tractable data-generating process. We assume a *spherical affine model*, where features are uniformly distributed on a sphere and class probabilities vary linearly around a uniform prior. Let $A = [a_1, \ldots, a_C] \in \mathbb{R}^{d \times C}$ be the signal matrix. We define the posterior vector $\eta(u) \in \mathbb{R}^C$, with elements $\eta_c(u) = \Pr(Y = c \mid U = u)$, as:

$$\eta(u) = \frac{1}{C}\mathbf{1}_C + A^\top u, \quad U \sim \mathrm{Unif}(\sqrt{d}\,\mathbb{S}^{d-1}). \tag{3}$$

To ensure that $\eta(u)$ is a valid probability distribution, we require $A\mathbf{1}_C = 0$ and $\sqrt{d}\max_c \|a_c\|_2 \le \frac{1}{C}$. The labels are then sampled via $Y \mid U = u \sim \mathrm{Categorical}(\eta(u))$. The learner observes only (U, Y), remaining agnostic to A and η. Because spherical symmetry yields $\mathbb{E}[UU^\top] = I_d$ and a uniform marginal class distribution, the best constant classifier achieves an error of $1 - 1/C$. Under this model, the conditional coded mean is $\mathbb{E}[z_Y \mid U = u] = M^\top \eta(u)$. While reconstructing this mean yields $H_C\eta(u)$ and preserves the Bayes optimal decision, we are specifically interested in the error introduced when our rank constraint prevents full recovery.

In class coordinates, the squared loss weights this reconstruction by $K_M = MM^\top \in \mathbb{R}^{C \times C}$, which we refer to as the *coding metric* on centered class scores. Ultimately, the leading eigenspace of $AK_MA^\top$ dictates the retained directions.

Proposition 1. *Let $K_M = MM^\top$, and let $P_{M,r} \in \mathbb{R}^{d \times d}$ be the orthogonal projector onto the leading r-dimensional eigenspace of $AK_MA^\top$. Then a population risk minimizer and its reconstructed class-score coefficient matrix are*

$$\Theta_{M,r} = P_{M,r}AM, \qquad \widetilde{A}_{M,r} := \Theta_{M,r}M^\dagger = P_{M,r}A. \tag{4}$$

For $r < d$, the selected subspace is unique when the r-th and $(r+1)$-th eigenvalues of $AK_MA^\top$ are distinct. For $r = d$, the projector is I_d. Moreover, if $r \ge \mathrm{rank}(A)$, then $\widetilde{A}_{M,r} = A$.

Proof. See Appendix A. □

Thus, the selected feature directions are determined jointly by the class signal A and the coding metric K_M.

3 GEOMETRIC ANALYSIS

We quantify the information retained by a bottleneck via the expected maximum class score. Let $\mathcal{V} = \mathrm{conv}\{a_1, \ldots, a_C\}$ denote the *posterior coefficient polytope*. For an orthogonal projector $P \in \mathbb{R}^{d\times d}$, the projected vector Pu captures the retained input information. We define the reconstructed-score classifier and its associated risk as $h_P(u) = \mathrm{argmax}_c\, a_c^\top Pu$ and $R(P) = \Pr\{h_P(U) \neq Y\}$, respectively. Let $R^* = R(I_d)$ denote the Bayes risk based on the full input, while $R(h) = \Pr\{h(U) \neq Y\}$ denotes the risk of any generic classifier h. For a standard Gaussian vector $G \sim \mathcal{N}(0, I_d)$, we define the one-sided *Gaussian width* $w(\mathcal{V})$ and the spherical conversion factor κ_d by:

$$w(\mathcal{V}) = \mathbb{E}\big[\max_{v\in\mathcal{V}} G^\top v\big], \qquad \kappa_d = \sqrt{d}/\mathbb{E}\big[\|G\|_2\big] = \sqrt{d}\,\Gamma(d/2)/(\sqrt{2}\,\Gamma((d+1)/2)).$$

Because Gaussian width is unchanged by translating all class coefficient vectors by the same amount, $w(\mathcal{V})$ depends only on their relative differences. We call the difference $a_c - a_j$ between two class coefficient vectors a *class contrast*; it determines how the relative scores of classes c and j vary with the input. Ultimately, in Theorem 2, we show that the width lost under projection corresponds to the classification regret, up to the scaling factor κ_d.

Theorem 2 (Exact decision regret)**.** *Under the data-generating process in Equation (3), for any orthogonal projector P:*

$$R(P) = 1 - C^{-1} - \kappa_d w(P\mathcal{V}), \tag{5}$$

$$R(P) - R^* = \kappa_d\big[w(\mathcal{V}) - w(P\mathcal{V})\big], \tag{6}$$

$$R(P) = R^* \quad\Longleftrightarrow\quad (I-P)(a_c - a_j) = 0 \quad \textit{for all } c, j. \tag{7}$$

The classifier h_P achieves Bayes optimality among all decision rules measurable with respect to PU. Given the centering constraint $A\mathbf{1}_C = 0$, Equation (7) is equivalent to $PA = A$.

Proof. See Appendix A. □

Theorem 2 shows that the excess classification risk is exactly proportional to the lost Gaussian width. For rank one, this identity also yields an explicit characterization of the optimal projection in terms of the largest pairwise class-coefficient difference. Specifically, let $D_A = \max_{i,j} \|a_i - a_j\|_2$ denote the maximum pairwise distance, and let $R_r^*(A) = \min_{\mathrm{rank}(P)=r} R(P)$ be the minimum risk over all orthogonal projectors of rank r. We then obtain:

$$R_1^*(A) = 1 - C^{-1} - \kappa_d D_A/\sqrt{2\pi}. \tag{8}$$

When $D_A > 0$, the optimal projection direction aligns with a maximal pairwise difference $a_i - a_j$; if $D_A = 0$, any direction is trivially optimal; see Appendix H.1. Naturally, enlarging the retained subspace cannot increase the optimal classification risk. Theorem 2 also yields two useful consequences, summarized in Corollary 3: the optimal risk is monotone with the retained subspace, and the excess risk admits a simple bound in terms of the discarded class coefficients.

Corollary 3 (Excess-risk bound)**.** *For any orthogonal projector P,*

$$0 \le R(P) - R^* \le \kappa_d w((I-P)\mathcal{V}) \le \kappa_d\sqrt{2\log C}\,\max_c \|(I-P)a_c\|_2. \tag{9}$$

Proof. See Appendix A.3. □

Since h_P is Bayes optimal given PU, retaining more feature directions cannot increase the best achievable classification risk. Appendix A.3 further bounds the excess risk in terms of the class-coefficient components discarded by the projection.

While squared loss minimizes the discarded target energy $\mathcal{A}_r(M; A) = \sum_{j>r} \sigma_j(AM)^2$ (where σ_j denotes the j-th singular value), classification regret depends fundamentally on the lost Gaussian width. Consequently, even under isotropic coding ($K_M = \rho_K H_C$ for $\rho_K > 0$), the selected leading eigenspace of $AA^\top$ does not necessarily maximize the projected width. In Appendix H.1, we construct a sequence of classification problems for which one-hot or rescaled squared loss retains an arbitrarily small fraction of the largest classification-error reduction achievable by any rank-one projection relative to the best classifier that ignores the input. Beyond the spherical affine model, Appendix H.2 bounds the effects of posterior-approximation error and the discrepancy between $\mathbb{E}[U \mid PU]$ and PU. The exact expression of classification risk in terms of Gaussian width, however, relies on spherical features and on the affine class probabilities having the uniform baseline $\mathbf{1}_C/C$.

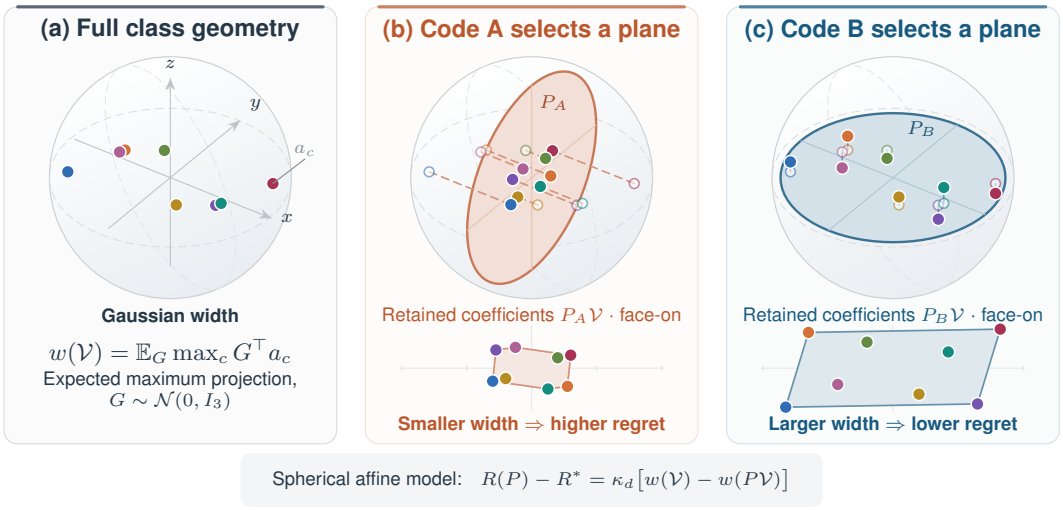

Figure 2: Both codes share the posterior coefficient polytope (shaded), but their coding metrics select different principal axes: vertical for Code A and horizontal for Code B. Dashed lines show projected coefficients. Code B retains a wider interval, hence greater Gaussian width and lower optimal risk. The full construction perturbs this base geometry to distinguish all classes while certifying the separation.

4 Matched Codes Can Select Different Feature Subspaces

We now ask whether conventional code summaries determine the projector selected by training. Conventional code summaries describe properties of M but do not capture how M interacts with the class-signal matrix A. Write $R(M, r) = R(P_{M,r})$ for the resulting risk. For code column $m_j = M_{:,j}$, the binary target has conditional mean $(Am_j)^\top U$ and Bayes error

$$q^*_{M,j} = \frac{1}{2} - \frac{1}{2}\mathbb{E}\big[|(Am_j)^\top U|\big] = \frac{1}{2} - \frac{1}{2}\kappa_d\sqrt{2/\pi}\,\|Am_j\|_2,$$

where $\|Am_j\|_2$ is the signal norm of binary task j; i.e., j-th *bit norm*. Thus, matching bit norms matches intrinsic binary-task difficulty, which differs from multiclass risk R^*. We construct codes with $C-1$ bits that match these norms and the stated summaries but select different subspaces. In this regard, we first illustrate the mechanism with an eight-class base model (Figure 2). For each target bottleneck rank r, we then use $m = 2^{\lceil \log_2 r \rceil}$ groups of eight classes. We also place a two-dimensional class-distinguishing signal of scale τ in each of r active groups and add centered coordinates of scale ϵ to make all class coefficients distinct.

Theorem 4. *For any bottleneck rank $r \geq 1$, let $m = 2^{\lceil \log_2 r \rceil}$, $C = 8m$, and $d = C + 2r$. There exist distinct balanced binary coding matrices with full centered span $M_A^{(m)}, M_B^{(m)} \in \{-1, +1\}^{C\times(C-1)}$ that have identical singular spectra, identical row- and column-Hamming-distance histograms, and no repeated columns up to sign.*

For any $0 < \tau \leq 1/(C\sqrt{6d})$ and $0 < \epsilon \leq \tau$, there exists a valid spherical affine posterior with C distinct class coefficients and $\mathrm{rank}(A) = C - 1$ such that the two codes have equal Bayes errors for each pair of corresponding binary tasks, with all such errors strictly below $1/2$. Despite these matched properties, the two codes select unique rank-r leading eigenspaces whose intersection is $\{0\}$. Moreover, letting $\alpha_r = \mathbb{E}[\max_{j\leq r} |Z_j|]$ for independent standard normal variables Z_j, their risks satisfy

$$R(M_A^{(m)}, r) - R(M_B^{(m)}, r) \geq \frac{3}{5}\kappa_d\tau\alpha_r \geq \frac{3}{35}\left(1 - \frac{1}{C} - R^*\right).$$

The separation is a bottleneck effect: at any feasible rank $k \geq C - 1$, both codes recover the global Bayes risk R^*.

Proof. See Appendix C. □

By Theorem 2, this separation persists even when each retained representation is evaluated with its Bayes-optimal decoder.

Appendix C gives the full construction, identifies the selected subspaces, derives the normalized risk gap, and provides an exact certificate for the separation over the full perturbation range. It also shows that the separation persists under sufficiently small perturbations that preserve the matching conditions, including perturbations that introduce cross-group couplings.

Appendix D gives a complementary separation under Gaussian class-conditional distributions with nonlinear posteriors. The codes retain matching spectra and distance histograms and yield different optimal-decoder risks, although this extension does not impose the matched bit Bayes errors of Theorem 4.

5 Finite-Sample Estimation

In finite samples, we separate two questions: how good is the representation learned with a given code when training and decoding are controlled, and how well does the complete code-selection procedure perform when all methods are given the same total number of labeled examples? To isolate these effects and weight all centered class contrasts equally, we first rescale the coded targets. For $n = C - 1$, a balanced code with a full centered span guarantees that $M^\top M$ is positive definite, which ensures its symmetric inverse square root is well-defined. Letting $g \in \mathbb{R}^C$ denote a centered class-score vector (such that $\mathbf{1}_C^\top g = 0$), we establish the resulting equivalence in Proposition 5.

Proposition 5. *For a balanced code M with full centered span and $n = C - 1$, defining the whitened code as $Q_M = M(M^\top M)^{-1/2}$ yields $Q_M Q_M^\top = H_C$. For any centered class-score vector g, this ensures:*

$$\|Q_M^\top (e_Y - g)\|_2^2 = \|H_C e_Y - g\|_2^2. \tag{10}$$

Consequently, for a fixed class of centered score functions, using Q_M yields a code-independent loss and preserves the exact same set of minimizing class-score functions across any dataset. Under a rank constraint, this equivalence also holds with a common Frobenius ridge penalty, provided the corresponding change of coordinates is applied.

Proof. See Appendix I.3. □

Note that this controlled equivalence assumes the fitting data, rank, regularization, and decoding strategy all remain fixed; varying any of these factors can independently alter performance.

Given N labeled training examples (u_i, y_i), let $X \in \mathbb{R}^{N \times d}$ be the input matrix centered across examples in each feature column and $Y \in \mathbb{R}^{N \times C}$ the class-indicator matrix centered across examples in each class column. The centered coded targets are $YM \in \mathbb{R}^{N \times n}$. We fit the linear bottleneck by solving

$$\min_{\Theta} \frac{1}{N} \|YM - X\Theta\|_F^2 + \lambda \|\Theta\|_F^2, \quad \text{s.t. } \operatorname{rank}(\Theta) \le r.$$

This reduced-rank ridge problem has a global solution obtained by truncated SVD; Appendix I.2 gives the closed-form reconstructed class-score coefficients, the fitted intercept (i.e., estimated constant offset), and the treatment of singular covariance matrices and eigenvalue ties. If $r \ge C - 1$ is feasible, the rank constraint no longer truncates the centered class signal, and all admissible full-span codes yield the same reconstructed coefficient.

6 Experiments

We test whether changing the output code alters the representation learned under a fixed fitting dataset, bottleneck rank, and decoder. We use a frozen ResNet-18 encoder (He et al., 2016) to map each image to 512 features, learning only the linear bottleneck and downstream decoder.

Design and primary comparison. For each task, we use five independent stratified training draws. We allocate nested label budgets of 1,000, 3,000, and 6,000 for Fashion-MNIST (Xiao et al., 2017), and 2,000 and 6,000 for CIFAR-100 (using its 100 fine classes (Krizhevsky, 2009)). Before feature

Table 1: Test error (%) as mean ± SE across five runs (lower is better). Budgets include fitting, pilot, and validation labels. All use frozen ResNet-18 (10k test images/task); the first three use a Gaussian decoder.

	Fashion-MNIST ($r = 4$)			CIFAR-100 ($r = 16$)	
Method	1,000	3,000	6,000	2,000	6,000
Representation comparison: same fitting labels (60% of budget)					
Gaussian-selected code	21.92 ± 0.25	20.96 ± 0.43	18.40 ± 0.10	64.72 ± 0.18	58.79 ± 0.87
One-hot RRR	31.95 ± 0.67	32.35 ± 0.65	32.02 ± 0.72	63.28 ± 0.27	59.96 ± 0.11
Alternatives: pilot labels available for fitting (80% of budget)					
Validation-selected code	21.55 ± 0.43	20.03 ± 0.35	18.39 ± 0.35	62.24 ± 0.23	56.76 ± 0.23
Projected LDA	32.31 ± 0.83	31.37 ± 0.48	31.72 ± 0.39	62.35 ± 0.26	59.33 ± 0.24
Rank-r softmax	21.35 ± 0.50	17.65 ± 0.26	16.43 ± 0.12	57.52 ± 0.07	49.50 ± 0.11

extraction, we fix the bottleneck ranks at 4 and 16, respectively, along with 32 candidate codes and three ridge penalties. Each candidate matrix contains $n = C - 1$ bits (9 and 99, respectively). Our primary comparison evaluates the selected code against one-hot RRR on Fashion-MNIST with 3,000 total labels, using the common decoder family described below; all other contrasts are considered secondary. We partition the available labels into 60% for predictor fitting, 20% for a pilot class-distribution model, and 20% for penalty validation. Comparators relying on the same fitting subset receive this exact 60% split, whereas full-budget baselines train on the combined fitting and pilot data (80%) and validate on the remaining 20%. When performing validation-code search, the validation set is also used to select the optimal code.

Code selection and representation evaluation. For each code M and penalty λ, let $\widehat{h}_{M,\lambda}$ be the fitted reconstructed-score classifier with intercept. From pilot data, we estimate class proportions $\widehat{\pi}_c^{\mathrm{pil}}$, means $\widehat{\mu}_c^{\mathrm{pil}}$, and pooled shrinkage covariance $\widehat{\Sigma}_{\mathrm{pool}}^{\mathrm{pil}}$. We score candidates by

$$\widehat{R}_G(M, \lambda) = \sum_{c=1}^{C} \widehat{\pi}_c^{\mathrm{pil}} \Pr_{U \sim \mathcal{N}(\widehat{\mu}_c^{\mathrm{pil}}, \widehat{\Sigma}_{\mathrm{pool}}^{\mathrm{pil}})} \left\{ \widehat{h}_{M,\lambda}(U) \neq c \right\}. \tag{11}$$

We approximate this risk using shared simulation draws across all candidate codes, minimize it for each penalty, and ultimately select the penalty based on validation error. Notably, this model-based scoring mechanism does not strictly require the underlying image features to be perfectly Gaussian. Once codes and penalties are fixed, we fit an identical projected linear discriminant analysis (LDA) (Hastie et al., 2009) decoder to both the coded and one-hot representations, ensuring they use the exact same fitting examples and covariance estimates. We do not re-run selection after swapping decoders. This strategy effectively controls for the decoding process while preserving estimation error, though it does not perfectly measure optimal-decoder risk. For the softmax baseline, we use the factored implementation described in Appendix K. Table 1 reports mean test error ± one standard error across the five draws; Appendix K gives the complete experimental protocol and diagnostics.

Effects of Code Choice and Procedure Comparisons The results reveal substantial but task-dependent code-induced representation effects. In our primary Fashion-MNIST comparison, the selected code reduces test error by 11.39 pp relative to one-hot RRR, yielding an approximate paired 95% confidence interval of $[10.19, 12.59]$; this favorable effect persists across all evaluated draws and budgets (Figure 3). On CIFAR-100, however, the effect is unfavorable at 2,000 labels and unresolved at 6,000 labels. This variation supports our central claim that output codes interact with task geometry rather than providing a universal improvement over one-hot training. Because the dataset and bottleneck rank differ across tasks, the experiments cannot isolate the precise source of this variation (secondary comparisons are reported without adjustment for multiple testing).

When controlling for equal total label budgets, a full-budget validation code search remains competitive with our Gaussian selection strategy, and direct rank-constrained softmax achieves lower error at every reported budget (Table 1). Interestingly, Gaussian scoring and validation search (when restricted to the same fitting labels) select the exact same final Fashion-MNIST models at both 3,000 and 6,000 labels. Ultimately, while coding yields clear representation gains, these results do not demonstrate an independent benefit for Gaussian selection or a strict advantage under equal computational budgets.

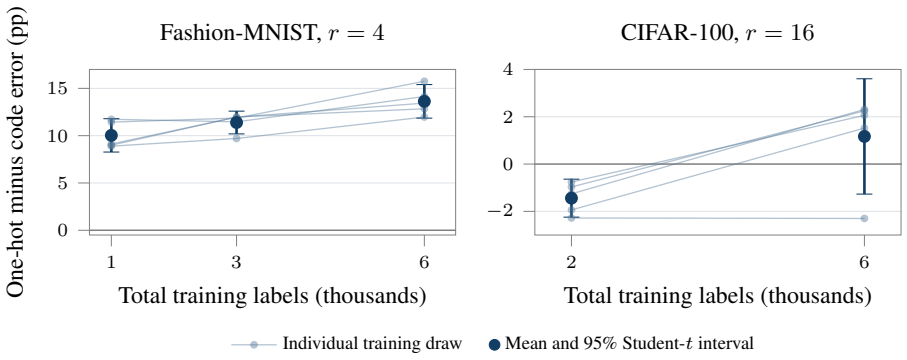

Figure 3: Coding improves on one-hot RRR fitted on the same labels on Fashion-MNIST, but increases CIFAR-100 error at 2,000 labels; the 6,000-label effect is unresolved. Positive values favor coding under the common Gaussian decoder. Light lines show five draws; dark points show means and approximate paired Student-t 95% intervals, conditional on the eligible training pool and fixed test set. Budgets are nested. Dataset and rank differ between panels.

We report cross-encoder confirmation in Appendix L, and scoring diagnostics and paired audits in Supplementary Section S9.

7 Discussion and Conclusion

We show that output coding can change which class distinctions are preserved under a shared low-rank bottleneck, even when the compared codes can represent the same centered class-score functions. Under the spherical affine model, the increase in classification error is exactly proportional to the Gaussian width lost by projection. Our matched-code construction sharpens this result: the two codes have identical singular spectra, identical distance histograms, and equal Bayes errors for corresponding binary tasks, yet they retain subspaces with no nonzero direction in common and achieve different multiclass risks. Thus, conventional code summaries do not determine which class information is preserved; the outcome also depends on how the coding metric interacts with the distribution of class information across feature directions. The Gaussian-width identity quantifies the classification cost of the information discarded by the selected subspace. Rescaling the coded targets to remove code-dependent weighting makes the squared-loss objective invariant to the particular full-span code, but this choice need not minimize classification error. These observations also motivate two distinct empirical comparisons: when the goal is to isolate the effect of code choice, compare representations under a common decoder; when comparing complete learning procedures, use the same total number of labeled examples.

Our theoretical construction applies to every bottleneck rank, with the number of classes increasing with the rank. Although the absolute risk difference can decrease as the posterior-validity constraints become more restrictive, the construction guarantees a gap of at least $3/35$ of the improvement achieved by the full-input Bayes classifier over the best classifier that ignores the input. The separation also persists under sufficiently small perturbations that preserve the matching conditions. The experiments show that sensitivity to code choice can also be substantial and task-dependent with real frozen features. On Fashion-MNIST, the selected code reduces test error by 11.39 percentage points relative to the matched one-hot low-rank baseline, whereas on CIFAR-100 it performs worse with 2,000 labeled examples and shows no clear difference with 6,000 labeled examples. These results support the central conclusion that code choice can materially affect which class information is retained under a shared bottleneck, without implying that output coding is uniformly better than one-hot training. Natural next steps include extending the theory beyond the spherical affine model, studying end-to-end representation learning rather than frozen features, and developing methods to select or design codes from properties of the data distribution.

AI-USE DISCLOSURE

We used OpenAI's ChatGPT for language editing, coding assistance, and identifying related works. The authors have fully reviewed and remain responsible for all content, mathematical statements, and experimental results.

ETHICS STATEMENT

We investigate output coding under shared representation bottlenecks using synthetic data and publicly available benchmarks. We do not collect new personal data or conduct interventions involving human participants. Our results concern predictive risk under the stated assumptions; they do not establish fairness or safety in downstream applications, where class-specific performance and potential disparities require separate assessment.

REPRODUCIBILITY STATEMENT

We provide proofs, certificate constructions, and experimental protocols in the Appendix (Appendices K and L, and Supplementary Sections S9–S11). The accompanying code archive includes implementations and setup instructions. We also include the experimental source code, protocols, and setup instructions with the submission. Follow these instructions to obtain the required datasets and pretrained weights, regenerate synthetic data and features, and rerun model fitting and evaluation. To meet the submission size limit, we omit saved data and experiment outputs. We will make these omitted materials publicly available through the project's GitHub repository upon acceptance.

REFERENCES

Erin L. Allwein, Robert E. Schapire, and Yoram Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1:113–141, 2000. URL https://www.jmlr.org/papers/v1/allwein00a.html.

Miguel Ángel Bautista, Oriol Pujol, Fernando de la Torre, and Sergio Escalera. Error-correcting factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(10):2388–2401, 2018. doi: 10.1109/TPAMI.2017.2763146. URL https://doi.org/10.1109/TPAMI.2017.2763146.

Xin Bing, Bingqing Li, and Marten Wegkamp. A new regression lens on multi-class classification. *Bernoulli*, 2026. URL https://arxiv.org/abs/2402.14260. To appear.

Derik T. Boonstra, Rakheon Kim, and Dean M. Young. Subspace ordering for maximum response preservation in sufficient dimension reduction. arXiv preprint arXiv:2510.27593, 2025. URL https://arxiv.org/abs/2510.27593.

Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995. doi: 10.1137/0916069. URL https://doi.org/10.1137/0916069.

R. Dennis Cook and Xiangrong Yin. Dimension reduction and visualization in discriminant analysis (with discussion). *Australian & New Zealand Journal of Statistics*, 43(2):147–199, 2001. doi: 10.1111/1467-842X.00164. URL https://doi.org/10.1111/1467-842X.00164.

Koby Crammer and Yoram Singer. On the learnability and design of output codes for multiclass problems. *Machine Learning*, 47(2–3):201–233, 2002. doi: 10.1023/A:1013637720281. URL https://doi.org/10.1023/A:1013637720281.

Thomas G. Dietterich and Ghulum Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286, 1995. doi: 10.1613/jair.105. URL https://doi.org/10.1613/jair.105.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936. doi: 10.1007/BF02288367. URL https://doi.org/10.1007/BF02288367.

Samarth Gupta and Saurabh Amin. Integer programming-based error-correcting output code design for robust classification. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 1724–1734. PMLR, 2021. URL https://proceedings.mlr.press/v161/gupta21c.html.

Trevor Hastie, Robert Tibshirani, and Andreas Buja. Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association*, 89(428):1255–1270, 1994. doi: 10.1080/01621459.1994.10476866. URL https://doi.org/10.1080/01621459.1994.10476866.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, second edition, 2009. doi: 10.1007/978-0-387-84858-7. URL https://doi.org/10.1007/978-0-387-84858-7.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830. URL https://doi.org/10.1080/01621459.1963.10500830.

Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248–264, 1975. doi: 10.1016/0047-259X(75)90042-1. URL https://doi.org/10.1016/0047-259X(75)90042-1.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/cifar.html.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. doi: 10.1016/S0047-259X(03)00096-4. URL https://doi.org/10.1016/S0047-259X(03)00096-4.

Michel Ledoux. Concentration of measure and logarithmic Sobolev inequalities. *Séminaire de probabilités*, 33:120–216, 1999. URL https://www.numdam.org/item/SPS_1999__33__120_0/.

Mingxia Liu, Daoqiang Zhang, Songcan Chen, and Hui Xue. Joint binary classifier learning for ECOC-based multi-class classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(11):2335–2341, 2016. doi: 10.1109/TPAMI.2015.2430325. URL https://doi.org/10.1109/TPAMI.2015.2430325.

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://www.cs.mcgill.ca/~colt2009/papers/012.pdf.

Paul Mineiro and Nikos Karampatziakis. Fast label embeddings via randomized linear algebra. In *Machine Learning and Knowledge Discovery in Databases*, volume 9284 of *Lecture Notes in Computer Science*, pp. 37–51. Springer, 2015. doi: 10.1007/978-3-319-23528-8_3. URL https://doi.org/10.1007/978-3-319-23528-8_3.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Martin Raič. A multivariate Berry–Esseen theorem with explicit constants. *Bernoulli*, 25(4A):2824–2853, 2019. doi: 10.3150/18-BEJ1072. URL https://doi.org/10.3150/18-BEJ1072.

Harish G. Ramaswamy and Shivani Agarwal. Convex calibration dimension for multiclass loss matrices. *Journal of Machine Learning Research*, 17(14):1–45, 2016. URL https://jmlr.org/papers/v17/14-316.html.

Yang Song, Qiyu Kang, and Wee Peng Tay. Error-correcting output codes with ensemble diversity for robust learning in neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9722–9729, 2021. doi: 10.1609/aaai.v35i11.17169. URL https://doi.org/10.1609/aaai.v35i11.17169.

Gunjan Verma and Ananthram Swami. Error correcting output codes improve probability estimation and adversarial robustness of deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, pp. 8643–8653, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/cd61a580392a70389e27b0bc2b439f49-Abstract.html.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747, 2017. URL https://arxiv.org/abs/1708.07747.

Anlan Yu, Shusen Jing, Ning Lyu, Wujie Wen, and Zhiyuan Yan. Error correction output codes for robust neural networks against weight-errors: A neural tangent kernel point of view. In *Advances in Neural Information Processing Systems*, volume 37, pp. 82493–82513, 2024. doi: 10.52202/079017-2623. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/964ce727dccdc18aa9bad54604699bbe-Abstract-Conference.html.

Yi Yu, Tengyao Wang, and Richard J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015. doi: 10.1093/biomet/asv008. URL https://doi.org/10.1093/biomet/asv008.

Hao Zhang, Joey Tianyi Zhou, Tianying Wang, Ivor W. Tsang, and Rick Siow Mong Goh. Deep N-ary error correcting output codes. In *Proceedings of the 13th EAI International Conference on Mobile Multimedia Communications*, pp. 409–427, 2020a. doi: 10.4108/eai.27-8-2020.2299197. URL https://doi.org/10.4108/eai.27-8-2020.2299197.

Xin Zhang, Qing Mai, and Hui Zou. The maximum separation subspace in sufficient dimension reduction with categorical response. *Journal of Machine Learning Research*, 21(29):1–36, 2020b. URL https://jmlr.org/papers/v21/17-788.html.

GUIDE TO THE APPENDIX

We present the proofs and empirical details supporting our two primary claims: how a shared bottleneck alters optimal classification risk, and why matched code summaries fundamentally fail to capture this change. To this end, we first derive the decision-risk identity before constructing and formally certifying our matched-code examples. We then explore how these theoretical conclusions generalize beyond the spherical affine model and comprehensively detail our experimental protocol. Finally, all additional guarantees, numerical illustrations, and extended dataset results—alongside their explicit assumptions and evaluation scopes—are deferred to the Supplementary Material.

Question	Where to find the argument or evidence
How does width determine regret?	Appendix A: conditional Bayes optimality, width conversion, and the equality condition.
What matches, and what changes?	Appendices B and C: code identities, all-direction separation, exact certificate, and local coupling.
What extends beyond the model?	Appendices D and H.2: Gaussian class models and distribution-dependent approximation defects.
Which guarantees apply?	Appendices H.1, J, and G: direct decision comparators, equal score capacity, and paired audits.
What varies across training draws?	Appendix K: the primary Fashion-MNIST comparison, CIFAR-100, all budgets, and all prescribed methods.
What varies across encoders?	Appendix L: the earlier ResNet/DINOv2 study and its conditional uncertainty.

The accompanying source package contains the experiment scripts, fixed protocols, execution records, and CSV data used to generate the figures. It does not include the original saved feature arrays, fitted models, or individual test predictions. We use numerical implementation checks to complement the proofs; our empirical comparisons do not assume the exact population model.

In the experimental tables, "same fit" means a comparator uses the same final fitting examples as the selected-code method. "Full budget" adds the pilot examples to that fitting set. Both comparisons use the same validation partition.

NOTATION USED THROUGHOUT THE PAPER

We use lowercase letters for realizations of random inputs and labels. The following symbols retain the same meaning throughout the main model; extensions define their additional quantities locally. Subscripts distinguish code-dependent quantities from general class-score coefficients.

Symbol	Meaning
C, d, n, N, r	Number of classes, input dimension, code length, training sample size, and bottleneck rank.
U, u, Y, y	Random input, input realization, random class label, and label realization.
M, m_j, z_y	Binary coding matrix, its jth column, and encoded label $M^\top e_y$.
H_C, K_M	Class-centering matrix and coding metric $MM^\top$.
V, W, Θ	Shared feature map, coded-output readout, and predictor coefficient $\Theta = VW$.
$A, a_c, \eta(u)$	True affine posterior coefficient matrix, its cth column, and class-probability vector.
$B, \widehat{B}_{M,r}$	A centered class-score coefficient matrix and its empirical RRR estimate.
$P_{M,r}, \mathcal{V}$	Code-selected feature projector and coefficient polytope $\mathrm{conv}\{a_c\}_{c=1}^{C}$.
$R(h), R(P), R(M, r)$	Classifier risk, projected-score risk, and code-selected population risk.
$R^*, q^*_{M,j}$	Full-input multiclass Bayes risk and Bayes error of coded binary task j.
$w(\mathcal{V}), \kappa_d$	Gaussian width and the factor converting Gaussian support means to spherical support means.
S_N, T	Empirical regularized input covariance normalizer and normalized feature–label cross-moment matrix.
$T_M, \widehat{T}_M$	Population coded moment AM and its moment estimator.
B_0	Fixed integer coefficient matrix in the eight-class construction.

A Complete proofs for decision preservation

We derive the population reduced-rank solution, then identify the class probabilities available from the retained features. We show that class-score reconstruction attains the conditional Bayes risk and use Gaussian width to characterize exactly when projection preserves the full-input Bayes risk.

A.1 Proof of Proposition 1

Using code balance and the affine posterior, we obtain $\mathbb{E}[z_Y \mid U] = M^\top A^\top U$. Conditioning on U and using $\mathbb{E}[UU^\top] = I_d$, we decompose the population loss as

$$\mathbb{E}\big[\|z_Y - \Theta^\top U\|_2^2\big] = \mathbb{E}\big[\|z_Y - \mathbb{E}[z_Y \mid U]\|_2^2\big] + \|AM - \Theta\|_F^2.$$

The first term does not depend on Θ. Applying the Eckart–Young theorem (Eckart & Young, 1936), we obtain a best rank-r approximation by retaining the leading left singular directions of AM, or equivalently the leading eigendirections of $AK_MA^\top$. Thus $\Theta_{M,r} = P_{M,r}AM$. Reconstructing the class coefficients gives

$$\widetilde{A}_{M,r} = P_{M,r}AMM^\dagger = P_{M,r}AH_C = P_{M,r}A,$$

where we use $MM^\dagger = H_C$ and $A\mathbf{1} = 0$. An eigengap at the truncation boundary makes the selected subspace unique. Finally, $\mathrm{range}(AM) = \mathrm{range}(A)$ because $A = AMM^\dagger$. Hence $r \geq \mathrm{rank}(A)$ retains the entire signal, giving $\widetilde{A}_{M,r} = A$.

A.2 Proof of Theorem 2

We use spherical symmetry to obtain $\mathbb{E}[U \mid PU] = PU$. The posterior conditional on the retained features is therefore $\frac{1}{C}\mathbf{1}_C + A^\top PU$. Choosing its largest coordinate maximizes conditional success, so $R(P) = 1 - \frac{1}{C} - \mathbb{E}\big[\max_c a_c^\top PU\big]$. We then decompose a standard Gaussian vector into an independent radius and a uniform direction to derive Equation (5).

To characterize equality in Equation (7), we decompose $G = PG + (I - P)G$. Conditional on PG, distinct discarded class coefficients give at least one competitor a nonconstant Gaussian score difference relative to a maximizer of the projected scores. The competitor can overtake the projected maximizer with positive probability, so the conditional Jensen gap is strictly positive. Equality therefore holds exactly when all discarded coefficients coincide.

Conditional information in a spherical projection. We first determine what the retained features reveal about the label. Let $P \in \mathbb{R}^{d\times d}$ be the orthogonal projector onto their subspace, so $P = P^\top = P^2$, and let $P^\perp = I_d - P$ project onto the discarded subspace. Reflection across $\mathrm{range}(P)$ maps U to $PU - P^\perp U$ while preserving both its distribution and its projection PU. The discarded component has conditional mean zero: $\mathbb{E}[P^\perp U \mid PU] = 0$ and $\mathbb{E}[U \mid PU] = PU$. The same identities hold at rank zero and full rank. Applying the tower property to the affine posterior, we obtain

$$\Pr(Y = c \mid PU) = C^{-1} + a_c^\top PU.$$

Thus, h_P attains the Bayes risk given the retained representation:

$$R(P) = 1 - C^{-1} - \mathbb{E}_U\big[\max_c a_c^\top PU\big]. \tag{12}$$

Coincident projected classes have equal conditional posterior probabilities. The identity therefore remains valid at ties, provided the tie rule depends only on the projected scores.

Conversion to width and strict equality conditions. We next express the expected largest class score on the feature sphere as Gaussian width. Write a standard Gaussian vector $G \sim \mathcal{N}(0, I_d)$ as $G = \rho_G S$, where S is uniform on $\mathbb{S}^{d-1}$ and $\rho_G = \|G\|_2$ is its radius; ρ_G is independent of S. We use the support function $h_{\mathcal{V}}(z) = \max_c a_c^\top z$ to denote the largest class score in direction $z \in \mathbb{R}^d$. Positive homogeneity then gives

$$\mathbb{E}\big[h_{\mathcal{V}}(PG)\big] = \mathbb{E}\big[\rho_G\big]\,\mathbb{E}\big[h_{\mathcal{V}}(PS)\big] = \frac{\mathbb{E}\big[\rho_G\big]}{\sqrt{d}}\mathbb{E}\big[h_{\mathcal{V}}(PU)\big].$$

Substituting this relation into Equation (12), we obtain Equation (5). Setting $P = I_d$ gives the full-input Bayes risk; subtracting it yields Equation (6).

To prove strictness, suppose the discarded class coefficients $q_c = P^\perp a_c \in \mathbb{R}^d$ are not all identical. Conditional on the retained Gaussian coordinates $PG = x$, write $m_c = a_c^\top x$ for class c's retained score and fix any maximizing class i. We can choose j with $q_j \neq q_i$. The discarded component $P^\perp G$ is independent of PG and has a nondegenerate Gaussian law on $\mathrm{range}(P^\perp)$. Thus the full-score advantage of class j over class i,

$$\omega_{ji} = (m_j - m_i) + (q_j - q_i)^\top P^\perp G$$

has positive variance, so $\mathbb{E}[\max(0, \omega_{ji}) \mid PG = x] > 0$. Pointwise,

$$\max_c (m_c + q_c^\top P^\perp G) - (m_i + q_i^\top P^\perp G) \geq \max(0, \omega_{ji}).$$

Taking conditional expectations removes the centered discarded score of the projected maximizer. We therefore obtain a strict increase in the expected maximum for every x, and hence $w(\mathcal{V}) > w(P\mathcal{V})$. Conversely, if every q_c equals a common vector q, the full maximum differs from the projected maximum only by $q^\top P^\perp G$, which has mean zero. The widths then agree. Finally, centering gives $Cq = \sum_c P^\perp a_c = 0$, so equality holds exactly when $P^\perp a_c = 0$ for every c. This completes our proof of Theorem 2.

A.3 Proof of Corollary 3

We compare two retained-feature projectors $P, Q \in \mathbb{R}^{d\times d}$. If $\mathrm{range}(P) \subseteq \mathrm{range}(Q)$, then $PQ = P$, so a QU-measurable classifier can compute PU. Conditional Bayes optimality therefore gives $R(Q) \leq R(P)$: access to more retained features cannot increase the best achievable risk. To quantify the loss under projection, we use the following inequality for every $g \in \mathbb{R}^d$:

$$h_{\mathcal{V}}(g) \leq h_{P\mathcal{V}}(g) + h_{(I-P)\mathcal{V}}(g).$$

Taking expectations gives the first inequality in Equation (9). We then set $s = \max_c \|(I - P)a_c\|_2$. Each $G^\top(I - P)a_c$ is centered Gaussian with variance at most s^2. For $t > 0$, we combine the Gaussian exponential-moment bound with Jensen's inequality to obtain

$$\mathbb{E}[\max_c G^\top (I - P)a_c] \leq \frac{\log C}{t} + \frac{ts^2}{2}.$$

Minimizing over t, we obtain $s\sqrt{2\log C}$, with the case $s = 0$ following by continuity. This argument does not require independent class scores.

A.4 A pointwise margin statement

To examine individual decisions beyond the averaged width identity, fix an input u with unique Bayes class $c^* = \arg\max_c \eta_c(u)$. For each competitor $j \neq c^*$, we define the posterior margin $\gamma_j(u) = \eta_{c^*}(u) - \eta_j(u) = (a_{c^*} - a_j)^\top u > 0$. Comparing the projected scores shows that the decision remains c^* whenever

$$\gamma_j(u) > (a_{c^*} - a_j)^\top (I - P)u \qquad \text{for every } j \neq c^*.$$

We obtain a stronger sufficient condition by replacing the right-hand side with its absolute value. Conversely, if the projected rule selects $j \neq c^*$, its conditional regret satisfies

$$0 < \eta_{c^*}(u) - \eta_j(u) \leq (a_{c^*} - a_j)^\top (I - P)u.$$

These pointwise comparisons do not use spherical symmetry. We need the assumptions of Theorem 2 to obtain the exact averaged risk identity and its strict global equality condition.

B Base family and exact code identities

We build the all-rank construction from eight classes, choosing two coding metrics that select different feature directions despite matching the stated code summaries and bit Bayes errors. After establishing the risk gap, we derive the binary-code identities that extend it to larger bottlenecks.

B.1 THE TWO BASE CODES

For $C = 8$ classes and $n = 7$ binary tasks, we use the following code matrices, where the signs denote ± 1:

$$M_A = \begin{pmatrix} + & + & + & + & + & + & + \\ + & + & + & + & - & - & - \\ + & + & - & - & + & - & - \\ + & - & - & - & - & + & - \\ - & - & + & + & + & - & - \\ - & - & + & - & - & + & + \\ - & + & - & - & + & + & + \\ - & - & - & + & - & - & + \end{pmatrix}, \qquad M_B = \begin{pmatrix} + & + & + & + & + & + & + \\ + & + & - & - & - & - & - \\ + & - & + & + & - & - & - \\ + & - & - & - & + & + & - \\ - & - & - & - & + & - & + \\ - & + & + & - & - & - & + \\ - & - & + & + & - & + & + \\ - & + & - & + & + & + & - \end{pmatrix}.$$

We hold the eight-class prediction problem fixed and change only the code used to train the rank-one predictor. The matrix $B_0 \in \mathbb{R}^{2\times 8}$ below specifies how class probabilities vary along two informative input coordinates. We scale this signal by τ and add eight centered coordinates with scale ϵ, obtaining the posterior coefficient matrix $A_{\tau,\epsilon} \in \mathbb{R}^{10\times 8}$ in Equation (3). These additional coordinates distinguish classes that coincide in the base problem while preserving matched bit difficulty. With $n = 7$ bits, each learner must predict a seven-dimensional target from a ten-dimensional input through a shared one-dimensional representation.

Proposition 6 (Matched separation with certified robustness)**.** *For the binary matrices M_A, M_B in Appendix B.1, define*

$$B_0 = \begin{pmatrix} 1 & 0 & -2 & 1 & 0 & -1 & 2 & -1 \\ 1 & 0 & 1 & 1 & 0 & -1 & -1 & -1 \end{pmatrix}, \qquad A_{\tau,\epsilon} = \begin{pmatrix} \tau B_0 \\ \epsilon H_8 \end{pmatrix}.$$

Let $d = 10$, $0 < \tau \le 1/64$, and $0 \le \epsilon \le \tau$. The resulting centered affine posterior is valid on the sphere of radius $\sqrt{10}$. The two codes have identical spectra, distance histograms, and corresponding bit Bayes errors, and neither has repeated columns up to sign. At rank one,

$$R(M_A, 1) - R(M_B, 1) \ge \tfrac{1}{2}\kappa_{10}\tau > 0. \tag{13}$$

For $\epsilon > 0$, all eight posterior coefficient vectors are distinct, all seven bits have nonzero signal, and $\operatorname{rank}(A_{\tau,\epsilon}) = 7$. Both codes attain Bayes risk at rank seven. At $\epsilon = 0$, even the nonzero singular spectra of $A_{\tau,0}M_A$ and $A_{\tau,0}M_B$ agree, yet their exact rank-one risk gap is $\kappa_{10}\tau\sqrt{2/\pi}$.

We obtain this separation while controlling code geometry and marginal bit difficulty. At $\tau = 1/64$, the certified risk gap exceeds 0.80 percentage points throughout the interval. We give a hand proof for $\epsilon/\tau \le 10^{-4}$ in Appendix B.2, then certify every point of $[0, 1]$ in Appendix B.3 using exact rational spectral checks and bounds between grid centers. The nondegenerate separation therefore extends beyond infinitesimal perturbations. We establish robustness of these examples without asserting their prevalence.

B.2 PROOF OF PROPOSITION 6: ANALYTIC CONSTRUCTION

We verify that M_A, M_B are balanced, span the centered class space, share the code statistics in Proposition 6, and contain no equal or opposite column pairs. These properties depend only on the integer entries; our executable certificate checks the characteristic polynomials, ranks, and distance histograms exactly. We now prove the risk separation by hand for $\epsilon/\tau \le 10^{-4}$; Appendix B.3 extends the bound to $\epsilon/\tau \le 1$.

Validity and exact bit matching. For either code, write $m_j \in \{-1, +1\}^8$ for its jth column, which assigns the binary targets to the eight classes for task j. We use $m_{A,j}$ and $m_{B,j}$ to compare corresponding columns. The vector $B_0 m_j \in \mathbb{R}^2$ describes the variation in the conditional mean of bit j along the two base coordinates. Direct multiplication gives $B_0\mathbf{1} = 0$ and, for both codes,

$$\big(\|B_0 m_1\|^2, \ldots, \|B_0 m_7\|^2\big) = (36, 8, 0, 0, 8, 36, 20) =: \nu.$$

Here ν_j is the squared signal norm of the jth base binary task. Since $H_8 m_j = m_j$ and every binary column has squared norm eight,

$$\|A_{\tau,\epsilon} m_{A,j}\|^2 = \|A_{\tau,\epsilon} m_{B,j}\|^2 = \tau^2 \nu_j + 8\epsilon^2.$$

For a balanced column m, the conditional bit mean is $(Am)^\top U$, so its Bayes error is

$$q^*_{M,j} = \frac{1}{2} - \frac{1}{2}\mathbb{E}\big[|(Am_j)^\top U|\big] = \frac{1}{2} - \frac{1}{2}\|Am_j\|\mathbb{E}\big[|U_1|\big].$$

We use spherical symmetry in the final equality. Equal signal norms therefore give exactly matching bit Bayes errors. For $\epsilon > 0$, every signal norm is positive. The lower block ϵH_8 also separates all coefficient columns and gives rank seven on the centered class space. To check posterior validity, we bound the column norms by

$$\max_c \|a_c\|^2 \le 5\tau^2 + \tfrac{7}{8}\epsilon^2 < 6\tau^2, \qquad \sqrt{10}\max_c \|a_c\| < 8\tau \le \tfrac{1}{8}.$$

With centering, this bound establishes posterior validity over the entire support.

A strict gap at the base point. At $\epsilon = 0$, the class probabilities depend only on the first two input coordinates. We identify the coordinate retained by training from the larger eigenvalue of $A_{\tau,0}MM^\top A_{\tau,0}^\top$. Factoring out τ^2, we obtain the following nonzero blocks of the two code-weighted feature covariances:

$$B_0 M_A M_A^\top B_0^\top = \begin{pmatrix} 48 & 0 \\ 0 & 60 \end{pmatrix}, \qquad B_0 M_B M_B^\top B_0^\top = \begin{pmatrix} 60 & 0 \\ 0 & 48 \end{pmatrix}.$$

Code A therefore selects the second feature direction, whereas code B selects the first. Both have eigengap $12\tau^2$ and nonzero coded-target singular values $\tau\sqrt{60}, \tau\sqrt{48}$. The second row of τB_0 has range 2τ, while the first has range 4τ. Applying the rank-one segment formula, we obtain

$$R(M_A, 1) - R(M_B, 1) = \frac{\kappa_{10}}{\sqrt{2\pi}}(4\tau - 2\tau) = \kappa_{10}\tau\sqrt{2/\pi}.$$

At the base point, some class posteriors coincide and two bits have zero signal. We next remove both degeneracies by perturbing the posterior while preserving every bit-matching equality.

A quantitative nondegenerate interval. Let $0 < \epsilon \le \tau/10^4$ and fix either code M. We write $T_\epsilon = A_{\tau,\epsilon}M \in \mathbb{R}^{10\times 7}$ for the coded conditional-mean coefficient, $S_\epsilon = T_\epsilon T_\epsilon^\top \in \mathbb{R}^{10\times 10}$ for its feature covariance, and P_ϵ for the leading rank-one projector selected by training. Subscript zero denotes the unperturbed quantity. Using $\|M\|_{\mathrm{op}} \le \|M\|_F = \sqrt{56} < 8$ and $\|T_0\|_{\mathrm{op}} = \tau\sqrt{60} < 8\tau$, we bound the perturbations as

$$\|T_\epsilon - T_0\|_{\mathrm{op}} < 8\epsilon, \qquad \|S_\epsilon - S_0\|_{\mathrm{op}} < 128\tau\epsilon + 64\epsilon^2 < \tfrac{13}{1000}\tau^2 =: \rho.$$

Since $\rho < 12\tau^2/2$, the leading eigenvalue remains separated and unique. Applying Yu et al. (2015, Corollary 3) to the rank-one projectors, we obtain

$$\|P_\epsilon - P_0\|_{\mathrm{op}} \le \frac{2\rho}{12\tau^2} < \frac{13}{6000}.$$

To bound the projected scores, let $a_{\epsilon,c}$ denote column c of $A_{\tau,\epsilon}$, with τ fixed. Each unperturbed column has norm below 3τ, and its perturbation has norm at most ϵ. We therefore obtain

$$\max_c \|P_\epsilon a_{\epsilon,c} - P_0 a_{0,c}\| < \left(\frac{39}{6000} + \frac{1}{10^4}\right)\tau = \frac{33}{5000}\tau.$$

We now apply the expected-maximum Lipschitz bound in Appendix H.1. With $\sqrt{2\log 16} < 3$, each code's projected-support score changes by less than $(99/5000)\kappa_{10}\tau$. The two-code risk difference can therefore decrease by less than $(198/5000)\kappa_{10}\tau$. Since $\sqrt{2/\pi} > 3/4$, we retain

$$R(M_A, 1) - R(M_B, 1) > \left(\frac{3}{4} - \frac{198}{5000}\right)\kappa_{10}\tau > \frac{1}{2}\kappa_{10}\tau.$$

This proves Equation (13). Both codes reconstruct the full coefficient matrix at rank seven: the matrix has rank seven when $\epsilon > 0$ and rank two at $\epsilon = 0$. We claim matching coded-target spectra only at $\epsilon = 0$; the code spectra and bit difficulties match throughout the parameter interval.

B.3 Proof of Proposition 6: interval certificate

We extend the range-gap certificate to every $s = \epsilon/\tau \in [0, 1]$, as required by Proposition 6. We work with the normalized coefficient matrix $\mathcal{A}(s) = [B_0^\top, sH_8^\top]^\top$ and covariance $S_M(s) = \mathcal{A}(s)MM^\top\mathcal{A}(s)^\top$. For $s \in [0, 1]$, we have $\|\mathcal{A}(s)\|_{\mathrm{op}} < 4$, and each column has norm below $5/2$. We verify $M^\top M \prec 16I$ for both codes by exact rational LDL factorization. Using $\|\mathcal{A}(s) - \mathcal{A}(t)\|_{\mathrm{op}} \le |s - t|$, we expand the covariance product and obtain

$$\|S_M(s) - S_M(t)\|_{\mathrm{op}} \le 128|s - t| + 16|s - t|^2.$$

We cover $[0, 1]$ with intervals centered at $s_j = j/400$, $j = 0, \ldots, 400$, of radius $h = 1/800$, and set $\rho = 128h + 16h^2 = 6401/40000$. For each center and code, we store a rational vector v, a threshold b, and its Rayleigh quotient $\mu = v^\top S_M(s_j)v/(v^\top v)$. We then verify, in exact arithmetic,

$$|\|v\| - 1| \le 10^{-8}, \qquad \|S_M(s_j)v - \mu v\| \le 10^{-5}, \qquad \mu > b, \qquad \mathrm{inertia}(S_M(s_j) - bI) = (1, 9).$$

The inertia check establishes that exactly one eigenvalue lies above b. With $g_M = \mu - b > 2\rho$, we expand v in the eigenbasis and bound its distance from a signed leading unit eigenvector by

$$e_M = \frac{2 \cdot 10^{-5}}{(1 - 10^{-8})g_M} + 10^{-8}.$$

The factor two converts the orthogonal residual bound into unit-vector distance; the final term accounts for normalization. The score range computed from v therefore differs from the exact range by at most $5e_M$. Within each interval, the leading unit eigenvector moves by at most $3\rho/g_M$ after sign alignment (Yu et al., 2015, Corollary 3), and each coefficient column changes by at most h. Together, these bounds limit the range change to $15\rho/g_M + 2h$.

Write r_M for the range of $v^\top\mathcal{A}(s_j)$ for code M. Accounting for both sources of error, we verify at every center that

$$r_B - r_A - 5(e_A + e_B) - 15\rho(1/g_A + 1/g_B) - 4h \ge 13/10.$$

The smallest verified lower bound exceeds 1.3887. Applying the rank-one range formula, we conclude that throughout the covered interval,

$$R(M_A, 1) - R(M_B, 1) \ge \frac{13\kappa_{10}\tau}{10\sqrt{2\pi}} > \frac{1}{2}\kappa_{10}\tau.$$

Our posterior-validity and bit-matching arguments apply over this same interval. The source package contains a generator and a standalone checker for all rational norm, residual, inertia, and range inequalities. We use floating-point eigenvectors to propose witnesses and exact arithmetic to establish the bounds. We detail the certificate construction in Supplementary Section S1. In Appendix C, we prove the stronger comparison of individual endpoints needed for the all-rank theorem.

B.4 Exact identities for the lifted binary codes

We now extend each base code to m groups of eight classes. The group index is part of the label and supplies no additional input. By combining the base code with orthogonal sign patterns across groups, we preserve both the binary entries and the matched code statistics.

Let $m \ge r$ be a power of two. We take a Sylvester Hadamard matrix $[\mathbf{1}_m, H_m^\circ] \in \{-1, +1\}^{m\times m}$, whose columns are orthogonal sign vectors and whose first column is constant. The remaining $m - 1$ columns form $H_m^\circ \in \{-1, +1\}^{m\times(m-1)}$, so $(H_m^\circ)^\top H_m^\circ = mI_{m-1}$ and $H_m^\circ(H_m^\circ)^\top = mH_m$. Here $H_m = I_m - m^{-1}\mathbf{1}_m\mathbf{1}_m^\top$ centers vectors over the group index. When $m = 1$, we take H_m° to be empty. For either base code $M \in \{-1, +1\}^{8\times 7}$, we define the lifted code using the Kronecker product $\otimes$:

$$M^{(m)} = [\, M \otimes \mathbf{1}_m,\ \mathbf{1}_8 \otimes H_m^\circ,\ M \otimes H_m^\circ \,].$$

The lifted matrix has $C = 8m$ rows and $n = C - 1$ binary columns. We index its rows by $(c, t) \in \{1, \ldots, 8\} \times \{1, \ldots, m\}$, where c denotes the base class and t its group. The three blocks contain 7, $m - 1$, and $7(m - 1)$ bits, respectively, encoding the base class, the group, and their joint

patterns. All columns are balanced. Writing $K_M = MM^\top$ and $J_8 = \mathbf{1}_8\mathbf{1}_8^\top$, we obtain by direct multiplication

$$M^{(m)}M^{(m)\top} = mK_M \otimes I_m + J_8 \otimes mH_m.$$

To read off the spectrum, we separate contrasts within groups from contrasts between group means; $\mathbf{1}_k^\perp$ denotes the vectors in $\mathbb{R}^k$ whose entries sum to zero. On $\mathbf{1}_8^\perp \otimes \mathbb{R}^m$, the eigenvalues are m times the seven positive eigenvalues of K_M, each repeated m times. On $\mathrm{span}(\mathbf{1}_8) \otimes \mathbf{1}_m^\perp$, the eigenvalue is $8m$. Only the overall constant vector lies in the nullspace. We therefore obtain full centered span and matching singular spectra for the lifted codes.

We verify column-distance matching through the block-diagonal column Gram matrices:

$$\mathrm{diag}\{mM^\top M,\ 8mI_{m-1},\ (M^\top M) \otimes mI_{m-1}\}.$$

These matrices transfer column-distance matching from the base codes to the lifted codes. No two columns agree up to sign: within a nonorthogonal block, the claim follows from the base code; all remaining distinct pairs are orthogonal. Finally, consider rows (c, t) and (c', t'). When $t = t'$, their Hamming distance is m times the base row distance; when $t \neq t'$, it is $4m$. We thus obtain matching row-distance histograms as well.

C Proof of Theorem 4

We extend the eight-class construction to make the selected direction differ in every active group of the retained representation. We verify the matching conditions, identify the selected subspaces, and bound their risk difference relative to the available Bayes improvement. We then certify the endpoint inequality underlying this bound and show that the separation persists under small perturbations that preserve matching.

C.1 Codes, posteriors, and matched quantities

We construct a common input–label distribution and two codes for it. Both learners receive the same d-dimensional input U and class label Y; only the encoding of Y changes during training. Let m be the smallest power of two at least the bottleneck rank r, and let $C = 8m$ be the number of classes. We index classes by (c, t), with base class $c \in \{1, \dots, 8\}$ and group $t \in \{1, \dots, m\}$. We use the eight-class codes M_A, M_B and integer signal matrix B_0 from Proposition 6. Let $[\mathbf{1}_m, H_m^\circ]$ be the Sylvester Hadamard matrix of Appendix B.4, where $H_m^\circ \in \mathbb{R}^{m\times(m-1)}$ contains its nonconstant columns. Let $E_r \in \mathbb{R}^{m\times r}$ contain the first r coordinate vectors of $\mathbb{R}^m$; it selects the r active groups. For either base code, we define the lifted code and common posterior coefficient matrix as

$$M^{(m)} = [M \otimes \mathbf{1}_m,\ \mathbf{1}_8 \otimes H_m^\circ,\ M \otimes H_m^\circ], \qquad A = \begin{pmatrix} \tau B_0 \otimes E_r^\top \\ \epsilon H_C \end{pmatrix}.$$

We use τ to scale the base signal and ϵ to scale its centered perturbation. The base signals occupy the first $2r$ input coordinates, and the perturbation occupies the last C. The resulting matrix $A \in \mathbb{R}^{d\times C}$, with $d = C + 2r$, defines the common class probabilities $\eta(u) = C^{-1}\mathbf{1} + A^\top u$ in Equation (3). Every row of A is centered. With $J_8 = \mathbf{1}_8\mathbf{1}_8^\top$ denoting the 8×8 all-ones matrix, the code Gram identities in Appendix B.4 apply unchanged:

$$M^{(m)}M^{(m)\top} = mMM^\top \otimes I_m + J_8 \otimes mH_m.$$

The lifted codes therefore have full centered span, matching singular spectra and row/column distance histograms, and no repeated columns up to sign. To check posterior validity, we bound the largest squared column norm by $5\tau^2 + \epsilon^2(1 - 1/C) < 6\tau^2$. The stated bound on τ then ensures a strictly positive posterior throughout the sphere. For positive ϵ, we also have $\mathrm{rank}(A) = C - 1$, and all class columns are distinct.

We next verify bit-difficulty matching. Write $\nu = (36, 8, 0, 0, 8, 36, 20)$ for the corresponding squared bit norms of B_0M. For a lifted code column $m_j \in \mathbb{R}^C$, the vector Am_j gives the coefficient of the target's conditional mean. Its squared norm measures the bit's signal strength under the isotropic input distribution. The first code block has squared signal norms $r\tau^2\nu_j + C\epsilon^2$; the middle block has $C\epsilon^2$; each column indexed by (j, k) in the final block has $r\tau^2\nu_j + C\epsilon^2$, where $j \in \{1, \dots, 7\}$

selects a base bit and $k \in \{1, \dots, m-1\}$ a column of H_m°. The last identity uses the fact that each of the first r entries of a Hadamard column has squared magnitude one. These expressions agree for corresponding columns of the two codes. Writing $M_{:,j}^{(m)}$ for column j of the full lifted code, we use spherical symmetry to express its bit Bayes error as

$$q^*_{M^{(m)},j} = \frac{1}{2} - \frac{1}{2}\kappa_d\sqrt{2/\pi}\,\|AM^{(m)}_{:,j}\|,$$

so the corresponding bit Bayes errors match, and every bit is informative for positive ϵ.

C.2 Exactly which directions are selected

We identify the direction selected within each group and then combine these directions into the rank-r representation. We write $s = \epsilon/\tau \in [0,1]$ for the perturbation size relative to the base signal, and let $\mathcal{A}(s) = \begin{pmatrix} B_0 \\ sH_8 \end{pmatrix} \in \mathbb{R}^{10\times 8}$ be the normalized eight-class coefficient matrix. Within each active group's class-contrast component, we recover the base coded covariance: up to zero directions and an orthogonal change of coordinates, it is

$$m\tau^2\mathcal{A}(s)MM^\top\mathcal{A}(s)^\top.$$

The normalized base covariance $\mathcal{A}(s)MM^\top\mathcal{A}(s)^\top$ has leading eigenvalue at least 60. The exact interval certificate establishes a unique leading line throughout $[0,1]$, so each of the r active groups contributes one such line. In an inactive group, eigenvalues are below $16m\epsilon^2$; the class-group-mean block has eigenvalues at most $8m\epsilon^2$. Both bounds lie below $60m\tau^2$. The top-r projector therefore retains exactly one leading line from each active group. Although these lines share the same eigenvalue, their combined r-dimensional eigenspace lies strictly above the truncation boundary, so the selected projector is unique.

Let $v_M(s) \in \mathbb{R}^{10}$ be a leading unit eigenvector of $\mathcal{A}(s)MM^\top\mathcal{A}(s)^\top$. We project the eight coefficient columns onto this line to obtain an interval containing zero. We denote the magnitudes of its right and left endpoints by $q_M^+(s)$ and $q_M^-(s)$:

$$q_M^+(s) = \max_c v_M(s)^\top\mathcal{A}(s)_{:,c}, \qquad q_M^-(s) = -\min_c v_M(s)^\top\mathcal{A}(s)_{:,c}.$$

Changing the eigenvector's sign exchanges the endpoints but preserves their minimum and maximum. We certify the following endpoint inequality:

$$\min\{q_B^+(s), q_B^-(s)\} - \max\{q_A^+(s), q_A^-(s)\} \geq 3/5 \qquad (0 \leq s \leq 1). \tag{14}$$

This inequality ensures that the two selected base lines are distinct. Because the active-group embeddings are orthogonal, the two selected rank-r spaces have zero intersection: the selected line changes in each of the r active groups.

C.3 Risk gap and normalization by available Bayes improvement

We compute the width using independent standard normal variables $Z_1, Z_2, \dots$, with the first r coordinates corresponding to the active groups. Write P_A and P_B for the selected rank-r projectors under the lifted codes, and recall $\alpha_k = \mathbb{E}\big[\max_{j\leq k}|Z_j|\big]$ for any positive integer k with independent standard normals. We suppress the argument s of the endpoint functions. In selected coordinates, the largest class score equals

$$\tau\max_{t\leq r}\{q_M^+ Z_t\mathbf{1}\{Z_t \geq 0\} + q_M^-|Z_t|\mathbf{1}\{Z_t < 0\}\}.$$

Here $\mathbf{1}\{\cdot\}$ denotes an indicator, equal to one when its condition holds and zero otherwise. Inactive classes contribute zero and cannot increase this nonnegative maximum. Let $b = \min(q_B^+, q_B^-)$ and $a = \max(q_A^+, q_A^-)$. We couple the two representations through the same independent standard-normal Z_t to obtain

$$w(P_B\mathcal{V}) - w(P_A\mathcal{V}) \geq \tau(b-a)\mathbb{E}\big[\max_{t\leq r}|Z_t|\big] \geq (3/5)\tau\alpha_r.$$

Applying Theorem 2 converts this width difference into the absolute risk gap. Its conditional Bayes-optimality statement also shows that the gap persists under the best decoder of each retained representation.

To bound the full width, we treat the signal and perturbation rows separately. Let G_{1t} and G_{2t} be independent standard Gaussian coordinates corresponding to the two base signal coordinates of active group t. The base signal has coefficient set $\{0, \pm(1,1), \pm(-2,1)\}$. Hence its support over r groups is at most $2\tau \max_t |G_{1t}| + \tau \max_t |G_{2t}|$, with expectation $3\tau\alpha_r$. For the perturbation, let $G_1, \dots, G_C$ be independent standard normals and $\bar{G} = C^{-1}\sum_c G_c$ their mean. The perturbation contributes width $\epsilon\mathbb{E}\big[\max_{c\le C}(G_c - \bar{G})\big] = \epsilon\mathbb{E}\big[\max_c G_c\big] \le \epsilon\alpha_C$.

We compare α_C and α_r using Gaussian concentration. Since $C < 16r$, we can bound the maximum using at most $16r$ independent coordinates arranged in sixteen blocks of size r. Write Z_{blk} for one such block maximum. Each block maximum of absolute coordinates is 1-Lipschitz, has mean α_r, and satisfies the Gaussian centered moment-generating bound (Ledoux, 1999, Section 2.3) $\mathbb{E}\big[\exp(t(Z_{\mathrm{blk}} - \alpha_r))\big] \le \exp(t^2/2)$. We apply the log-sum-exp bound and optimize over $t > 0$ to obtain

$$\alpha_C \le \alpha_r + \sqrt{2\log 16} < 4\alpha_r,$$

where we used $\alpha_r \ge \sqrt{2/\pi}$ and $\pi \log 16 < 9$. Combining this bound with subadditivity of support functions, we obtain

$$w(\mathcal{V}) \le 3\tau\alpha_r + \epsilon\alpha_C \le 7\tau\alpha_r.$$

We divide by the positive Bayes improvement $1 - C^{-1} - R^* = \kappa_d w(\mathcal{V})$ to obtain the relative constant $3/35$. The absolute gap may decrease as C grows; the guarantee preserves a fixed fraction of the available Bayes improvement.

C.4 Exact certificate of endpoint domination

We certify the endpoint inequality by controlling two errors: the discrepancy between a numerical eigenvector and the true leading direction at each sampled parameter, and the movement of that direction between samples. Normalizing the eight-class covariance makes these checks depend only on $s = \epsilon/\tau$. We use rational centers $s_j = j/400$ and radius $h = 1/800$, whose intervals cover $[0,1]$. We first bound the covariance variation within each interval. Write $D = \begin{pmatrix} 0_{2\times 8} \\ H_8 \end{pmatrix}$, so $\mathcal{A}(s) = \mathcal{A}(t) + (s-t)D$ and $\|D\|_{\mathrm{op}} = 1$. The identity $B_0 B_0^\top = \mathrm{diag}(12, 6)$ gives $\|\mathcal{A}(t)\|_{\mathrm{op}} \le \sqrt{12 + t^2} < 4$ for $t \in [0,1]$. For each code, we verify $M^\top M \prec 16 I_7$ by exact rational LDL factorization, which gives $\|K_M\|_{\mathrm{op}} < 16$. We expand $S(s) = \mathcal{A}(s) K_M \mathcal{A}(s)^\top$ around t and apply the operator-norm triangle inequality to obtain

$$\|S(s) - S(t)\|_{\mathrm{op}} \le 2\|\mathcal{A}(t)\|_{\mathrm{op}}\|K_M\|_{\mathrm{op}}|s-t| + \|K_M\|_{\mathrm{op}}|s-t|^2 \le 128|s-t| + 16|s-t|^2.$$

Thus $\rho = 128h + 16h^2$ bounds covariance variation within one interval. At each center and for each code, we set $S = \mathcal{A}(s_j) M M^\top \mathcal{A}(s_j)^\top \in \mathbb{R}^{10\times 10}$. We store a rational vector $v \in \mathbb{R}^{10}$ approximating the leading unit eigenvector and a rational threshold b separating the largest eigenvalue from the others. Using exact Fraction arithmetic, we verify

$$|\|v\| - 1| \le 10^{-8}, \quad \|Sv - \mu v\| \le 10^{-5}, \quad \mu = \frac{v^\top S v}{v^\top v} > b, \quad \mathrm{inertia}(S - bI) = (1, 9).$$

Here μ is the Rayleigh quotient of v, and the inertia pair $(1,9)$ means that $S - bI$ has one positive and nine negative eigenvalues. Let $g = \mu - b > 2\rho$ be the certified spectral separation and $e = 2 \cdot 10^{-5}/((1 - 10^{-8})g) + 10^{-8}$. We bound the distance from the rational vector to a signed leading unit eigenvector by e. Since each coefficient column has norm below $5/2$, each endpoint has error at most $5e/2$ at the center. Between the center and any point in its interval, the unit leading eigenvector changes by at most $3\rho/g$ after sign alignment (Yu et al., 2015, Corollary 3), and each column changes by at most h. Using subscripts M to identify the code in the bounds e and g, we obtain the following total error bound for either endpoint:

$$u_M = (5/2)e_M + (15/2)\rho/g_M + h.$$

If $q_M^{\pm}$ are instead evaluated using the rational center vector, the exact interval check is

$$\min(q_B^+, q_B^-) - \max(q_A^+, q_A^-) - u_A - u_B \ge 3/5.$$

All 401 checks pass, and the smallest certified lower bound exceeds 0.6325. We store every rational vector and threshold and verify them with a separate standard-library checker. It reconstructs the matrices and repeats the exact norm, residual, inertia, and endpoint checks without a floating eigensolver. We use numerical eigenvectors only to propose witnesses; the certificate relies entirely on exact checks.

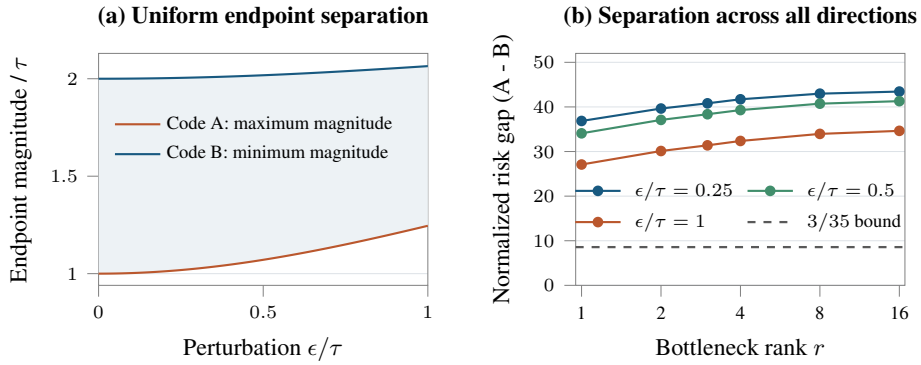

Figure 4: (a) Retained-axis endpoints normalized by τ. Code B's minimum magnitude exceeds Code A's maximum throughout the perturbation interval, with a certified gap of at least $3/5$. (b) Risk difference relative to the full-input Bayes improvement over the best constant classifier. Points use 65,536 Monte Carlo samples per configuration; the dashed line is the uniform $3/35$ bound.

C.5 Perturbations that preserve matching and couple groups

We formalize the local robustness of the separation under perturbations that preserve the matched binary-task signal norms.

Corollary 7 (Stability under the matching constraints)**.** *For every positive-perturbation construction satisfying Theorem 4, there exists an open neighborhood surrounding its posterior Gram matrix (relative to the linear constraints preserving equality of corresponding bit norms) that maintains valid, distinct class posteriors and sustains a relative risk gap strictly exceeding* $3/70$*. For* $r \geq 2$*, this robust neighborhood permits nonzero cross-group couplings between the active class clusters.*

Proof. We now perturb the class-probability coefficients while holding both codes fixed. We express the matching conditions through centered posterior Gram matrices $\mathcal{G} = A^\top A \in \mathbb{R}^{C\times C}$, whose entries record inner products of class coefficient vectors. For each bit index $j \in \{1, \ldots, C-1\}$, let $m_{A,j}, m_{B,j} \in \mathbb{R}^C$ denote the corresponding lifted code columns. We collect the matching conditions in the linear space

$$\mathcal{L} = \{\mathcal{G} = \mathcal{G}^\top : \mathcal{G}\mathbf{1} = 0, (m_{A,j}^\top \mathcal{G} m_{A,j}) = (m_{B,j}^\top \mathcal{G} m_{B,j}) \text{ for every } j\},$$

where the quadratic forms equal the squared conditional-mean coefficient norms of the corresponding bits. At every positive-ϵ instance, $\mathcal{G}$ is strictly positive definite on $\mathbf{1}^\perp$, and $\sqrt{d\mathcal{G}_{cc}} < 1/C$. Both properties persist in a sufficiently small neighborhood within $\mathcal{L}$. We take the square-root factorization of each perturbed $\mathcal{G}$ and pad it into dimension d. Under spherical inputs, this factorization preserves the posterior geometry up to an orthogonal transformation of the features.

We use continuity to carry the separation to this neighborhood. The separated spectral projectors vary continuously with $\mathcal{G}$, as do Gaussian width and the positive Bayes-improvement denominator. The relative gap, initially at least $3/35$, therefore remains above $3/70$ in a sufficiently small relative-open neighborhood. The zero-intersection property also persists, since the largest cross-subspace canonical correlation—the largest absolute inner product between unit vectors in the two retained spaces—is strictly below one. We thus establish the local assertion, without specifying a uniform radius for the neighborhood.

For $r \geq 2$, we restrict perturbations to symmetric off-diagonal blocks connecting the centered seven-dimensional contrast spaces of active groups. Each eight-class group has seven independent class contrasts, so each pair of groups allows a 7×7 coupling block. This space has dimension $49\binom{r}{2}$. Centering is automatic; at most $C-1$ additional linear constraints enforce bit-norm matching. Since $49\binom{r}{2} > C - 1$ for $C = 8m < 16r$ and $r \geq 2$, a nonzero cross-group direction lies in the constraint kernel. By scaling this direction sufficiently small, we stay within the neighborhood above and obtain matched separation with genuine cross-group posterior couplings. This establishes existence within the matching constraints; it does not quantify how frequently the effect occurs for unrestricted data distributions.

□

D BEYOND AFFINE POSTERIORS: CLASS-CONDITIONAL RISK

We retain the same prediction pipeline: encode each class label with M, regress its code vector on the input, and reconstruct class scores. We now replace the spherical input U with a general feature vector $X \in \mathbb{R}^d$ and allow nonlinear class probabilities. The squared-loss solution still depends only on means and covariances. We introduce additional distributional assumptions to evaluate the risk of the resulting classifier.

Let $Y \in \{1, \dots, C\}$ be the class label, $\pi_c = \Pr\{Y = c\} > 0$ its prior probability, and $\mu_c = \mathbb{E}[X \mid Y = c] \in \mathbb{R}^d$ its conditional feature mean. We use π to collect the C priors and retain a centered, full-span code $M \in \{-1, +1\}^{C \times n}$, where n counts binary targets, and a rank limit r on their shared linear predictor. Write $\bar{\mu} = \sum_c \pi_c \mu_c$, $\Sigma = \mathrm{Cov}(X) \succ 0$, and

$$D = [\pi_c(\mu_c - \bar{\mu})]_{c=1}^{C}, \quad S_\lambda = (\Sigma + \lambda I)^{-1/2}, \quad T_\lambda = S_\lambda D.$$

The matrix $D \in \mathbb{R}^{d \times C}$ records feature–label cross-covariances. We use ridge penalty $\lambda \geq 0$ and covariance normalizer $S_\lambda \in \mathbb{R}^{d \times d}$ to obtain the normalized cross-covariance $T_\lambda \in \mathbb{R}^{d \times C}$. We use $\Theta \in \mathbb{R}^{d \times n}$ to predict the n coded targets from the centered input $X - \bar{\mu}$. With a free intercept and penalty $\lambda \|\Theta\|_F^2$, we apply the same reduced-rank calculation to obtain the reconstructed class coefficients

$$B_{M,r} = S_\lambda P_{M,r} T_\lambda, \qquad P_{M,r} = \mathrm{Top}_r(T_\lambda M M^\top T_\lambda^\top).$$

The operator Top_r returns an orthogonal projector onto r leading eigenvectors. The matrix $B_{M,r} \in \mathbb{R}^{d \times C}$ contains reconstructed class-score coefficients; we denote its cth column by $b_{M,c}$. We predict $h_M(x) = \arg\max_c \{\pi_c + b_{M,c}^\top (x - \bar{\mu})\}$. We need only finite second moments for this calculation, so it remains valid without an affine posterior. Regression-based discriminant analysis and its estimation theory are established (Bing et al., 2026). We extend the risk analysis to the specified coded learner with its fixed reconstruction decoder.

We compute accuracy from the probability that the score comparison selects the true class. Under Gaussian class distributions, the $C - 1$ score differences are jointly Gaussian, allowing us to express accuracy as a decision-region probability. In the proposition, $\Sigma_c \in \mathbb{R}^{d \times d}$ is the within-class covariance, $L_c \in \mathbb{R}^{(C-1) \times d}$ contains contrasts against $k \neq c$, and $\delta_{\pi,c} \in \mathbb{R}^{C-1}$ contains prior contrasts. For the approximation statement, we use $V \in \mathbb{R}^{d \times r}$ to map inputs to retained coordinates $V^\top X$. We measure distributional discrepancy by the convex-set distance: the largest absolute difference in probabilities assigned to a convex region in these coordinates.

Proposition 8 (Class-conditional risk and its approximation)**.** *If $X \mid Y = c \sim \mathcal{N}(\mu_c, \Sigma_c)$ with $\Sigma_c \succ 0$, then*

$$R(h_M) = 1 - \sum_{c=1}^{C} \pi_c \Pr\{Z_c \in \mathcal{O}_c\}, \quad Z_c \sim \mathcal{N}(L_c(\mu_c - \bar{\mu}) + \delta_{\pi,c},\ L_c \Sigma_c L_c^\top), \tag{15}$$

where L_c has rows $(b_{M,c} - b_{M,k})^\top$, $\delta_{\pi,c,k} = \pi_c - \pi_k$, and $\mathcal{O}_c$ implements the fixed argmax tie rule. For a general conditional distribution and a classifier that maximizes affine class scores in $V^\top X \in \mathbb{R}^r$, replace each conditional distribution by a Gaussian with the same projected mean and covariance. If their convex-set distances are $\zeta_{c,V}$, the resulting risk R_G satisfies

$$|R(h) - R_G(h)| \leq \sum_c \pi_c \zeta_{c,V}.$$

Because each decision region is an intersection of halfspaces in the retained representation, we control the approximation in dimension r, rather than the ambient dimension d. For example, conditional on $Y = c$, suppose $X = \mu_c + \sum_i \xi_{c,i}$ with independent centered summands. Each $\xi_{c,i} \in \mathbb{R}^d$ contributes one independent component of the feature noise; i indexes these components, not training examples. We set $\Omega_{c,V} = \sum_i \mathrm{Cov}(V^\top \xi_{c,i}) \succ 0$, the covariance of the retained within-class noise. We apply the convex-set normal approximation of Raič (2019, Theorem 1.1) to obtain

$$\zeta_{c,V} \leq \min\left\{1, (42 r^{1/4} + 16) \sum_i \mathbb{E}\left[\|\Omega_{c,V}^{-1/2} V^\top \xi_{c,i}\|^3\right]\right\}. \tag{16}$$

This established theorem supplies a sufficient condition for the Gaussian risk approximation. We do not assume that its bound is small for pretrained features. In Appendix F, we derive the reduction and account for degenerate score covariances and ties.

We next construct a Gaussian class model in which the code-dependent loss of information persists. The two matched codes retain different coordinates, and their risk ordering remains unchanged when we use the best possible decoder of each coordinate. Here δ sets the separation of the class means, Φ is the standard normal cumulative distribution function, and $R^*_{M,1}$ denotes the smallest risk attainable by any decoder of code M's retained rank-one representation.

Proposition 9 (A separation with nonlinear Gaussian posteriors)**.** *Let Y be uniform on eight classes and $X \mid Y = c \sim \mathcal{N}(\delta(B_0)_{:,c}, I_2)$, where B_0 is the integer matrix in Proposition 6. Use $\lambda = 0$ and $0 < \delta < 2/3$. At rank one, code M_A retains the second coordinate and code M_B the first. With the prescribed reconstruction decoder,*

$$R(M_A, 1) - R(M_B, 1) = \frac{\Phi(2\delta) - \Phi(\delta)}{4} > 0.$$

If we instead use the Bayes-optimal decoder within each retained coordinate, the risk difference remains positive and equals

$$R^*_{M_A,1} - R^*_{M_B,1} = \frac{2\Phi(\delta/2) - 1}{4}.$$

We obtain a valid softmax posterior with affine logits on unbounded support. At $\delta = 1/2$, the two gaps are approximately 3.75 and 4.94 percentage points. The codes preserve their matched spectra and distance histograms, although we do not assert matched bit Bayes errors for this Gaussian construction. In Appendix F, we describe the coincident class means at the analytic base and their perturbations. By allowing the optimal decoder within each coordinate, the second formula isolates the loss caused by the retained representation.

E AUDITING SELECTION AFTER REFITTING

A code-selection procedure may compare candidates on one fitting sample and then refit the selected classifier on another. We assess the resulting final classifiers on a separate labeled audit sample. By comparing their losses on the same examples, we distinguish changes that affect correctness from changes between two incorrect predictions. Prediction disagreement alone does not make this distinction.

We fix the final classifiers $f_1, \dots, f_K$, with $K \geq 2$, independently of the audit. Each maps an input feature vector to one of C classes. Let $N_{\mathrm{aud}} \geq 2$ be the audit sample size. For independent labeled observations (X_i, Y_i), $i = 1, \dots, N_{\mathrm{aud}}$, we define

$$\widehat{\ell}_{f,j} = \frac{1}{N_{\mathrm{aud}}} \sum_i \mathbf{1}\{f_j(X_i) \neq Y_i\}, \qquad Z_{i,jk} = \mathbf{1}\{f_j(X_i) \neq Y_i\} - \mathbf{1}\{f_k(X_i) \neq Y_i\}.$$

We denote by $\widehat{v}_{jk}$ the sample variance of $Z_{i,jk}$, using the denominator $N_{\mathrm{aud}} - 1$, and set $t = \log(2K(K-1)/\alpha)$ for an allowed failure probability $\alpha \in (0, 1)$. The indicator $\mathbf{1}\{\cdot\}$ equals one when its event holds and zero otherwise. Thus $Z_{i,jk}$ records the error of candidate j minus that of candidate k on the same example.

We compare the deployed classifier with the best final classifier in the candidate family. By controlling all pairwise comparisons simultaneously, we allow the deployed classifier to be selected after inspecting the audit.

Proposition 10 (A paired audit of the deployed candidate family)**.** *With probability at least $1 - \alpha$, every audit-dependent selected index $\widehat{j}$ satisfies*

$$R(f_{\widehat{j}}) - \min_k R(f_k) \leq \max\left\{0, \max_{k \neq \widehat{j}} \left[\widehat{\ell}_{f,\widehat{j}} - \widehat{\ell}_{f,k} + \sqrt{\frac{2\widehat{v}_{\widehat{j}k}\, t}{N_{\mathrm{aud}}}} + \frac{14t}{3(N_{\mathrm{aud}} - 1)}\right]\right\}. \tag{17}$$

We obtain this guarantee by applying the empirical Bernstein inequality of Maurer & Pontil (2009, Theorem 11) to $(Z_{i,jk} + 1)/2$ and taking a union bound. The loss difference is zero when both

classifiers are wrong, even if their predictions differ. We include every candidate code and penalty in the audit family before selection. Independent stratified observations are also permitted when their allocation matches the target class proportions. If a subsequent fit uses audit labels, it requires a new independent audit. In Appendix G, we give the proof and derive a signed pilot-to-final transfer bound and a margin-dependent RRR perturbation bound. These guarantees quantify the uncertainty that remains after auditing; that uncertainty can still exceed the observed performance differences.

F CLASS-CONDITIONAL EXTENSIONS AND SELECTION/REFITTING PROOFS

We first derive the coded regression solution from second moments, then introduce the distributional assumptions needed to evaluate its classification risk. For Gaussian class distributions, we express risk through the probabilities of the classifier's decision regions. We then use a normal-approximation bound to state when this calculation also approximates risk under non-Gaussian features. Finally, we account for prediction changes between the pilot and final fits.

F.1 POPULATION CODED REGRESSION FOR ARBITRARY POSTERIORS

We fit the centered code target to the centered features. We write $e_Y \in \mathbb{R}^C$ for the observed class's one-hot vector, so that $M^\top e_Y \in \mathbb{R}^n$ is its code vector. With $X_0 = X - \bar{\mu} \in \mathbb{R}^d$ and $Y_0 = e_Y - \pi \in \mathbb{R}^C$, we have $\mathbb{E}[X_0 Y_0^\top] = D$ and $D\mathbf{1} = 0$. The free coded intercept is $M^\top \pi$. We optimize $\Theta \in \mathbb{R}^{d\times n}$ subject to the rank limit r used in the main model. After collecting terms that do not depend on Θ, we write the ridge objective as

$$\operatorname{tr}\{\Theta^\top(\Sigma + \lambda I)\Theta\} - 2\operatorname{tr}(\Theta^\top DM) = \|S_\lambda^{-1}\Theta - T_\lambda M\|_F^2 + \text{constant}.$$

Since multiplication by the invertible S_λ^{-1} preserves rank, we obtain a global solution $\Theta = S_\lambda(T_\lambda M)_r$, where $(Q)_r$ denotes a best rank-r approximation to a matrix Q in Frobenius norm. Using $(T_\lambda M)_r = P_{M,r}T_\lambda M$ and $MM^\dagger = H_C$, we recover $B_{M,r} = S_\lambda P_{M,r} T_\lambda$. This standard reduced-rank identity expresses the fit in covariance-normalized coordinates. At full centered rank, it establishes code invariance for any posterior. It does not establish Bayes optimality of the resulting affine classifier; Gaussian posteriors generally require a different decoder to attain their Bayes rule.

F.2 PROOF OF PROPOSITION 8

For each true class c, we characterize correct classification by comparing its score with every $k \neq c$. The tie rule determines which comparisons require strict inequalities. Under a Gaussian conditional distribution, the contrast vector has the mean and covariance in Equation (15). We obtain the risk formula by weighting the corresponding correctness probabilities by the class priors. When the score matrix has rank $r < C - 1$, the contrast covariance may be singular. We therefore interpret each probability as a Gaussian measure on its actual support, without requiring a nonsingular density.

For concreteness, we use the smallest-index argmax rule: the comparison is strict against $k < c$ and weak against $k > c$. The resulting half-open decision regions are measurable and convex. This convention assigns tied scores to one label, so identical score functions cannot contribute correctness for several labels at once. We use the same deterministic rule in Monte Carlo evaluation.

More generally, we write the scores as $F^\top V^\top x + b$, where $V \in \mathbb{R}^{d\times r}$ forms the retained features, $F \in \mathbb{R}^{r\times C}$ maps them to class scores, and $b \in \mathbb{R}^C$ contains the score intercepts. We take V to have full column rank r and use the effective dimension if the rank is smaller. Each correctness region is a convex subset of $\mathbb{R}^r$, so its probability changes by at most the convex-set distance between the two projected conditional laws. Weighting these changes by the class priors gives Proposition 8. For independent summands, the standardized projected terms $\Omega_{c,V}^{-1/2} V^\top \xi_{c,i}$ are centered and independent, and their covariances sum to I_r. We obtain Equation (16) by applying Raič (2019, Theorem 1.1) to these terms. The theorem covers measurable convex sets, including the half-open tie regions defined above. Because the Gaussian reference preserves the class means and covariances, it also preserves the population squared-loss fit. We thus compare the risk of the same classifier under two distributions. Estimating the moments introduces a separate error, which we assess on independent validation examples.

To illustrate the role of dimension, we take the within-class noise to be a normalized sum of m independent Rademacher vectors with independent coordinates, and take V to have orthonormal columns. The retained noise covariance is then I_r. Writing $\xi \in \{-1, +1\}^d$ for one unnormalized sign vector, we calculate $\mathbb{E}\big[\|V^\top \xi\|^4\big] \le r(r+2)$ and consequently $\mathbb{E}\big[\|V^\top \xi\|^3\big] \le r\sqrt{r+2}$. We obtain the sufficient risk-error bound

$$\min\{1, (42r^{1/4} + 16)r\sqrt{r+2}/\sqrt{m}\}.$$

The bound depends on the retained rank through the projected moments. Its numerical constant can be conservative, so we use this example to illustrate the approximation rather than to certify its accuracy for real representations.

F.3 Proof of Proposition 9

We use the two-by-eight matrix B_0 from Proposition 6. Its centered, orthogonal rows have squared norms 12 and 6, giving

$$\Sigma = I_2 + \delta^2 B_0 B_0^\top/8 = \mathrm{diag}(1 + 3\delta^2/2,\ 1 + 3\delta^2/4), \qquad D = \delta B_0/8.$$

After normalizing the features, we obtain diagonal code-weighted covariances with entries proportional to

$$M_A : \left(\frac{48}{1+3\delta^2/2}, \frac{60}{1+3\delta^2/4}\right), \quad M_B : \left(\frac{60}{1+3\delta^2/2}, \frac{48}{1+3\delta^2/4}\right).$$

These entries show that Code A always selects the second coordinate. Code B selects the first precisely when $12 > 27\delta^2$, giving $0 < \delta < 2/3$. Reversing the diagonal feature normalization preserves both coordinate lines.

The reconstructed rank-one scores share an intercept; their slopes are proportional to the retained class means. Thus the classifier selects a class with the largest projected mean on the positive half-line and one with the smallest projected mean on the negative half-line. For a unit retained direction v and class means μ_c, we obtain the exact correctness probability

$$1 - R = \{\Phi(\max_c v^\top \mu_c) + \Phi(-\min_c v^\top \mu_c)\}/8.$$

The fixed rule resolves ties among identical extrema. Since the tied labels have the same projected distribution, their ordering does not affect this formula. The first row of δB_0 has extrema $\pm 2\delta$, whereas the second has extrema $\pm\delta$. We obtain the first gap by subtracting the resulting risks.

We next evaluate the Bayes-optimal decoder within each representation by integrating the largest prior-weighted projected class density. Under equal priors, coincident projected means contribute the same density to this maximum. For k distinct, equally spaced means with separation δ and unit variance, this integral is $1 + (k-1)(2\Phi(\delta/2) - 1)$. To obtain it, we split the line at consecutive midpoints. Each interior interval contributes $2\Phi(\delta/2) - 1$, and the two outer intervals together contribute $2\Phi(\delta/2)$. There are five distinct means in the first coordinate and three in the second. We obtain the second gap after weighting each correctness probability by $1/8$.

The resulting posterior is

$$\eta_c(x) = \frac{\exp(\delta (B_0)_{:,c}^\top x - \delta^2 \|(B_0)_{:,c}\|^2/2)}{\sum_k \exp(\delta (B_0)_{:,k}^\top x - \delta^2 \|(B_0)_{:,k}\|^2/2)}.$$

This posterior is nonlinear and valid throughout $\mathbb{R}^2$. Although some class means coincide in this analytic construction, we separate them in the experiment through the ten-dimensional extension $\mu_c = (\delta (B_0)_{:,c}^\top, \epsilon (H_8)_{:,c}^\top)^\top$, with identity within-class covariance. We report every prescribed perturbation level. A positive ϵ makes all class means distinct. The perturbed outcomes provide numerical evidence; the exact Gaussian gap formulas above apply at $\epsilon = 0$. The marginal bit-difficulty matching established for the spherical affine family need not persist under this distributional change.

F.4 A disagreement-based selection/refitting certificate

A pilot risk score evaluates one classifier, whereas deployment may use a refitted classifier. We quantify this change without imposing an eigengap or a posterior model. Let g_j and f_j, $j \le K$,

denote the pilot and final classifiers for every candidate code and penalty. Their fitting samples may overlap, but the labeled audit sample must be independent of all fitting data. Conditional on those data, we evaluate the classifiers on N_{aud} independent audit examples. We write q_j for the numerical risk prediction obtained at the pilot stage and $\alpha \in (0,1)$ for the allowed failure probability. We denote the pilot's mean audit loss by $\widehat{\ell}_{g,j}$ and its fraction of prediction disagreements with the final classifier by $\widehat{d}_j$. We count the audit labels in the total sample budget.

We separate the observable discrepancy into pilot-score calibration and prediction changes caused by refitting. Concentration terms account for the uncertainty in estimating these two quantities from a finite audit sample.

Proposition 11 (A finite-sample selection/refitting certificate)**.** *For any pilot scores q_j and $a_{\mathrm{aud}} = \sqrt{\log(4K/\alpha)/(2N_{\mathrm{aud}})}$, with probability at least $1-\alpha$, simultaneously for all j,*

$$|R(f_j) - q_j| \le \underbrace{|q_j - \widehat{\ell}_{g,j}| + \widehat{d}_j + 2a_{\mathrm{aud}}}_{c_j}. \tag{18}$$

Thus any selected index $\widehat{j}$, including an audit-dependent choice, satisfies

$$R(f_{\widehat{j}}) - \min_j R(f_j) \le q_{\widehat{j}} + c_{\widehat{j}} - \min_j (q_j - c_j).$$

If $\widehat{j} \in \arg\min_j q_j$ and $j^* \in \arg\min_j R(f_j)$, the sharper bound $c_{\widehat{j}} + c_{j^*}$ holds.

We prove the certificate by combining finite-family concentration with $|R(f_j) - R(g_j)| \le \Pr\{f_j(X) \ne g_j(X)\}$. The observable calibration and disagreement terms account for the pilot fit, selection, and final refit, but need not be small. This application of standard concentration controls the transfer of risk between classifiers; it does not provide a parameter-estimation rate. In our refit-aware Gaussian selector, we instead evaluate the final candidate directly under the pilot class model. That scoring step therefore avoids classifier-transfer error, while the error in the class model still requires assessment.

F.5 Proof of Proposition 11

We condition on the pilot and final-fitting observations and on any simulation randomness used to construct q_j. The audit sample remains independent of those observations. We control $2K$ Bernoulli quantities: the errors of g_j and the prediction disagreements of g_j, f_j. Applying two-sided Hoeffding inequalities (Hoeffding, 1963, Theorem 2) and a union bound gives, with probability at least $1-\alpha$,

$$|R(g_j) - \widehat{\ell}_{g,j}| \le a_{\mathrm{aud}}, \qquad |\Pr\{f_j(X) \ne g_j(X)\} - \widehat{d}_j| \le a_{\mathrm{aud}} \quad (j \le K).$$

These quantities may use the same audit observations because the union bound does not require their mutual independence. For each example, changing a prediction changes the zero-one loss by at most the disagreement indicator. We therefore apply the triangle inequality to obtain

$$|R(f_j) - q_j| \le |R(g_j) - q_j| + \Pr\{f_j(X) \ne g_j(X)\} \le |q_j - \widehat{\ell}_{g,j}| + \widehat{d}_j + 2a_{\mathrm{aud}}.$$

This proves Equation (18). We obtain the oracle inequality by subtracting the smallest lower risk bound from the selected model's upper bound. When the selected index minimizes q_j, we compare it with a final-risk minimizer and use $q_{\widehat{j}} - q_{j^*} \le 0$ to obtain the sharper bound.

We must include every candidate penalty, rank, and code that may be selected after inspecting the audit labels. In particular, selecting a rank from the audit requires adjusting K to cover that choice. The models may share training observations, provided their construction remains independent of the audit. If a later fit also uses the audit labels, we need a new independent audit to certify that fit. The term $\widehat{d}_j$ measures changes in predictions directly, without assuming stability of the empirical eigenspaces.

We can use the same Hoeffding argument for independent stratified observations when their class proportions equal the target priors. Individual observations may have different expected losses; their average expectation still equals the target risk. Our balanced representation tasks use equal class counts. To interpret the bounds as population guarantees, we continue to assume that training, validation, and test examples share the relevant class-conditional distributions. The certificate does not cover dataset shift. We report its numerical size, including loose cases, to show whether the guarantee is informative in each experiment.

G PAIRED LOSS AUDITS AND MARGIN-DEPENDENT FITTING STABILITY

We now prove the paired audit bound and examine how refitting affects its interpretation. We first compare the final classifiers using an empirical Bernstein inequality, then retain the sign of pilot-to-final loss changes to refine the transfer calculation. Finally, we relate perturbations in the RRR fit to prediction changes through classification margins. Throughout these audit calculations, N_{aud} is the number of independent audit examples, K is the number of candidates, and $\alpha \in (0,1)$ is the allowed failure probability.

G.1 PROOF OF PROPOSITION 10

We condition on the fitting observations and simulation randomness used to construct the K candidates. For each ordered pair $j \neq k$, the variables $(Z_{i,jk}+1)/2$ lie in $[0,1]$ and are independent across audit observations; their sample variance is $\widehat{v}_{jk}/4$. We apply Maurer & Pontil (2009, Theorem 11) with failure probability $\alpha/(K(K-1))$ and rescale the result by two. This bounds $R(f_j) - R(f_k)$ by the expression in Equation (17). A union bound makes these comparisons simultaneous over all ordered pairs, allowing selection on the same audit. We then maximize over the competitors to prove the proposition. The $j = k$ term is zero and needs no concentration bound.

The cited inequality permits independent observations with different distributions. For stratified sampling, the average expected loss equals the target risk when the allocation proportions match the target priors; other allocations require a weighted analysis. Neither the candidate classifiers nor their pairwise statistics need to be mutually independent. Their construction must, however, remain independent of the audit examples. For comparison, we can apply a direct Hoeffding audit to the same candidate family and obtain the separate bound

$$\min\{1, \widehat{\ell}_{f,\widehat{j}} + a\} - \min_k \max\{0, \widehat{\ell}_{f,k} - a\}, \qquad a = \sqrt{\log(2K/\alpha)/(2N_{\mathrm{aud}})}.$$

Each procedure has its own confidence guarantee. To select the smaller bound after computing both, we would need to allocate the failure probability jointly; otherwise, the stated confidence level need not be preserved.

G.2 SIGNED TRANSFER FROM PILOT TO FINAL FITS

For a pilot classifier g_j, its final counterpart f_j, and a pilot risk score q_j, we define the signed loss change $D_{ij} = \mathbf{1}\{f_j(X_i) \neq Y_i\} - \mathbf{1}\{g_j(X_i) \neq Y_i\}$, with audit mean $\widehat{\Delta}_j$ and sample variance $s^2_{D,j}$. We write $\widehat{\ell}_{g,j}$ and $s^2_{g,j}$ for the pilot loss mean and variance. With $t_0 = \log(8K/\alpha)$, we apply Maurer & Pontil (2009, Theorem 11) to both tails of all $2K$ quantities and take a union bound. This gives the simultaneous guarantee

$$|R(f_j) - q_j| \leq |q_j - \widehat{\ell}_{g,j}| + |\widehat{\Delta}_j| + \sqrt{2s^2_{g,j}t_0/N_{\mathrm{aud}}} + \frac{7t_0}{3(N_{\mathrm{aud}}-1)} + \sqrt{2s^2_{D,j}t_0/N_{\mathrm{aud}}} + \frac{14t_0}{3(N_{\mathrm{aud}}-1)}.$$

As in Proposition 11, we obtain the oracle comparison by subtracting the smallest resulting lower risk bound from the selected upper bound. Retaining the sign can reduce the empirical transfer term through cancellation, while pairing can reduce its variance. The resulting bound need not improve at every sample size because its concentration constants and union-bound allocation also contribute.

Prediction disagreement can overstate the transfer error when the two classifiers predict different incorrect labels: then $D_{ij} = 0$, although $f_j(X_i) \neq g_j(X_i)$. A direct audit of final losses also avoids passing through pilot calibration and an additional triangle inequality. Indeed, adding the empirical signed correction to the empirical pilot loss gives the empirical final loss exactly. This identity explains the relation between the audits; it does not define a new selector.

G.3 A MARGIN-DEPENDENT PERTURBATION BOUND FOR CODED RRR

We next bound how changes in a fitted representation can alter its class predictions. The relevant comparison is between the score perturbation and the margin separating the largest score from the second largest. For a fixed coding metric $K_M = MM^\top \in \mathbb{R}^{C\times C}$, we compare two covariance-normalized geometries (S,T) and $(\widehat{S},\widehat{T})$. Here $S, \widehat{S} \in \mathbb{R}^{d\times d}$ normalize the features, and $T, \widehat{T} \in$

$\mathbb{R}^{d\times C}$ are the normalized feature–label cross-covariances. We denote by P and $\widehat{P}$ their leading rank-r projectors, and define $B = SPT$ and $\widehat{B} = \widehat{S}\widehat{P}\widehat{T}$. We include any feature-centering correction in the score intercepts $b, \widehat{b} \in \mathbb{R}^C$. The reference score vector is then $B^\top x + b$, with $B \in \mathbb{R}^{d\times C}$. The classifier h selects its largest entry, while $\widehat{h}$ uses the estimated scores. We define

$$e_T = \|\widehat{T} - T\|_{\mathrm{op}}, \quad e_S = \|\widehat{S} - S\|_{\mathrm{op}}, \quad \delta_S = \|K_M\|_{\mathrm{op}}(2\|T\|_{\mathrm{op}} e_T + e_T^2).$$

Suppose $TK_MT^\top$ has truncation gap $\gamma > 2\delta_S$, where γ is its rth eigenvalue minus its $(r+1)$st. Weyl's inequality bounds the reduction in spectral separation by the covariance perturbation. Under the stated gap condition, the separation therefore remains larger than half the reference gap. We apply the operator-norm version of Yu et al. (2015, Theorem 1 and the following discussion) to obtain $\|\widehat{P} - P\|_{\mathrm{op}} \le 2\delta_S/\gamma$. We then expand the reconstructed coefficient difference as

$$\widehat{B} - B = (\widehat{S} - S)PT + \widehat{S}(\widehat{P} - P)T + \widehat{S}\widehat{P}(\widehat{T} - T)$$

and bound its three terms to obtain

$$\|\widehat{B} - B\|_{2,\infty} \le e_S\|T\|_{2,\infty} + \|\widehat{S}\|_{\mathrm{op}}\left(\frac{2\delta_S}{\gamma}\|T\|_{2,\infty} + \|\widehat{T} - T\|_{2,\infty}\right) =: e_B.$$

Here $\|A\|_{2,\infty}$ denotes the largest Euclidean column norm. If $\Pr\{\|X\| > L\} \le \xi$, every class-score error outside this tail event is at most $e = \|\widehat{b} - b\|_\infty + Le_B$. We write $m(X)$ for the gap between the largest and second-largest reference scores. The predictions agree whenever $m(X) > 2e$, which gives

$$|R(\widehat{h}) - R(h)| \le \xi + \Pr\{m(X) \le 2e\}.$$

Under the explicit margin condition $\Pr\{m(X) \le t\} \le C_m t^a$, with $C_m > 0$ and $a > 0$ controlling the probability of nearly tied scores, we obtain $\xi + C_m(2e)^a$. When the reference scores equal the true posterior, we can also bound the posterior mass lost on each changed decision. This gives the sharper excess-risk bound $\xi + C_m(2e)^{a+1}$. The additional conclusion relies on the scores being posterior probabilities; it does not follow for an arbitrary regression solution.

This calculation separates the effects of the coding metric and the classification problem. The metric contributes through $\|K_M\|_{\mathrm{op}}$ and the selected spectral gap, while the margin distribution determines how sensitive predictions are to the resulting score perturbations. We obtain the bound by combining standard subspace perturbation and margin arguments under the stated assumptions. It does not provide an unconditional rate for learned representations. Small ridge penalties or spectral gaps, heavy feature tails, and substantial mass near ties can all weaken the guarantee. We therefore report the empirical audit sizes without claiming a sharp sample-complexity improvement.

H Supporting decision-optimality and robustness results

H.1 Direct decision selection

We first derive the rank-one decision identity from Section 3. We then use it to show why whitening need not select the best decision direction and to control the error of direct selection from estimated coefficients. Throughout, we work under the spherical affine model in Equation (3) and define $S_A(P) = \kappa_d w(P\mathcal{V})$ and $R_r^*(A) = \min_{\mathrm{rank}(P)=r} R(P)$, with optimization over orthogonal projectors.

Rank-one identity and optimal direction. For a rank-one projector $P = vv^\top$ with a unit feature direction $v \in \mathbb{R}^d$, we write $\alpha_+ = \max_c v^\top a_c$ and $\alpha_- = \min_c v^\top a_c$. The projection reduces the Gaussian-width calculation to one dimension. For a standard Gaussian vector $G \sim \mathcal{N}(0, I_d)$, $Z = v^\top G \sim \mathcal{N}(0,1)$, and

$$\mathbb{E}\big[\max_c Zv^\top a_c\big] = \alpha_+\mathbb{E}[Z\mathbf{1}_{Z\ge 0}] + \alpha_-\mathbb{E}[Z\mathbf{1}_{Z<0}] = \frac{\alpha_+ - \alpha_-}{\sqrt{2\pi}}.$$

To optimize the direction, we interchange the finite maximum over column pairs with the supremum over unit v and apply Cauchy–Schwarz. We obtain $\max_{i,j}\|a_i - a_j\|_2/\sqrt{2\pi}$, so a longest pairwise difference determines an optimal direction. If all columns coincide, centering implies $A = 0$. Every projector is then optimal, and Equation (8) continues to hold with $D_A = 0$.

Why whitening need not preserve the best decision direction. We now construct a posterior for which the direction of greatest squared-loss variation retains an arbitrarily small fraction of the optimal decision improvement. Let $a, b > 0$ and take $d = 2$, $C = 2m + 2$, with an integer $m > 2$. We construct A using columns $\pm ae_1$ once each and $\pm be_2$ repeated m times each, where e_1, e_2 are the standard coordinate vectors. We choose $a/b = \sqrt{m/2}$ and scale both magnitudes so that $\sqrt{2}a \le 1/C$. These choices give a centered, valid affine posterior. The coefficient polytope has vertices $\pm ae_1, \pm be_2$ and diameter $2a$. Our rank-one identity therefore selects e_1 as the decision-optimal direction.

Squared loss weights the repeated columns through their contribution to the covariance: $AA^\top = 2\,\mathrm{diag}(a^2, mb^2)$ has eigenvalues $2a^2$ and $2mb^2 = 4a^2$. One-hot and whitened squared loss therefore select e_2. To compare the resulting decisions, we note that projection onto a coordinate with half-range t improves on the best constant classifier's risk by $2\kappa_2 t/\sqrt{2\pi}$. Squared loss consequently retains the following fraction of the optimal improvement:

$$\frac{b}{a} = \sqrt{\frac{2}{m}} \longrightarrow 0.$$

The vanishing fraction describes a relative loss of decision information; posterior validity also makes the absolute improvement shrink along this family. We use repeated posterior columns to represent distinct classes with identical conditional probabilities. We do not require repeated code columns.

Estimation and optimization error without an eigengap. We next quantify how coefficient estimation and approximate optimization affect direct decision selection. We define $S_{\widehat{A}}$ analogously to S_A using $\widehat{\mathcal{V}} = \mathrm{conv}\{\widehat{a}_1, \dots, \widehat{a}_C\}$. Suppose $\max_c \|\widehat{a}_c - a_c\|_2 \le \epsilon_A$ and a rank-r orthogonal projector $\widehat{P}$ satisfies

$$S_{\widehat{A}}(\widehat{P}) \ge \max_{\mathrm{rank}(P)=r} S_{\widehat{A}}(P) - \xi,$$

where the maximum ranges over orthogonal projectors and $\xi \ge 0$ measures the optimization error. Combining the width characterization with a Lipschitz bound, we obtain the following guarantee for $\widehat{h}(u) = \arg\max_c \widehat{a}_c^\top \widehat{P} u$:

$$R(\widehat{h}) - R_r^*(A) \le 2\kappa_d \sqrt{2\log(2C)}\,\epsilon_A + \xi. \tag{19}$$

The bound requires no eigengap. At rank one, we can set $\xi = 0$ by selecting a longest pairwise difference of the estimated coefficients; if all estimated columns coincide, any rank-one projector is optimal. At higher ranks, we retain the optimization error explicitly.

To prove the bound, we compare projected scores under the true coefficient matrix A and its estimate $\widehat{A}$, using the column-error bound ϵ_A. We set $L_C = \sqrt{2\log(2C)}$. For any deterministic orthogonal projector P, each $G^\top P(\widehat{a}_c - a_c)$ is a centered Gaussian with variance at most ϵ_A^2. We apply the standard exponential-moment bound to these C variables and their negatives to obtain

$$\mathbb{E}_G\big[\max_c |G^\top P(\widehat{a}_c - a_c)|\big] \le L_C \epsilon_A.$$

The maximum of absolute values is positively homogeneous, so we can apply the same radial identity used for Gaussian width. For every P, this gives

$$\mathbb{E}_U\big[\max_c |U^\top P(\widehat{a}_c - a_c)|\big] \le \kappa_d L_C \epsilon_A, \qquad |S_A(P) - S_{\widehat{A}}(P)| \le \kappa_d L_C \epsilon_A.$$

Because the bound uses the same constant for every P, we can apply it to the selected $\widehat{P}$ after conditioning on the coefficient estimate. We evaluate risk on a fresh test feature independent of the pilot estimate.

Spherical symmetry gives $\mathbb{E}[U \mid \widehat{P}U] = \widehat{P}U$. We use this identity and the measurability of $\widehat{h}$ with respect to $\widehat{P}U$ to compare its success probability with the optimized width:

$$\begin{aligned} 1 - C^{-1} - R(\widehat{h}) &= \mathbb{E}[a_{\widehat{h}(U)}^\top \widehat{P}U] \\ &\ge S_{\widehat{A}}(\widehat{P}) - \kappa_d L_C \epsilon_A \\ &\ge \sup_{\mathrm{rank}(P)=r} S_A(P) - \xi - 2\kappa_d L_C \epsilon_A. \end{aligned}$$

Substituting the definition of $R_r^*(A)$, we obtain Equation (19) without requiring continuity of a selected eigenspace. We also do not require the coefficient estimate to define a valid posterior. For a random estimate, we condition on its coefficient-error event; the argument does not provide the probability of that event.

This result separates a statistical guarantee from the optimization needed to use it. At rank one, we attain the maximum of $S_{\widehat{A}}$ by applying the diameter construction to $\widehat{A}$, which gives $\xi = 0$. At higher ranks, the bound does not establish global convergence. Our direct-width baseline uses local optimization, so we do not have a certified ξ for that baseline. Nor does Equation (19) justify refitting a selected binary code: changing the final fitting sample can change its learned subspace. We therefore evaluate that pipeline separately.

H.2 What survives outside spherical affine posteriors?

We now examine which parts of the decision-risk identity survive beyond the spherical affine model. We isolate two possible sources of error: nonlinear class probabilities and a nonzero conditional mean of the discarded features given the retained ones. Both concern the data distribution, even when we hold the code-selected projector P fixed. We allow a general integrable feature vector $U \in \mathbb{R}^d$ and write the true posterior as

$$\eta(u) = \pi + A^\top u + e(u),$$

where $\pi \in \mathbb{R}^C$ is a fixed probability vector, $A \in \mathbb{R}^{d\times C}$ contains the affine coefficients, and $e(u) \in \mathbb{R}^C$ is the approximation residual. Only η must be a valid posterior. For a fixed orthogonal projector $P \in \mathbb{R}^{d\times d}$, we define $h_P(u) = \arg\max_c [\pi + A^\top Pu]_c$,

$$b_P(PU) = \mathbb{E}[U \mid PU] - PU, \quad \beta_P = \mathbb{E}\big[\|A^\top b_P(PU)\|_\infty\big], \quad \varepsilon = \mathbb{E}\big[\|e(U)\|_\infty\big].$$

The vector $b_P(PU)$ records the conditional mean of the discarded feature component, and β_P measures its effect on the class scores. The quantity ε measures the mean largest absolute error in the affine approximation to the class probabilities. In the following proposition, we use these quantities to bound the difference between actual regret and the affine model's predicted loss of decision information.

Proposition 12 (A distribution-dependent approximation guarantee)**.** *Let*

$$\Delta_{A,\pi}(P) = \mathbb{E}\big[\max_c [\pi + A^\top U]_c\big] - \mathbb{E}\big[\max_c [\pi + A^\top PU]_c\big].$$

If the displayed expectations are finite, then

$$|R(h_P) - R^* - \Delta_{A,\pi}(P)| \le \beta_P + 2\varepsilon. \tag{20}$$

The excess risk of h_P over the best classifier measurable with respect to PU is at most $2(\beta_P + \varepsilon)$.

For a spherical affine posterior, $\beta_P = \varepsilon = 0$, even with nonuniform π. We therefore retain the exact support expectations, although unequal intercepts prevent us from reducing them to homogeneous Gaussian width. For a general distribution, $\Delta_{A,\pi}$ is a signed reference quantity and need not be a nonnegative regret. Our guarantee thus identifies two requirements for extending the affine prediction: an accurate posterior approximation and an appropriate conditional mean. Isotropy alone supplies neither. We illustrate this distinction in the stress experiments through an isotropic discrete example for which Gaussian rankings fail.

I Proofs for approximation guarantees and empirical training

We first prove the approximation guarantee from Appendix H.2, including its comparison with the best classifier based on the retained features. We then derive the empirical coded-regression fit, establish the whitening identity, and characterize the feature subspaces that real-valued codes can select.

I.1 Proof of Proposition 12

We write $s_{\text{aff}}(u) = \pi + A^\top u \in \mathbb{R}^C$ for the affine reference score vector and $h = h_P(U)$ for the predicted class index. Conditioning on the retained feature, we decompose the classifier's success

probability as

$$\mathbb{E}\big[\eta_h(U)\big] = \mathbb{E}\big[\max_c [s_{\mathrm{aff}}(PU)]_c\big] + \mathbb{E}[a_h^\top b_P(PU)] + \mathbb{E}\big[e_h(U)\big].$$

To compare this quantity with the Bayes success probability, we use the fact that the maximum is 1-Lipschitz in the infinity norm:

$$|\mathbb{E}\big[\max_c \eta_c(U)\big] - \mathbb{E}\big[\max_c [s_{\mathrm{aff}}(U)]_c\big]| \le \varepsilon.$$

We obtain Equation (20) by subtracting the two success probabilities and bounding the conditional-mean and residual terms. For the second conclusion, we compare the affine prediction with the true posterior conditional on PU:

$$s_{\mathrm{aff}}(PU) + A^\top b_P(PU) + \mathbb{E}[e(U) \mid PU].$$

Maximizing an unperturbed score vector incurs regret at most twice the infinity norm of its perturbation. We apply this inequality conditionally on the retained feature and then use Jensen's inequality to obtain $2\beta_P + 2\varepsilon$. The resulting bound remains valid whether or not the approximation errors are small. We do not assume that these errors are observable; estimating them on real features requires a separate statistical analysis.

I.2 Exact empirical reduced-rank ridge training

We now derive the coded-regression solution used in our empirical studies. Each training example contributes one row to the data matrices: N denotes the number of fitting examples, while n remains the number of code bits. We let $\mathbf{X} \in \mathbb{R}^{N\times d}$ contain input features after subtracting each feature's sample mean, and let $\mathbf{Y} \in \mathbb{R}^{N\times C}$ contain one-hot labels after subtracting the empirical class-prior vector from every row. Thus $\mathbf{Y}M \in \mathbb{R}^{N\times n}$ contains the centered coded training targets. We set $\widehat{\Sigma} = \mathbf{X}^\top\mathbf{X}/N$ and, for a ridge penalty $\lambda > 0$, define

$$S_N = (\widehat{\Sigma} + \lambda I)^{-1/2}, \qquad T = S_N\mathbf{X}^\top\mathbf{Y}/N.$$

Here $S_N \in \mathbb{R}^{d\times d}$ normalizes feature covariance and $T \in \mathbb{R}^{d\times C}$ is the normalized empirical feature–label cross-covariance. We fit $\Theta \in \mathbb{R}^{d\times n}$ to predict all code targets jointly. By completing the square, we transform

$$\min_{\mathrm{rank}(\Theta)\le r} N^{-1}\|\mathbf{Y}M - \mathbf{X}\Theta\|_F^2 + \lambda\|\Theta\|_F^2$$

into minimizing $\|S_N^{-1}\Theta - TM\|_F^2$ up to an additive constant. The invertibility of S_N makes this a rank-preserving change of variables. We can therefore apply the best low-rank approximation theorem of Eckart & Young (1936) to obtain a global solution and its reconstructed coefficient:

$$\widehat{\Theta}_{M,r} = S_N(TM)_r, \qquad \widehat{B}_{M,r} = S_N(TM)_r M^\dagger \tag{21}$$

The notation $(TM)_r$ means a best rank-r approximation in Frobenius norm, and $M^\dagger$ is the Moore–Penrose pseudoinverse. The reconstructed coefficient $\widehat{B}_{M,r} \in \mathbb{R}^{d\times C}$ maps a feature vector, centered using the fitting-sample mean, to class scores. If $\widehat{\pi}$ is the empirical class-prior vector, centered reconstruction has intercept $H_C\widehat{\pi}$. We recover the equivalent prediction rule with intercept $\widehat{\pi}$ by adding the same constant $1/C$ to every class score.

For $\lambda = 0$, we use the same derivation whenever $\widehat{\Sigma}$ is positive definite. If it is singular, we restrict the fit to its positive-eigenvalue subspace and use $S_N = (\widehat{\Sigma}^\dagger)^{1/2}$. The empirical cross-moment $\mathbf{X}^\top\mathbf{Y}/N$ lies in that subspace, where the normalization is invertible. We therefore retain the displayed formulas by setting coefficient components in $\ker(\mathbf{X})$ to zero. This choice gives the minimum-norm coefficient for a given fitted response; spectral ties can still produce multiple optimal responses. In the implementation, we use an eigenvalue floor 10^{-12} solely to handle numerical singularity. All reported real-data penalties are positive.

I.3 Proof of Proposition 5

For this proposition, we assume $n = C - 1$ and full column rank, so $M^\top M$ is positive definite. We define $Q_M = M(M^\top M)^{-1/2}$ and obtain

$$Q_M^\top Q_M = I_{C-1}, \qquad Q_M Q_M^\top = M(M^\top M)^{-1}M^\top = H_C.$$

For the second identity, we recognize the middle matrix as the orthogonal projector onto $\mathrm{range}(M)$. Balance and rank $C-1$ give $\mathrm{range}(M) = \mathbf{1}^{\perp}$. For any centered class-score vector g, we have $H_C g = g$, and therefore

$$\|Q_M^{\top}(e_Y - g)\|_2^2 = (e_Y - g)^{\top} H_C (e_Y - g) = \|H_C e_Y - g\|_2^2.$$

Because this equality holds pointwise, we obtain the same loss on every dataset and the same set of minimizing score functions within a fixed centered function class.

To include the rank constraint and ridge term, we use the bijection $\Theta = BQ_M$ and $B = \Theta Q_M^{\top}$ between code coefficients and centered class-score coefficients $B \in \mathbb{R}^{d\times C}$ with $B\mathbf{1} = 0$. The identities above imply $\mathrm{rank}(\Theta) = \mathrm{rank}(B)$ and $\|\Theta\|_F = \|B\|_F$. Since $\mathbf{Y}H_C = \mathbf{Y}$, we also have

$$N^{-1}\|\mathbf{Y}Q_M - \mathbf{X}\Theta\|_F^2 + \lambda\|\Theta\|_F^2 = N^{-1}\|\mathbf{Y} - \mathbf{X}B\|_F^2 + \lambda\|B\|_F^2.$$

We have therefore matched the entire rank-constrained objective with centered one-hot RRR through a change of coordinates. For any common penalty $\lambda \geq 0$, the reconstructed solution sets agree, including at spectral ties.

More generally, if $r \geq C-1$, then TM has rank at most $C-1$ and $T\mathbf{1} = 0$, so $\widehat{B}_{M,r} = S_N TMM^{\dagger} = S_N T$ for every code with full centered span. We establish these identities at a fixed sample and penalty; different tuning decisions or fitting data can still yield different solutions. Accordingly, we compare coefficients at the same sample, penalty, and rank in our checks.

I.4 Loss metrics can realize any informative subspace with real codes

Finally, we ask which feature directions a training metric can select when we allow real-valued code entries. We retain the posterior matrix $A \in \mathbb{R}^{d\times C}$ but relax the binary-code constraint. Suppose A has full row rank and $A\mathbf{1} = 0$. We prescribe a positive definite matrix $S \in \mathbb{R}^{d\times d}$ whose leading eigenvectors give the desired feature directions, and set

$$K = A^{\dagger}S(A^{\dagger})^{\top} + \rho_K(H_C - A^{\dagger}A), \qquad \rho_K > 0.$$

Here, $A^{\dagger} \in \mathbb{R}^{C\times d}$ is the pseudoinverse, ρ_K is a positive scale, and $K \in \mathbb{R}^{C\times C}$ is the proposed coding metric. The two summands act on orthogonal subspaces of $\mathbf{1}^{\perp}$. Thus K is positive definite on that space and satisfies $AKA^{\top} = S$. We factor $K = MM^{\top}$ to obtain a real-valued centered code of rank $C-1$. By prescribing the leading eigenspace of S, we can realize any desired rank-r feature projector. An arbitrary K constructed this way need not admit a binary-code factorization. This distinction explains why a finite binary candidate pool can exclude a better decision subspace.

J Expressivity does not imply a training ordering

We now make explicit the distinction between the class-score functions a code can represent and the functions its training loss selects. Let $\phi(u) \in \mathbb{R}^k$ be a shared representation from a fixed encoder class and let $W_i \in \mathbb{R}^{k\times n_i}$ be an unrestricted linear readout for code $M_i \in \{-1,+1\}^{C\times n_i}$. Its coded scores are $f_i(u) = W_i^{\top}\phi(u)$.

For the reconstruction decoder in Equation (2), we can represent every centered class-score coefficient $B \in \mathbb{R}^{k\times C}$ with $B\mathbf{1}_C = 0$ by choosing $W_i = BM_i$:

$$W_i M_i^{\dagger} = BM_i M_i^{\dagger} = BH_C = B.$$

Conversely, the coefficient $B = W_i M_i^{\dagger}$ is centered and has rank at most $\mathrm{rank}(W_i)$. These two directions show that every code with full centered span represents the same centered class-score functions under a common rank limit.

We obtain the same expressivity conclusion for soft Euclidean nearest-codeword decoding. Every row of a binary code with n bits has squared norm n, so this decoder predicts $\arg\max_c [Mf]_c$. If $M_1 = M_2 J_{21}$ with $J_{21} \in \mathbb{R}^{n_2\times n_1}$, we can reproduce the scores $M_1 f_1(u)$ by choosing $f_2(u) = J_{21}W_1^{\top}\phi(u)$. This change of readout does not increase the representation dimension. When both code column spaces equal $\mathbf{1}_C^{\perp}$, we can apply the construction in both directions.

These reparameterizations establish equal expressivity, but they generally change the squared loss and readout regularization. We therefore cannot infer an ordering of fitted classifiers from expressivity alone. Our main theorems use the reconstruction decoder in Equation (2); we do not claim the same risk identity for uncalibrated Euclidean or hard Hamming decoding.

K FRESH-TRAINING REPLICATION PROTOCOL AND COMPLETE RESULTS

We repeat the image-to-class pipeline on new training draws while keeping the image encoder fixed. For each draw, we fit only the low-rank prediction coefficients, intercepts, and subsequent decoders. We define a candidate as one code matrix paired with a ridge penalty; candidate selection therefore does not require retraining the image encoder. We fixed the local protocol in `cohesion_protocol.md` before extracting the new features or observing their outcomes. Its SHA-256 is `d683b12b9ceae31f8f63c95c81f1e8d71ce90caa1b5d3fb85bd242dacb89e2c6`. We document numerical implementation details in the accompanying execution record. The protocol is a prospective local record, not an externally registered preregistration.

K.1 SAMPLING, FEATURES, AND UNCERTAINTY

For Fashion-MNIST (Xiao et al., 2017), we exclude all 3,000 training rows used in the earlier confirmation, leaving 57,000 eligible rows from the official training partition. With seeds 27100–27104, we independently draw 600 examples per class; the first 100 and 300 examples in each sampled class define nested smaller budgets. The union of rows requiring features contains 24,268 examples. For CIFAR-100 fine labels (Krizhevsky, 2009), we exclude all 4,000 training rows used in the earlier coarse-label study, leaving 46,000 eligible official training rows. With seeds 27200–27204, we draw 60 examples per class and use the first 20 per class for the smaller budget. This feature union contains 23,149 rows. Independent draws from a fixed eligible pool may overlap; the nested budgets within each draw are dependent.

We extract features with the pinned `microsoft/resnet-18` model (He et al., 2016) at revision `65a5785d9156231087c481e0c7dd33a5ff6f7e3e`, using its published processor and 512-dimensional pooled output. We repeat each Fashion image's grayscale values across three channels and keep all encoder weights fixed. For both tasks, we evaluate on all 10,000 official test rows. We reuse the Fashion test features exactly and newly extract the CIFAR test features, although some CIFAR images were previously evaluated with coarse labels. Neither task therefore supplies a new independent test dataset. We record source/model and feature hashes in the provenance file. We also saved source-row indices and per-draw mappings with the original feature arrays; those arrays are absent from the present archive.

We summarize the five training draws by their mean and one standard error (SE). For a given comparator, let D_i denote its test error minus the selected-code error, in percentage points, on draw $i \in \{1, \ldots, 5\}$. We denote the mean of these paired differences by $\bar{D}$ and their sample standard deviation by s_D, and report the 95% confidence interval (CI)

$$\bar{D} \pm t_{4,0.975}\, s_D/\sqrt{5}.$$

Here $t_{4,0.975}$ is the 97.5th percentile of the Student-t distribution with four degrees of freedom. This conventional approximate Student-t interval uses only five independent draws from the fixed eligible pool and conditions on a fixed test array. Its coverage is not distribution-free, and it excludes dataset-level and test-sampling uncertainty. We report every paired observation and make no multiplicity-adjusted confirmatory claim for the secondary budgets or CIFAR analyses. Our primary contrast compares the selected code with one-hot regression on the same fitting labels under a common decoder on Fashion at 3,000 labels. The CIFAR comparison changes both dataset and rank, so it assesses transfer without isolating class count as a causal factor.

K.2 LEARNERS AND SELECTION ORDER

We use stratified split seeds 27300+draw for the 80/20 development/validation split and 27400+draw to divide the development set into fitting and pilot subsets. We estimate standardization parameters from the fitting rows alone and share them across methods. We use rank four for Fashion and rank sixteen for CIFAR, and search the penalty grid $10^{-4}, 10^{-2}, 1$ in that order. We generate the 32-code pools with seeds 18511 and 27511. Each matrix has C class rows and $n = C - 1$ binary-task columns, where C denotes the number of classes. We obtain balanced binary codes with full centered span by permuting the class rows of one base code, and fix this pool design without using test outcomes.

For each penalty, we fit coded RRR candidates on the final-fitting examples and fit a separate set on the full development set. To score the final-fitting candidates, we estimate class means and a pooled Ledoit–Wolf covariance (Ledoit & Wolf, 2004) from the pilot examples. Pooling assumes a shared within-class covariance, while shrinkage stabilizes its estimate. We then estimate each candidate's risk under this Gaussian model using 8,192 common draws (seed 28110+draw). For the fixed-random comparator, we set the code index to the draw index modulo 32. Validation code selection minimizes raw validation error over the appropriate code pool. We also use validation to choose each method's penalty, breaking ties by grid order.

After selection, we fit common Gaussian decoders on the selected representations using each method's original fitting rows, without retuning codes or penalties. Projected linear discriminant analysis (LDA) fits a Gaussian class decoder in the one-hot RRR subspace and selects its penalty separately using its own decoder.

For rank-constrained softmax, we retain the earlier study's penalty on the coefficient-matrix product. We use a one-hot initialization and a random initialization (seed 28900+draw), allowing at most 5,000 L-BFGS-B iterations per start (Byrd et al., 1995), and choose the start with the lower training objective. We retain all capped fits. With centered fitting features $\mathbf{X} \in \mathbb{R}^{N \times d}$, we factor the slope scores as $(\mathbf{X}V)F$, where N is the number of fitting examples, d is the feature dimension, $V \in \mathbb{R}^{d \times r}$, and $F \in \mathbb{R}^{r \times C}$. Thus, $\mathbf{X}V$ contains the r learned feature combinations, and F maps them to class scores; we fit the intercept separately. This factorization avoids redundant multiplication by a dense class coefficient. We verify its equivalence to the dense formulation by separately comparing the objectives, analytic gradients, and finite-difference derivatives. The computational change therefore does not introduce a new learner. We record all 150 starts and every selected model.

Table 2: All prescribed methods for Fashion-MNIST / ResNet-18: test error (%), mean ± one SE over training draws. Gaussian decoders reuse the representations and penalties selected with the reconstruction decoder. Rows without an explicit decoder use their original prediction rule.

Method	1,000 labels	3,000 labels	6,000 labels
Random code, reconstruction	30.21 ± 0.68	27.50 ± 1.35	25.80 ± 1.35
Random code, Gaussian	25.68 ± 0.98	23.78 ± 0.69	21.03 ± 0.64
Selected code, reconstruction	24.65 ± 0.32	23.51 ± 0.57	21.43 ± 0.25
Selected code, Gaussian	21.92 ± 0.25	20.96 ± 0.43	18.40 ± 0.10
Validation code, same fit	25.46 ± 0.59	23.51 ± 0.57	21.43 ± 0.25
Validation code, same-fit Gaussian	22.54 ± 0.44	20.96 ± 0.43	18.40 ± 0.10
Validation code, full budget	24.27 ± 0.42	22.75 ± 0.13	21.21 ± 0.39
Validation code, full-budget Gaussian	21.55 ± 0.43	20.03 ± 0.35	18.39 ± 0.35
One-hot, same fit	43.50 ± 1.26	44.44 ± 0.94	45.06 ± 1.06
One-hot, same-fit Gaussian	31.95 ± 0.67	32.35 ± 0.65	32.02 ± 0.72
One-hot, full budget	44.38 ± 1.29	44.83 ± 1.17	44.55 ± 1.20
One-hot, full-budget Gaussian	33.16 ± 1.45	32.12 ± 0.78	31.72 ± 0.39
Projected LDA, full budget	32.31 ± 0.83	31.37 ± 0.48	31.72 ± 0.39
Softmax, full budget	21.35 ± 0.50	17.65 ± 0.26	16.43 ± 0.12

Table 3: All prescribed methods for CIFAR-100 / ResNet-18: test error (%), mean ± one SE over training draws. Gaussian decoders reuse the representations and penalties selected with the reconstruction decoder. Rows without an explicit decoder use their original prediction rule.

Method	2,000 labels	6,000 labels
Random code, reconstruction	73.61 ± 0.35	71.43 ± 0.66
Random code, Gaussian	65.30 ± 0.28	59.73 ± 0.95
Selected code, reconstruction	72.55 ± 0.47	69.62 ± 0.53
Selected code, Gaussian	64.72 ± 0.18	58.79 ± 0.87
Validation code, same fit	72.97 ± 0.65	69.63 ± 0.46
Validation code, same-fit Gaussian	65.15 ± 0.32	58.20 ± 0.13
Validation code, full budget	71.57 ± 0.56	69.72 ± 0.44
Validation code, full-budget Gaussian	62.24 ± 0.23	56.76 ± 0.23
One-hot, same fit	76.09 ± 0.50	75.61 ± 0.39
One-hot, same-fit Gaussian	63.28 ± 0.27	59.96 ± 0.11
One-hot, full budget	77.15 ± 0.50	75.57 ± 0.43
One-hot, full-budget Gaussian	62.35 ± 0.26	59.33 ± 0.24
Projected LDA, full budget	62.35 ± 0.26	59.33 ± 0.24
Softmax, full budget	57.52 ± 0.07	49.50 ± 0.11

Table 4: Paired same-fit one-hot minus selected-code error under the common Gaussian decoder, in percentage points. All five training-draw effects are shown. Intervals are approximate paired Student-t 95% intervals conditional on the eligible source pool and fixed test set.

Task	Labels	Draw 1	Draw 2	Draw 3	Draw 4	Draw 5	Mean [95% CI]
Fashion-MNIST	1,000	8.89	8.99	11.73	9.12	11.43	10.03 [8.27, 11.79]
Fashion-MNIST	3,000	9.70	11.97	11.46	11.96	11.86	11.39 [10.19, 12.59]
Fashion-MNIST	6,000	11.97	12.84	14.14	13.44	15.75	13.63 [11.85, 15.40]
CIFAR-100	2,000	-0.97	-0.76	-2.28	-1.94	-1.26	-1.44 [-2.24, -0.64]
CIFAR-100	6,000	2.25	2.06	-2.30	1.52	2.31	1.17 [-1.27, 3.61]

K.3 Verification, convergence, and computation

In the replay audit, we check all 350 selected model/decoder variants against saved predictions and score ranks, together with all 4,800 candidate test evaluations, all 25 split partitions, and 150 whitening/full-rank controls. We verify that every new training draw excludes previously used training rows and that feature/source hashes and nested budget mappings agree with the recorded artifacts. These checks establish artifact consistency; they do not establish the correctness of every modeling assumption.

All 25 validation-selected softmax fits satisfy a solver stopping criterion. Among the 75 per-penalty fits chosen by training objective, 65 satisfy such a criterion, as do 113 of all 150 starts. We retain capped or otherwise nonconverged fits in the comparisons. Solver convergence does not certify global optimality. We perform the dense-versus-factored objective and gradient check before fitting and record it separately.

We independently time complete searches on draw zero at 6,000 labels, after feature extraction and all other predictor fitting have finished. Each timed search reproduces the saved penalty, code where applicable, and test predictions. Before timing, we load the arrays and standardize features using the shared fitting subset. Each search includes the full three-penalty grid, validation, pilot estimation where applicable, and common decoder fitting when reported. We exclude encoder extraction and test prediction from these predictor-selection times. Extracting the new Fashion training features takes 422.2 seconds, with existing test features reused; extracting CIFAR training and test features takes 608.7 seconds. We use four CPU threads for extraction and one BLAS/OpenMP thread for predictor timing. These measurements describe one search per task and method; they do not establish an advantage across hardware or under equal computation budgets.

Table 5: Complete independent search time (seconds), at 6,000 labels on draw zero. Gaussian-decoder fitting is included for the first three rows.

Method	Fashion, rank 4	CIFAR-100, rank 16
Selected code	0.54	1.51
Validation code, full budget	0.42	0.84
One-hot, full budget	0.41	0.51
Softmax, full budget	125.58	357.93

L Prespecified confirmation: design, uncertainty, and complete results

We test whether changing the code changes which useful directions a low-rank linear predictor retains from fixed image features. We first select the coded representation using the reconstruction decoder, then fit a common Gaussian decoder to compare representations under the same decoding rule. We use *raw* for each method's original decoder: reconstruction for coded and one-hot RRR, and the prescribed Gaussian or class-score decoder for projected LDA and softmax, respectively. For the *Gaussian* control, we fix the original representation and penalty before fitting the common Gaussian decoder. All methods produce multiclass predictions, although coded fitting uses binary target vectors internally.

L.1 Protocol, datasets, and feature extraction

We fixed `confirmation_protocol.md` before extracting the new features or inspecting their test outcomes, and retain its SHA-256 in the provenance and execution record. This is a prospective local protocol, not an externally registered preregistration. We designed the confirmation in response to the earlier DINOv2/CIFAR study and its exploratory decoder ablation, and retain all prescribed tasks, ranks, methods, and unfavorable outcomes.

We use Fashion-MNIST (Xiao et al., 2017) as the new dataset. With seed 2026091901, we select 300 examples per class from its official 60,000-image training partition and evaluate on all 10,000 official test images in source order. We extract ImageNet-pretrained ResNet-18 (He et al., 2016)

Table 6: Prespecified confirmation at rank four: test error (%), mean ± one SE over five splits. All tasks use 3,000 training images and the complete 10,000-image official test set. F-R/F-D denote Fashion-MNIST with ResNet-18/DINOv2; C-R denotes CIFAR-10 with ResNet-18.

Method	F-R	F-D	C-R
Selected code, reconstruction	22.94 ± 0.42	19.61 ± 0.21	29.88 ± 0.54
Selected code, Gaussian decoder	20.42 ± 0.13	16.27 ± 0.09	26.23 ± 0.28
One-hot, same-fit Gaussian decoder	32.24 ± 0.30	30.04 ± 0.39	35.76 ± 0.35
Validation code, full-budget Gaussian	20.70 ± 0.48	15.62 ± 0.07	26.18 ± 0.11
Projected LDA, full budget	32.29 ± 0.13	30.02 ± 0.13	35.22 ± 0.22
Softmax, full budget	17.45 ± 0.06	13.91 ± 0.07	21.58 ± 0.13

global-average-pooled features (512 dimensions) and DINOv2-small (Oquab et al., 2024) final normalized classification-token (CLS) features (384 dimensions). Before applying the published image processor, we repeat grayscale values across three channels. We keep the encoders frozen, with no task fine-tuning or additional feature normalization.

For the secondary CIFAR-10/ResNet study (Krizhevsky, 2009), we reuse the same 3,000 training source rows as in the earlier CIFAR-10/DINOv2 study and evaluate on the complete 10,000-image official test set. Some CIFAR test images had already been examined in that study; Fashion supplies the confirmation on a new dataset. Both studies use the full test partition and a training subset, rather than the full training partition.

We pin `microsoft/resnet-18` to revision

`65a5785d9156231087c481e0c7dd33a5ff6f7e3e` and `facebook/dinov2-small` to revision

`ed25f3a31f01632728cabb09d1542f84ab7b0056`.

We record publisher Parquet sources, model/processor hashes, and feature hashes in the provenance files. We saved source-row indices with the original feature arrays, which are absent from the present archive. We extract features in evaluation mode using CPU PyTorch 2.8.0 and four threads, and retain the extraction scripts in the source package.

L.2 Splits, estimators, and budgets

For each task, we use seeds 19300–19304 for the development/validation splits and seeds 19500–19504 for the fit/pilot subdivision. Each split contains 1,800 fitting, 600 pilot, and 600 validation examples, stratified equally by class. We use the fitting examples to estimate low-rank coefficients, the pilot examples to estimate the class-distribution model that scores codes, and the validation examples to select hyperparameters. We measure test error only after fixing all choices. We estimate standardization parameters from fitting features alone and share them across methods. For the full-budget baselines, we combine the 1,800 fitting and 600 pilot labels for training and retain the same 600 validation labels. Every method selects a penalty from $10^{-4}, 10^{-2}, 1$ by validation error, with ties broken by grid order. We fixed ranks $2, 4, 9$ in advance and excluded rank one because the two-label limitation of the raw squared-loss decoder with a common intercept was already known.

For every task, we use the same pool of 32 balanced binary code matrices with full centered span, generated by class-row permutations with seed 18511. With $C = 10$ classes, each matrix has C rows and $n = C - 1 = 9$ binary-task columns. For each penalty, the Gaussian and mixture methods first select a code using the pilot risk model; validation then selects among the resulting three code/penalty pairs. Validation search on the same fitting labels directly compares all 96 candidates trained on 1,800 labels. For the full-budget validation-code comparator, we instead train these candidates on all 2,400 development labels and select them using the same validation set. We set the random code index to the split index modulo 32. For width scoring, we use 8,192 common standard-normal draws in the pilot score span.

We fit the class-conditional Gaussian model using pilot class means and a pooled Ledoit–Wolf covariance (Ledoit & Wolf, 2004). The pilot variant scores pilot-fitted classifiers, whereas the final-fit

variant scores the actual final-fitting candidates under the pilot distribution. The latter removes classifier-transfer error from that scoring step. For both variants, we use 8,192 common integration draws and Rao–Blackwellized loss: at each simulated feature, we average the conditional error probability rather than sample a class label.

For the mixture model, we first project pilot features onto the common $C - 1$ score span: the feature subspace needed to evaluate all centered linear class scores for the given fitting geometry. We then whiten these features by their pooled covariance and fit two diagonal Gaussian components per class. We fix two initializations, 200 iterations, and covariance regularization 0.05. To estimate risk, we use 8,192 common mixture draws and average conditional loss using the exact fitted-mixture posterior probabilities. We choose neither the component count nor the regularization using test outcomes. This comparison assesses the adequacy of the risk model; it does not test whether pretrained features satisfy the theorem's convex-distance assumption.

For nine methods, we also fix the representation and penalty selected with the raw decoder, then fit the same pooled Gaussian decoder using each method's original fitting labels. We do not reselect the representation or penalty after changing the decoder. This prescribed control examines the mechanism: the one-hot comparison uses the same fitting labels, while the full-budget comparators assess practical label use. Projected LDA uses its own specified Gaussian decoder.

For rank-constrained softmax, we use the existing product-ridge objective, two starts selected by training objective, and a uniform 5,000-iteration cap. All 45 selected softmax fits satisfy a solver stopping criterion, with a maximum reported gradient infinity norm of 9.50×10^{-5}. Of the 135 best-start penalty fits, 129 satisfy such a criterion. We retain the six remaining fits, none of which validation selects. All 1,350 class-mixture fits converge within at most 38 expectation-maximization (EM) iterations. Solver termination does not certify a global optimum.

L.3 Uncertainty and primary comparisons

Our two primary contrasts compare representations selected by the Gaussian final-fit risk criterion with one-hot representations fitted on the same labels, using the common Gaussian decoder at rank four. We make this comparison separately for the two Fashion encoders. We average test error over five fitted models and report one standard error across data splits, conditional on the common training and test arrays.

For paired intervals, we first average each test example's loss difference across the five fitted models. We then resample test images within class 5,000 times while preserving the official class proportions. We assign each of the two primary intervals a nominal coverage of 97.5%, allocating the family error by Bonferroni. The resulting percentile bootstrap intervals have approximate coverage; Bonferroni allocation does not make their marginal bootstrap coverage exact.

These intervals condition on the fitted models and estimate test-sampling uncertainty. They exclude uncertainty from drawing a new training dataset. Because the two encoders use the same Fashion examples, their results do not constitute independent dataset replications. We report all other displayed intervals as pointwise 95% descriptive intervals. We do not tune ranks, candidates, or methods using these test outcomes.

Table 7: Paired test-set bootstrap: same-fit one-hot error minus selected-code error under the common Gaussian decoder. Positive values favor coding. Intervals condition on the five fitted models; the two Fashion rank-four intervals use a Bonferroni allocation.

Task	Advantage (pp)	Interval	Level
Fashion-MNIST / ResNet-18	11.82	[11.03, 12.63]	97.5%
Fashion-MNIST / DINOv2	13.77	[12.90, 14.62]	97.5%
CIFAR-10 / ResNet-18	9.54	[8.88, 10.20]	95%

We provide results for every method and rank, together with all scoring diagnostics, audits, and timings, in Supplementary Section S9.