THE OPTIMISATION LANDSCAPE OF MCR^2

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding how neural networks organise learned representations is important for improving their generalisation, robustness, interpretability, and ultimately safety. The recent proposed framework Maximal Coding Rate Reduction (MCR^2) offers an information-theoretic perspective: it aims to capture the underlying structure of each class while arranging different classes in orthogonal subspaces. Despite promising applications, the optimisation landscape of MCR^2 remains poorly understood. Due to the requirement of every feature should have the same scale of information, the landscape of MCR^2 is difficult to analyse. Previous work has therefore studied a relaxed Frobenius-regularised variant. We show that this variant can perform poorly in practice, limiting the practical guidance offered by its landscape analysis. In this work, we characterise key features of local and global optima of MCR^2 itself and analyse other critical points. In particular, every local optimum whose occupied class ranks sum to at most the representation dimension separates the classes orthogonally. We also describe global solutions in several important settings, show how classes may overlap when feature space is limited, and identify configurations where local optimisation can stall or second-order tests are inconclusive. These results explain when MCR^2 favours discriminative representations under first-order optimisation and where its guarantees stop. Extensive experiments with synthetic data and real image datasets support the theoretical analysis.

1 INTRODUCTION

Deep learning has transformed natural language processing, biology, and robotics (Gemini Team Google et al., 2024; Abramson et al., 2024; Kaufmann et al., 2023). Yet choices about network depth, width, and training still rely heavily on trial and error, making model behaviour difficult to explain and formal guarantees hard to establish. Specifically, neural networks can be sensitive to simple geometric transformations (Azulay & Weiss, 2018; Engstrom et al., 2019), memorise arbitrary labels (Zhang et al., 2017), be vulnerable to adversarial perturbations (Szegedy et al., 2013), and forget earlier tasks during sequential training (McCloskey & Cohen, 1989; De Lange et al., 2022; Wu et al., 2021). Even the relative merits of multilayer perceptrons, convolutional neural networks, and Transformers remain debated (Tay et al., 2021; Xu et al., 2023). These issues call for a mathematical framework that explains network behaviour and guides design beyond empirical "alchemy" (Ma et al., 2022).

Theory has examined overparameterisation (Arora et al., 2019), optimisation dynamics (Gidel et al., 2019), expressive power (Rolnick & Tegmark, 2017; Hanin, 2017), and generalisation (Bartlett et al., 2017; Golowich et al., 2017). Yet many results rely on idealised settings, such as shallow networks (Zhang et al., 2019; Mei & Montanari, 2019) or infinite-width limits (Allen-Zhu et al., 2019; Buchanan et al., 2020), or focus on individual mechanisms such as dropout (Cavazza et al., 2018). With limited guidance for designing complete systems, model development continues to draw on neural architecture search (Zoph & Le, 2017; Baker et al., 2017), automated machine learning (Hutter et al., 2019), and meta-learning (Andrychowicz et al., 2016). This leaves a more fundamental question: what structure should a useful learned representation have?

The recently proposed *maximal coding rate reduction* (MCR^2) framework offers an information-theoretic answer (Yu et al., 2020; Chan et al., 2022; Yu et al., 2023; 2024). It starts from the premise that real-world samples within a class often share a low-dimensional submanifold. MCR^2 seeks to preserve this structure while separating the classes by maximising the difference between the cod-

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.

ing rate of all features and the average coding rate within classes. The resulting geometric principle guides representation learning and can be translated into interpretable deep architectures by unrolling the objective's optimisation into network layers (Chan et al., 2022). More recently, Buchanan et al. (2026) have placed these principles within a broader mathematical theory of memory.

This geometric perspective has supported supervised, self-supervised, and unsupervised learning, with reported robustness to corrupted labels (Yu et al., 2020). It has also inspired more efficient variational formulations (Baek et al., 2022) and white-box transformers based on sparse rate reduction (Yu et al., 2023; 2024). Coding rate principles have been explored in interpretable medical imaging (Zhou et al., 2024). Reliable representations are likewise important in high-stakes settings such as wearable physiological time-series analysis (Wang et al., 2026).

Despite these promising applications, the optimisation landscape of MCR^2 remains poorly understood. Wang et al. (2024) analyse a Frobenius-regularised variant of MCR^2. However, we show that this variant can perform poorly in practice (Table 1). Its landscape analysis therefore offers limited guidance for understanding and using MCR^2. To address this gap, we study the MCR^2 objective directly and characterise its landscape in several important settings. We ask whether the class separation favoured by global optima also holds at local optima and what happens when the representation space is too small.

To the best of our knowledge, this work is the first systematic analysis of the optimisation landscape of MCR^2 itself. We use geometric perturbations with bounds on coding rates to establish the following results:

- **Orthogonal class separation under a rank budget.** Every local optimum whose occupied class ranks sum to at most the representation dimension separates the classes orthogonally. Global optima are also orthogonal when the individually preferred class ranks fit. We prove local separation by showing that a nonorthogonal critical point under the rank budget can be improved through a rotation that leaves each class's internal geometry unchanged. This makes class separation a favourable property of the MCR^2 landscape in the stated regimes.
- **Structure in limited dimension.** When the feature space is too small for all classes to be orthogonal, MCR^2 still has structured optima. With sufficiently balanced class sizes, every global optimum aligns each class along one direction and uses the available feature dimensions evenly. When classes are restricted to one direction, every local optimum is global. If classes retain fixed higher-dimensional structure, some overlap is unavoidable and obeys a sharp lower bound.
- **A remaining challenge for optimisation.** MCR^2 can have local optima that separate classes orthogonally but are not globally optimal. Our three-sample example shows that local optimisation may stop at such a representation even though a better one exists. How to reach a global optimum reliably therefore remains an open question.

Together, our results precisely characterise key properties of the MCR^2 landscape both when the representation space can accommodate orthogonal classes and when its dimension is limited. We also identify a potential obstacle in practice: optimisation can settle at a local optimum without reaching the best overall representation.

2 Landscape with sufficient representation dimension

Throughout, bold lowercase letters denote vectors and bold uppercase letters denote matrices. Scalar quantities, functions, sets, and subspaces use ordinary math type.

2.1 Preliminaries

Let $K \geq 2$ classes contain $m_k \geq 1$ samples, with $m = \sum_{k=1}^{K} m_k$, proportions $p_k = m_k/m$, and feature dimension $d \geq 2$. Write $\boldsymbol{Z} = [\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_K] \in \mathbb{R}^{d \times m}$ and require every column $\boldsymbol{z}_i$ to have unit Euclidean norm. For precision $\epsilon > 0$, set $\alpha = d/(m\epsilon^2)$, so $d/(m_k\epsilon^2) = \alpha/p_k$. The

optimisation objective of MCR^2 is

$$\max_{\|\boldsymbol{z}_i\|_2=1\,\forall i} F(\boldsymbol{Z}) = \frac{1}{2}\log\det\left(\boldsymbol{I}_d + \alpha \boldsymbol{Z}\boldsymbol{Z}^\top\right) - \sum_{k=1}^{K} \frac{p_k}{2}\log\det\left(\boldsymbol{I}_d + \frac{\alpha}{p_k}\boldsymbol{Z}_k\boldsymbol{Z}_k^\top\right). \tag{1}$$

When speaking of loss minimisation, we use $L(\boldsymbol{Z}) = -F(\boldsymbol{Z})$: a local minimiser of L is exactly a local maximiser of F. The unit-norm constraints define a compact feasible set, so global maximisers exist. As Yu et al. (2020) note, coding rate reduction is monotonic in feature scale, so they propose either fixing the total squared norm of each class, $\|\boldsymbol{Z}_k\|_F^2 = m_k$, or normalising each feature to unit length. Their released implementation code uses the latter choice. We follow this sample-wise constraint which also implies the former class-wise constraints. The class-wise constraints alone permit unequal sample norms and define a larger feasible set.

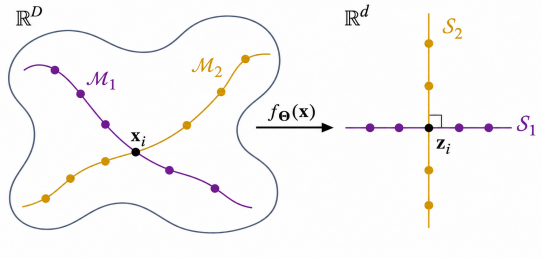


Figure 1: Geometric intuition for MCR^2. Samples from different class manifolds in the input space are mapped into mutually orthogonal class subspaces in the representation space.

Geometrically, MCR^2 seeks to map samples from each class into a class-specific subspace of the learned representation. When these subspaces fit within the feature space, the local and global optima studied below separate them into mutually orthogonal directions, as illustrated in Figure 1.

The Frobenius-regularised variant. Wang et al. (2024) analyse the following unconstrained, Frobenius-regularised variant of the MCR^2 objective:

$$\max_{\boldsymbol{Z}\in\mathbb{R}^{d\times m}} F_\lambda(\boldsymbol{Z}), \qquad F_\lambda(\boldsymbol{Z}) = F(\boldsymbol{Z}) - \frac{\lambda}{2}\|\boldsymbol{Z}\|_F^2, \qquad \lambda > 0. \tag{2}$$

Here F is the coding rate reduction in (1), evaluated without the unit-norm constraints. Individual feature norms are therefore free to vary, and the landscape results for F_λ concern a different optimisation problem. Following the experimental setup of Yu et al. (2020), we test this variant on CIFAR-10 dataset. Appendix I gives the full protocol and reporting details.

Table 1: CIFAR-10 test accuracy (%).

Objective	Accuracy
MCR^2 (Yu et al., 2020)	93.70
Frobenius-regularised variant	45.57

Table 1 shows that the Frobenius-regularised variant achieves only 45.57% test accuracy, compared with 93.70% for MCR^2. Figure 2 reveals a second weakness. MCR^2 produces clear class blocks in the feature Gram matrix, whereas the regularised variant does not. Its leading class subspaces also overlap, so the learned classes do not have the orthogonal structure sought by MCR^2. The poor classification result and the weak geometric separation together show that this relaxation has limited practical value. Consequently, analysing its optimisation landscape offers limited guidance for understanding and using MCR^2.

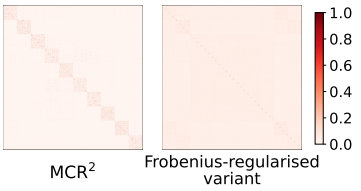


Figure 2: Entrywise Gram matrices $|\widetilde{\boldsymbol{Z}}^\top\widetilde{\boldsymbol{Z}}|$ of CIFAR-10 features sorted by class, where $\widetilde{\boldsymbol{Z}}$ has unit-norm columns.

2.2 Local optima inherit class orthogonality

Define $\boldsymbol{A} = \boldsymbol{I}_d + \alpha\boldsymbol{Z}\boldsymbol{Z}^\top$ and $\boldsymbol{B}_k = \boldsymbol{I}_d + (\alpha/p_k)\boldsymbol{Z}_k\boldsymbol{Z}_k^\top$. For a column $\boldsymbol{z}_i$ in class k, the Euclidean and Riemannian gradients are

$$\nabla_{\boldsymbol{z}_i} F = \alpha(\boldsymbol{A}^{-1} - \boldsymbol{B}_k^{-1})\boldsymbol{z}_i, \qquad \operatorname{grad}_{\boldsymbol{z}_i} F = (\boldsymbol{I}_d - \boldsymbol{z}_i\boldsymbol{z}_i^\top)\alpha(\boldsymbol{A}^{-1} - \boldsymbol{B}_k^{-1})\boldsymbol{z}_i. \tag{3}$$

Thus a critical point satisfies $(\boldsymbol{A}^{-1} - \boldsymbol{B}_k^{-1})\boldsymbol{z}_i = \mu_i \boldsymbol{z}_i$ for some scalar μ_i at every sample. In particular, Euclidean stationarity, which underlies the regularised analysis, is not the correct stationary condition here.

We first study when the class subspaces can fit without overlap. The local result requires the sum of the occupied class ranks to be at most the feature dimension. The global result is proved when the feature dimension is at least the sample count, and also applies in smaller spaces when the preferred class ranks fit.

Put $\boldsymbol{C}_k = \boldsymbol{Z}_k \boldsymbol{Z}_k^\top$, $\boldsymbol{S} = \sum_k \boldsymbol{C}_k = \boldsymbol{Z}\boldsymbol{Z}^\top$, and $r_k = \operatorname{rank}(\boldsymbol{Z}_k)$. The matrix $\boldsymbol{C}_k$ records the directional energy of class k: for any unit vector $\boldsymbol{w}$,

$$\boldsymbol{w}^\top \boldsymbol{C}_k \boldsymbol{w} = \|\boldsymbol{Z}_k^\top \boldsymbol{w}\|_2^2 = \sum_{i \text{ in class } k} (\boldsymbol{w}^\top \boldsymbol{z}_i)^2. \tag{4}$$

Thus $\boldsymbol{S}$ records the energy of all classes, and r_k counts the independent directions occupied by class k. These are uncentred second-moment matrices. Class orthogonality means that every feature of one class is perpendicular to every feature of another, namely $\boldsymbol{Z}_k^\top \boldsymbol{Z}_\ell = 0$.

The perturbation we use rotates an entire class by the same orthogonal matrix. It preserves every sample norm and every within-class inner product. Consequently all within-class coding rates are fixed, and only the total coding rate can change. The proof shows that, at a nonorthogonal stationary point under the dimension budget, some class can rotate part of its energy from a shared direction into an unused direction and strictly increase the total rate.

Theorem 1 (Strict saddle for nonorthogonal rank-deficient critical points)**.** *Let $\boldsymbol{Z}$ be a Riemannian critical point of (1). Suppose $\operatorname{rank}(\boldsymbol{Z}) < d$ and $\boldsymbol{Z}_k^\top \boldsymbol{Z}_\ell \neq 0$ for some $k \neq \ell$. Then there exist a class j, orthonormal vectors $\boldsymbol{u}, \boldsymbol{v}$, and numbers $0 < \mu < \lambda$ such that a plane rotation $\boldsymbol{Q}(t)$ of $\boldsymbol{u}$ toward $\boldsymbol{v}$ gives the feasible path $\boldsymbol{Z}_j(t) = \boldsymbol{Q}(t)\boldsymbol{Z}_j$ with all other classes fixed, and*

$$F(\boldsymbol{Z}(t)) - F(\boldsymbol{Z}) = \frac{1}{2} \log\left(1 + \frac{\alpha^2 \mu(\lambda - \mu)}{1 + \alpha\lambda} \sin^2 t\right). \tag{5}$$

In particular,

$$\frac{d^2}{dt^2} F(\boldsymbol{Z}(t))\Big|_{t=0} = \frac{\alpha^2 \mu(\lambda - \mu)}{1 + \alpha\lambda} > 0. \tag{6}$$

Every within-class Gram matrix $\boldsymbol{Z}_k(t)^\top \boldsymbol{Z}_k(t)$, and hence every sample norm and class rank, is constant along the path.

Corollary 1 (Local-optimum orthogonality under a total rank budget)**.** *Every local maximiser $\boldsymbol{Z}$ of (1) satisfying*

$$\sum_{k=1}^{K} \operatorname{rank}(\boldsymbol{Z}_k) \leq d \tag{7}$$

has $\boldsymbol{Z}_k^\top \boldsymbol{Z}_\ell = 0$ for all $k \neq \ell$. More generally, every nonorthogonal critical point satisfying (7) is a strict saddle.

The sufficient dimension condition is a sum of class ranks, not a sum of class sample counts. For example, K rank-r class representations need only satisfy $Kr \leq d$, even when each class contains many more than r samples. If prescribed class-rank bounds $r_k \leq d_k$ are part of the problem and $\sum_k d_k \leq d$, every local maximiser under those bounds is likewise class orthogonal: the rotation in Theorem 1 preserves the bounds. Thus the conclusion applies under the same type of ambient-dimension budget used for global class separation in Yu et al. (2020).

Proof roadmap. There are four steps. First, at a local optimum every infinitesimal class rotation has zero first derivative. This implies that $\boldsymbol{C}_k$ commutes with $\boldsymbol{S}$. Second, a rank argument using this commutation shows that a nonorthogonal stationary point under $\sum_k r_k \leq d$ leaves an unused ambient direction. Third, nonorthogonality supplies a shared direction with energy $\mu > 0$ from one class and energy $\lambda - \mu > 0$ from the others. Finally, rotating that class toward the unused direction gives the exact improvement in (5), contradicting local maximality. The coefficient is strictly positive for every $\alpha > 0$, so no additional precision threshold is needed. Appendix A proves each step and explains why the unused direction exists even when $\sum_k r_k = d$.

2.3 GLOBAL OPTIMA WHEN THE PREFERRED CLASS RANKS FIT

Write $r_k = \operatorname{rank}(\boldsymbol{Z}_k)$ and let $\boldsymbol{P}_k$ denote the orthogonal projector onto its column span. The next theorem reduces the global problem to K finite rank choices. A class with many samples can select a subspace of rank greater than one, while every class remains nonzero.

Theorem 2 (All global maximisers for $d \geq m$)**.** *Suppose $d \geq m$ and $K \geq 2$. For $1 \leq r \leq m_k$, define*

$$G_k(r) = \frac{r}{2}\left[\log\Big(1 + \frac{\alpha m_k}{r}\Big) - p_k \log\Big(1 + \frac{\alpha m_k}{p_k r}\Big)\right]. \tag{8}$$

The optimum value of (1) is $\sum_{k=1}^{K} \max_{1\leq r\leq m_k} G_k(r)$. A feasible $\boldsymbol{Z}$ is a global maximiser if and only if, for every k,

$$\boldsymbol{Z}_k \boldsymbol{Z}_k^\top = \frac{m_k}{r_k}\boldsymbol{P}_k, \qquad r_k \in \operatorname*{arg\,max}_{1\leq r\leq m_k} G_k(r), \qquad \boldsymbol{P}_k \boldsymbol{P}_\ell = 0 \quad (k \neq \ell). \tag{9}$$

Each $\boldsymbol{Z}_k$ is therefore a unit-norm tight frame for its class subspace.

Theorem A.6 of Yu et al. (2020) also allows one smaller singular value among equal larger ones. Its scalar lemma leaves this as a candidate. The pairwise second-variation argument in Appendix B excludes that candidate for the MCR^2 rate function. Schur–Horn ensures the spectral perturbation remains feasible with unit-norm columns.

The local-orthogonality conclusion when $d \geq m$ is an immediate, weaker consequence of Corollary 1, since $\sum_k r_k \leq m$. Appendix C also gives a direct Gram-matrix proof under the sharper condition $\sum_k r_k \leq d$.

The global-maximiser characterisation also holds beyond $d \geq m$ whenever the individually preferred class ranks fit into the ambient space. In detail, define G_k by (8) for $1 \leq r \leq \min\{m_k, d\}$. If one can select $r_k \in \arg\max_r G_k(r)$ with $\sum_k r_k \leq d$, the upper bound in the proof of Theorem 2 is attainable, and (9) again characterises every global maximiser. This gives an exact result for some $d < m$ problems.

3 LANDSCAPE WITH LIMITED REPRESENTATION DIMENSION

The sufficient-dimension theorem concerns the ranks occupied by a representation. If $d < \sum_k d_k$, prescribed rank *upper bounds* alone do not force overlap: an optimiser may choose smaller ranks. The unavoidable case $K > d$ is particularly informative, because every nonempty class has rank at least one, so the class subspaces cannot all be orthogonal. The question is then how much overlap the objective permits and how it distributes the total class energy. We first solve the full sample-wise problem in a balanced regime, then give a complete landscape theorem on the rank-one class family, and finally treat fixed higher-rank class geometry.

Theorem 3 (All global optima with sufficiently balanced class masses)**.** *Suppose $\max_k p_k \leq 1/d$. Then, for every $\alpha > 0$, the global value of the full objective (1) is*

$$F^* = \frac{d}{2}\log\Big(1 + \frac{\alpha m}{d}\Big) - \frac{1}{2}\log(1 + \alpha m). \tag{10}$$

A feasible $\boldsymbol{Z}$ is globally optimal if and only if

$$\operatorname{rank}(\boldsymbol{Z}_k) = 1 \text{ for every } k, \qquad \boldsymbol{Z}\boldsymbol{Z}^\top = \frac{m}{d}\boldsymbol{I}_d. \tag{11}$$

Equivalently, $\boldsymbol{Z}_k = \boldsymbol{u}_k \boldsymbol{s}_k^\top$ with $\|\boldsymbol{u}_k\| = 1$, $\boldsymbol{s}_k \in \{-1, 1\}^{m_k}$, and $\sum_k m_k \boldsymbol{u}_k \boldsymbol{u}_k^\top = (m/d)\boldsymbol{I}_d$.

This theorem imposes no rank-one constraint on the optimisation problem: rank one follows from equality conditions. Since every sample has unit norm, $\operatorname{tr} \boldsymbol{S} = m$ and $\operatorname{tr} \boldsymbol{C}_k = m_k$. Concavity of the log determinant makes the total coding rate largest when the d eigenvalues of $\boldsymbol{S}$ all equal m/d. Conversely, at fixed class trace m_k, the class coding rate is smallest when $\boldsymbol{C}_k$ has rank one. A global optimiser must attain both bounds at once. The condition $p_k \leq 1/d$ says precisely that the weight m_k of any rank-one class does not exceed the target eigenvalue m/d; it permits the class directions

to form a weighted tight frame satisfying (11). If one class were heavier than m/d, its rank-one contribution alone would preclude total isotropy, so the two bounds could not be attained together.

For $K > d$ the resulting class directions are necessarily nonorthogonal, but equal pairwise angles are not required. For equal class sizes, the balance condition is simply $K \geq d$. Global optimality need not even imply distinct class directions: with $d = 2$, four equally sized classes can take $\boldsymbol{u}_1 = \boldsymbol{u}_2 = \boldsymbol{e}_1$ and $\boldsymbol{u}_3 = \boldsymbol{u}_4 = \boldsymbol{e}_2$, and still satisfy (11). Thus dimension shortage can permit exact class collisions at a global optimum.

Theorem 4 (Complete landscape within the rank-one class family)**.** *Restrict to $\boldsymbol{Z}_k = \boldsymbol{u}_k \boldsymbol{s}_k^\top$ with fixed sign vectors $\boldsymbol{s}_k$ and unit class directions $\boldsymbol{u}_k$. Suppose $K > d$ and sort the class sizes as $w_1 \geq \cdots \geq w_K > 0$. Let $r \in \{0, \ldots, d-1\}$ and τ be determined by*

$$\tau = \frac{\sum_{k>r} w_k}{d-r}, \qquad w_r > \tau \geq w_{r+1}, \tag{12}$$

where the first inequality is omitted for $r = 0$. All global optima have mutually orthogonal directions for the r heavy classes, perpendicular to all remaining classes, and

$$\sum_{k>r} w_k \boldsymbol{u}_k \boldsymbol{u}_k^\top = \tau \boldsymbol{P}, \tag{13}$$

where $\boldsymbol{P}$ projects onto the remaining $d - r$ dimensions. Every local maximum on this family is global, and every nonglobal critical point has strictly positive curvature for F. The optimal total spectrum is $(w_1, \ldots, w_r, \tau, \ldots, \tau)$.

Within this rank-one family, each class coding rate is the same constant, $\frac{1}{2}\log(1 + \alpha m)$ after summing with the weights p_k. Optimising F therefore amounts to distributing the eigenvalues of $\boldsymbol{S} = \sum_k w_k \boldsymbol{u}_k \boldsymbol{u}_k^\top$ as evenly as the prescribed weights allow. A class with $w_k > \tau$ cannot fit into a direction whose total eigenvalue is τ, so it occupies its own axis. The lighter classes share the remaining axes as a weighted tight frame. For example, with $d = 2$ and weights $(4, 1, 1)$, the first class is isolated and the optimal spectrum is $(4, 2)$; with weights $(2, 2, 2)$, no class is isolated and the spectrum is $(3, 3)$.

This is a prescribed-norm frame problem, connected to classical convex frame potentials (Massey et al., 2013; 2016). The landscape statement is stronger than a description of global optima: at any other critical point, a redundant collection of class directions in a higher-eigenvalue subspace can rotate towards a lower-eigenvalue direction and increase F to second order. Appendix E gives a direct real-valued proof and the explicit curvature. The local conclusion here is for the rank-one family; it is not a classification of all critical points of the unrestricted multiclass problem. For singleton classes the family equals the full feasible set, so the theorem proves a complete landscape result even when $d < K = m$.

Proposition 1 (Unavoidable overlap and optimal fixed-class orientations)**.** *For any feasible $\boldsymbol{Z}$,*

$$\sum_{k \neq \ell} \operatorname{tr}(\boldsymbol{C}_k \boldsymbol{C}_\ell) \geq \frac{m^2}{d} - \sum_k \operatorname{tr}(\boldsymbol{C}_k^2), \tag{14}$$

with equality if and only if $\boldsymbol{S} = (m/d)\boldsymbol{I}_d$. If within-class Gram matrices are fixed and such an isotropic orientation is feasible, all global orientation optima are exactly the isotropic ones. In particular, for equal class sizes n and fixed tight class frames $\boldsymbol{C}_k = (n/r)\boldsymbol{P}_k$,

$$\frac{1}{K(K-1)r} \sum_{k \neq \ell} \operatorname{tr}(\boldsymbol{P}_k \boldsymbol{P}_\ell) \geq \frac{Kr - d}{d(K-1)}. \tag{15}$$

When $Kr > d$, the positive bound is attained precisely by a tight fusion frame $\sum_k \boldsymbol{P}_k = (Kr/d)\boldsymbol{I}_d$.

The overlap measure has a direct geometric meaning: $\operatorname{tr}(\boldsymbol{C}_k \boldsymbol{C}_\ell) = \|\boldsymbol{Z}_k^\top \boldsymbol{Z}_\ell\|_F^2$. Expanding $\operatorname{tr}(\boldsymbol{S}^2)$ into within-class and cross-class terms gives (14); equality requires all eigenvalues of $\boldsymbol{S}$ to be equal. When the within-class Gram matrices are fixed, their coding-rate terms are constant, so an attainable isotropic orientation also maximises the full objective over these orientations.

Thus, when fixed multidimensional classes cannot fit orthogonally, total isotropy provides an exact global target whenever feasible. For example, take $K = d$ and let $\boldsymbol{P}_k$ select r consecutive coordinate axes cyclically. Then $\sum_k \boldsymbol{P}_k = r\boldsymbol{I}_d$, so equality is attainable for any integer $1 < r < d$, although $Kr > d$. This orientation result does not optimise the within-class spectra, and general higher-rank local classification and unrestricted imbalanced global classification are not claimed.

4 A REMAINING CHALLENGE FOR OPTIMISATION

Class orthogonality describes how different classes are arranged, but does not determine the value of every local optimum. A three-sample model exhibits a suboptimal local maximum and a degenerate saddle at different coding precisions. The first phenomenon shows that local optimisation can settle on a class-separated representation without reaching the global value.

Proposition 2 (A suboptimal local maximum)**.** *Take $d = m = 3$ and class sizes $(2, 1)$. For every $\alpha > 0$, the configuration $\boldsymbol{Z}_{\mathrm{col}} = [\boldsymbol{e}_1, \boldsymbol{e}_1, \boldsymbol{e}_2]$ is a local maximiser of (1). At $\alpha = 2$, it is strictly suboptimal: $\boldsymbol{Z}_{\mathrm{orth}} = [\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3]$ is globally optimal, and*

$$F(\boldsymbol{Z}_{\mathrm{orth}}) - F(\boldsymbol{Z}_{\mathrm{col}}) = \frac{1}{6} \log \frac{35721}{32000} > 0. \tag{16}$$

The proof in Appendix F works directly with the within-class correlation and the determinant upper bound. The two identical features at $\boldsymbol{Z}_{\mathrm{col}}$ remain on the unit sphere. There is no zero-column collapse.

4.1 A NON-STRICT, NONMAXIMAL STATIONARY POINT

For a maximisation problem, call a nonglobal stationary point a *strict saddle* when the Riemannian Hessian has a strictly positive direction. This convention includes local minima with positive curvature. A non-strict saddle may still admit an improving higher-order path.

Theorem 5 (Degenerate saddle under sample-wise normalisation)**.** *Take $d = m = 3$, two classes of sizes $(m_1, m_2) = (2, 1)$, and*

$$p = \frac{2}{3}, \qquad \alpha = \sqrt{p}, \qquad \epsilon^2 = \frac{1}{\sqrt{p}}. \tag{17}$$

At $\boldsymbol{Z}_0 = [\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3]$, the Riemannian gradient of F vanishes and the Riemannian Hessian is negative semidefinite. Nevertheless $\boldsymbol{Z}_0$ is not a local maximiser: there are feasible matrices arbitrarily near it with strictly larger objective value. Consequently, $\boldsymbol{Z}_0$ is not a strict saddle either.

This example is not created by a zero sample, an absent class, or an insufficient feature dimension. Every column has norm one, both classes are represented, and $d = m$. The degeneracy arises at a precision where the quadratic change from class-internal correlation cancels, leaving a positive fourth-order term.

5 NUMERICAL VALIDATION

We report 6,440 feature-optimisation runs, 8,775 direct rotation evaluations, and independent automatic-differentiation checks. All configurations and seeds in this numerical landscape suite were retained. The experiments compare terminal representations with the exact values and geometric conditions predicted by the theorems. They distinguish unrestricted feature optimisation from fixed-Gram orientation and rank-one problems, whose optima need not coincide. A small gradient at a numerical endpoint indicates approximate stationarity, not local optimality. Table 2 summarises the runs; the value gap is measured against the exact optimum on each stated feasible set. Detailed grids, stopping criteria, data processing, and additional diagnostics appear in Appendix H.

Table 2: All feature-optimisation runs. $g_{\mathrm{rel}} = \|\operatorname{grad} F\|_F/(1 + |F|)$ is measured on the stated manifold, independently of the optimiser's success flag. The value gap uses the exact optimum for that experiment's feasible set. All 11 runs with a negative optimiser success flag are included.

Experiment	Runs	Maximum g_{rel}	Maximum value gap
Full spheres, sufficient dimension	900	4.91×10^{-8}	1.72×10^{-1}
Synthetic fixed-Gram orientations	120	3.95×10^{-8}	1.43×10^{-14}
Digits fixed-Gram orientations	90	4.68×10^{-8}	1.59×10^{-13}
Full spheres, balanced $K > d$	450	5.44×10^{-8}	3.47×10^{-14}
Rank-one classes, $K > d$	540	2.68×10^{-8}	7.11×10^{-15}
Higher-rank orientations, $Kr > d$	240	1.57×10^{-8}	1.43×10^{-14}
Three-sample precision sweep	4,100	2.54×10^{-8}	1.72×10^{-1}

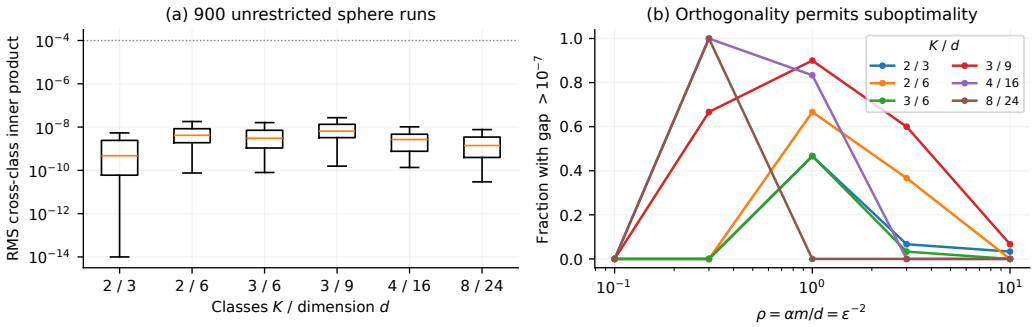


Figure 3: Sufficient-dimension sphere runs: (a) cross-class residuals across 150 runs per configuration; (b) fraction of starts with a global-value gap above 10^{-7} at each precision.

Class separation with sufficient dimension. We optimise every sample on its sphere for six class-size configurations, five precisions, and 30 independent starts per setting. The configurations range from $(m_1, m_2) = (2, 1)$ in $d = 3$ to eight classes of size three in $d = 24$, with $d = m$. For

$$E_{\text{cross}} = \left(\frac{\sum_{k<\ell} \|\boldsymbol{Z}_k^\top \boldsymbol{Z}_\ell\|_F^2}{\sum_{k<\ell} m_k m_\ell} \right)^{1/2}, \tag{18}$$

the maximum across all 900 endpoints is 1.31×10^{-7}. Because each column is unit norm, E_{cross} is the root-mean-square cross-class inner product. Its small value shows that different classes are nearly perpendicular at every observed endpoint. Nevertheless, 215 endpoints have a gap greater than 10^{-7} to the exact global value. Figure 3 therefore checks both class separation and the distinction between orthogonality and global optimality: obtaining the predicted class geometry does not by itself select the best within-class ranks.

Many samples and real-data class geometry. To test the sharper rank budget when $m \gg d$, we fix unit-column matrices $\boldsymbol{H}_k \in \mathbb{R}^{r_k \times m_k}$ and optimise $\boldsymbol{Z}_k = \boldsymbol{U}_k \boldsymbol{H}_k$ over $\boldsymbol{U}_k^\top \boldsymbol{U}_k = \boldsymbol{I}$. These runs preserve sample norms and every class Gram matrix, so improvements can only come from rearranging the class subspaces. Synthetic settings use up to 600 samples in $d = 40$ or 64. For real-data geometry, we use all 1,797 images of the ten-class digits dataset (scikit-learn developers, 2026). Within each class we retain four uncentred singular directions and normalise projected columns. Thus the total class rank is 40, while $d \in \{40, 64, 80\}$; the $d = 40$ case saturates the rank budget despite having far more samples than dimensions. Across 90 digits runs, the largest subspace overlap $\max_{k \neq \ell} \|\boldsymbol{U}_k^\top \boldsymbol{U}_\ell\|_2$ is 3.12×10^{-5} and the largest relative class-Gram change is 1.04×10^{-15}. Figure 4 shows the resulting separation. This experiment tests orientation of data-derived class geometry, without training an encoder or measuring classification accuracy.

Rotation formula checks. We construct 675 nonorthogonal critical configurations with noncommuting class matrices and evaluate 13 rotation angles per configuration. Each test rotates one whole class toward an unused direction, preserving its sample norms and within-class Gram matrix. The observed gain can therefore be compared directly with the closed-form change in the total coding rate. Direct objective differences agree with (5) within 3.19×10^{-13} in absolute error. Twenty-five independent automatic-differentiation checks agree with the proved curvature to relative error 8.58×10^{-14}. Figure 6 reports these checks across the sampled rotations and precisions.

Dimension shortage. For five balanced class configurations with $K > d$, all 450 unrestricted sphere runs reach the value in Theorem 3 to within 3.47×10^{-14}. Every estimated class rank is one, and the maximum relative error of $\boldsymbol{S}$ from $(m/d)\boldsymbol{I}$ is 3.30×10^{-7}. This checks both equality conditions of the full-problem theorem, including the possibility of overlapping class directions. Separately, 540 rank-one runs with balanced or highly unequal class masses reach the spectrum in Theorem 4; the largest relative spectrum error is 4.08×10^{-7}. These runs test the heavy-class and flat-tail prediction within the restricted rank-one family. Finally, 240 higher-rank orientation runs with $K = d$, $r > 1$, and cyclically feasible tight fusion frames attain the overlap bound

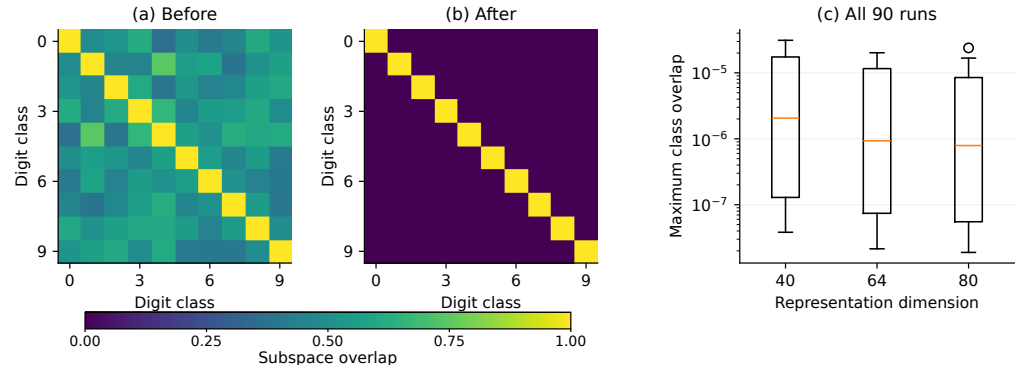


Figure 4: Digits class-subspace overlap before and after optimisation at $d = 40$ (left and centre; shared scale), and terminal maximum overlap across 30 runs per dimension (right).

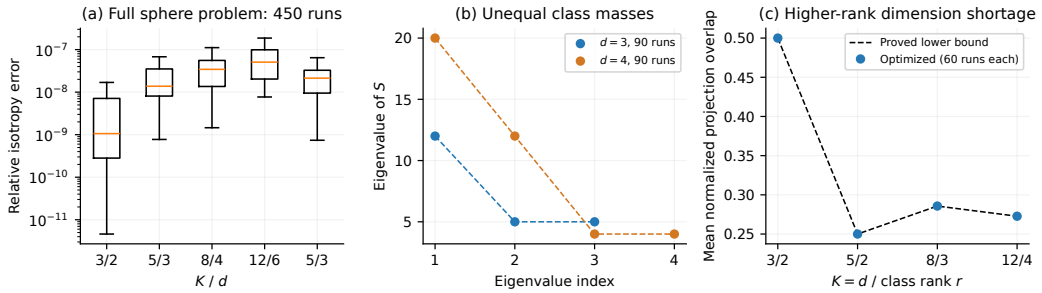


Figure 5: Dimension-limited results: (a) isotropy error for balanced full-sphere runs; (b) predicted and observed spectra for imbalanced rank-one classes; (c) proved and observed higher-rank overlap.

in Proposition 1 within 4.25×10^{-14}. Here the class Gram matrices are fixed, so the comparison concerns orientation rather than the unrestricted global objective. Figure 5 displays the three feasible sets and their distinct predictions.

Suboptimal local optima and degenerate stationary points. In the three-sample example, 4,100 starts map how often optimisation reaches each class-rank branch as precision changes. The rank-two branch changes quadratic stability at $\alpha = \sqrt{2/3}$, whereas the two branches exchange global value only at $\alpha \simeq 1.254884445$; these are different events. Forty-two full tangent-Hessian calculations check the predicted eigenvalues within 8.89×10^{-16}. An 80-digit evaluation resolves the fourth-order improvement where the Hessian has a zero direction. Figure 7 reports these checks in the appendix. The endpoint frequencies describe this initialisation and optimiser, while the analytical proofs establish the landscape statements.

6 Conclusion

We analyse the MCR^2 coding-rate objective itself with unit norm on every feature. When occupied class ranks fit in the ambient space, every nonorthogonal critical point admits an improving class rotation; consequently, local maxima separate classes orthogonally. When classes cannot fit, balanced masses produce rank-one, isotropic global optima, sometimes with coincident class directions. For $K > d$, the rank-one family has no suboptimal local maxima, and fixed higher-rank geometry obeys a sharp overlap bound. These structural guarantees do not ensure global optimisation: a three-sample example has an orthogonal suboptimal local maximum and a nonglobal stationary point with a negative semidefinite Hessian but a fourth-order ascent. The unrestricted imbalanced global problem and within-class local landscape remain open.

AI USE STATEMENT

We used OpenAI GPT-6 to check mathematical proofs and OpenAI Codex to write code for the numerical experiments. Codex also assisted with editing the manuscript and preparing its LaTeX source. The research questions and initial ideas originated with the authors. The authors reviewed the AI-assisted proofs, code, and text and take responsibility for all claims and artifacts in this paper.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630:493–500, 2024. doi: 10.1038/s41586-024-07487-w.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pp. 242–252. PMLR, 2019.

Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando de Freitas. Learning to learn by gradient descent by gradient descent. In *Advances in neural information processing systems*, volume 29, 2016.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, volume 32, pp. 7411–7422, 2019.

Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *arXiv preprint arXiv:1805.12177*, 2018.

Christina Baek, Ziyang Wu, Kwan Ho Ryan Chan, Tianjiao Ding, Yi Ma, and Benjamin D. Haeffele. Efficient maximal coding rate reduction by variational forms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2204.00077.

Bowen Baker, Otkrist Gupta, Nikhil Naik, and Ramesh Raskar. Designing neural network architectures using reinforcement learning. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1611.02167.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6241–6250, 2017.

Sam Buchanan, Dar Gilboa, and John Wright. Deep networks and the multiple manifold problem. *arXiv preprint arXiv:2008.11245*, 2020.

Sam Buchanan, Druv Pai, Peng Wang, and Yi Ma. Principles and practice of deep representation learning: or a mathematical theory of memory. *arXiv preprint arXiv:2606.06624*, 2026. URL https://arxiv.org/abs/2606.06624.

Jacopo Cavazza, Pietro Morerio, Benjamin Haeffele, Connor Lane, Vittorio Murino, and Rene Vidal. Dropout as a low-rank regularizer for matrix factorization. In *International Conference on Artificial Intelligence and Statistics*, pp. 435–444. PMLR, 2018.

Kwan Ho Ryan Chan, Yaodong Yu, Chong You, Haozhi Qi, John Wright, and Yi Ma. Redunet: A white-box deep network from the principle of maximizing rate reduction. *Journal of Machine Learning Research*, 23(114):1–103, 2022. URL https://jmlr.org/papers/v23/21-0631.html.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385, 2022. doi: 10.1109/TPAMI.2021.3057446.

Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1802–1811, 2019. URL https://proceedings.mlr.press/v97/engstrom19a.html.

Gemini Team Google, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. *CoRR*, abs/1712.06541, 2017. URL http://arxiv.org/abs/1712.06541.

Boris Hanin. Universal function approximation by deep neural nets with bounded width and ReLU activations. *arXiv preprint arXiv:1708.02691*, 2017.

Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren (eds.). *Automated Machine Learning: Methods, Systems, Challenges*. Springer, 2019. doi: 10.1007/978-3-030-05318-5.

Elia Kaufmann, Leonard Bauersfeld, Antonio Loquercio, Matthias Müller, Vladlen Koltun, and Davide Scaramuzza. Champion-level drone racing using deep reinforcement learning. *Nature*, 620(7976):982–987, 2023.

Yi Ma, Doris Tsao, and Heung-Yeung Shum. On the principles of parsimony and self-consistency for the emergence of intelligence. *arXiv preprint arXiv:2207.04630*, 2022.

Pedro Massey, Mariano Ruiz, and Demetrio Stojanoff. Optimal frame completions with prescribed norms for majorization, 2013. URL https://arxiv.org/abs/1302.3859.

Pedro G. Massey, Noelia B. Rios, and Demetrio Stojanoff. Frame completions with prescribed norms: local minimizers and applications, 2016. URL https://arxiv.org/abs/1610.02378.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pp. 109–165. Elsevier, 1989.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 08 2019. URL https://arxiv.org/abs/1908.05355.

David Rolnick and Max Tegmark. The power of deeper networks for expressing natural functions. *CoRR*, abs/1705.05502, 2017. URL http://arxiv.org/abs/1705.05502.

scikit-learn developers. Digits dataset: load_digits, 2026. URL https://scikit-learn.org/1.4/modules/generated/sklearn.datasets.load_digits.html. Accessed September 23, 2026.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Yi Tay, Mostafa Dehghani, Jai Gupta, Dara Bahri, Vamsi Aribandi, Zhen Qin, and Donald Metzler. Are pre-trained convolutions better than pre-trained transformers? *arXiv preprint arXiv:2105.03322*, 2021.

Peng Wang, Huikang Liu, Druv Pai, Yaodong Yu, Zhihui Zhu, Qing Qu, and Yi Ma. A global geometric analysis of maximal coding rate reduction. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 51012–51040, 2024. URL https://proceedings.mlr.press/v235/wang24as.html.

Xulong Wang, Xiyang Peng, Zheyuan Xu, Mingchang Xu, Yun Yang, Menghui Zhou, Zhong Zhao, Peng Yue, and Po Yang. Pdwearml: Leveraging daily activities for fast parkinson's disease severity assessment with wearable machine learning. *IEEE Transactions on Biomedical Engineering*, 73(9):3168–3180, 2026. doi: 10.1109/TBME.2025.3648564.

Ziyang Wu, Christina Baek, Chong You, and Yi Ma. Incremental learning via rate reduction. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1125–1133, 2021.

Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113–12132, 2023.

Yaodong Yu, Kwan Ho Ryan Chan, Chong You, Chaobing Song, and Yi Ma. Learning diverse and discriminative representations via the principle of maximal coding rate reduction. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/file/6ad4174eba19ecb5fed17411a34ff5e6-Paper.pdf.

Yaodong Yu, Sam Buchanan, Druv Pai, Tianzhe Chu, Ziyang Wu, Shengbang Tong, Benjamin D. Haeffele, and Yi Ma. White-box transformers via sparse rate reduction. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1e118ba9ee76c20df728b42a35fb4704-Abstract-Conference.html.

Yaodong Yu, Sam Buchanan, Druv Pai, Tianzhe Chu, Ziyang Wu, Shengbang Tong, Hao Bai, Yuexiang Zhai, Benjamin D Haeffele, and Yi Ma. White-box transformers via sparse rate reduction: Compression is all there is? *Journal of Machine Learning Research*, 25(300):1–128, 2024. URL https://www.jmlr.org/papers/v25/23-1547.html.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017.

Xiao Zhang, Yaodong Yu, Lingxiao Wang, and Quanquan Gu. Learning one-hidden-layer ReLU networks via gradient descent. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1524–1534. PMLR, 2019.

Menghui Zhou, Mingxia Wang, Yu Zhang, Zhipeng Yuan, Vitaveska Lanfranchi, and Po Yang. Learning interpretable continuous representation for alzheimer's disease classification. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 6545–6552. IEEE, 2024. doi: 10.1109/BIBM62325.2024.10821731.

Barret Zoph and Quoc V. Le. Neural architecture search with reinforcement learning. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1611.01578.

A Proof of the local-orthogonality and strict-saddle results

We write the proof for maximising F. Replacing F by $L = -F$ reverses every improvement and curvature sign. In particular, the conclusion of Corollary 1 concerns all local minima of the loss, not only global minima.

A.1 Why rotating a whole class is a feasible perturbation

Fix a class k, a skew-symmetric matrix $\boldsymbol{\Omega}^\top = -\boldsymbol{\Omega}$, and $\boldsymbol{Q}(t) = \exp(t\boldsymbol{\Omega})$. Then $\boldsymbol{Q}(0) = \boldsymbol{I}$ and $\boldsymbol{Q}(t)^\top \boldsymbol{Q}(t) = \boldsymbol{I}$. Set

$$\boldsymbol{Z}_k(t) = \boldsymbol{Q}(t)\boldsymbol{Z}_k, \qquad \boldsymbol{Z}_\ell(t) = \boldsymbol{Z}_\ell \quad (\ell \neq k). \tag{19}$$

For each sample in class k, $\|\boldsymbol{Q}(t)\boldsymbol{z}_i\|_2^2 = \boldsymbol{z}_i^\top \boldsymbol{Q}(t)^\top \boldsymbol{Q}(t)\boldsymbol{z}_i = 1$. More strongly,

$$\boldsymbol{Z}_k(t)^\top \boldsymbol{Z}_k(t) = \boldsymbol{Z}_k^\top \boldsymbol{Z}_k. \tag{20}$$

Thus all within-class inner products and all class ranks are fixed. The rotating class has second-moment matrix $\boldsymbol{C}_k(t) = \boldsymbol{Q}(t)\boldsymbol{C}_k\boldsymbol{Q}(t)^\top$, so

$$\begin{aligned} \det\left(\boldsymbol{I} + \frac{\alpha}{p_k}\boldsymbol{C}_k(t)\right) &= \det\left(\boldsymbol{Q}(t)\left(\boldsymbol{I} + \frac{\alpha}{p_k}\boldsymbol{C}_k\right)\boldsymbol{Q}(t)^\top\right) \\ &= \det\left(\boldsymbol{I} + \frac{\alpha}{p_k}\boldsymbol{C}_k\right). \end{aligned} \tag{21}$$

All other class terms are fixed as well. Therefore the only varying term of F is $\frac{1}{2}\log\det(\boldsymbol{I}+\alpha\boldsymbol{S}(t))$.

A.2 Step 1: the first-order condition for class rotations

Lemma 1 (Rotational stationarity)**.** *If $\boldsymbol{Z}$ is stationary under independent orthogonal rotations of its class blocks, then $\boldsymbol{C}_k\boldsymbol{S} = \boldsymbol{S}\boldsymbol{C}_k$ for every k.*

Proof. Since the path above is feasible for both positive and negative t, its derivative is zero at a local maximum. The same is true at every Riemannian critical point. Differentiate $\boldsymbol{C}_k(t)$ at zero, using $\boldsymbol{Q}'(0) = \boldsymbol{\Omega}$ and $(\boldsymbol{Q}^\top)'(0) = -\boldsymbol{\Omega}$:

$$\boldsymbol{S}'(0) = \boldsymbol{C}_k'(0) = \boldsymbol{\Omega}\boldsymbol{C}_k - \boldsymbol{C}_k\boldsymbol{\Omega}. \tag{22}$$

The identity $\frac{d}{dt}\log\det\boldsymbol{A}(t) = \operatorname{tr}(\boldsymbol{A}(t)^{-1}\boldsymbol{A}'(t))$, valid for positive definite $\boldsymbol{A}(t)$, now gives

$$\begin{aligned} 0 = \left.\frac{d}{dt}F(\boldsymbol{Z}(t))\right|_{t=0} &= \frac{\alpha}{2}\operatorname{tr}\left(\boldsymbol{A}^{-1}(\boldsymbol{\Omega}\boldsymbol{C}_k - \boldsymbol{C}_k\boldsymbol{\Omega})\right) \\ &= \frac{\alpha}{2}\operatorname{tr}\left((\boldsymbol{C}_k\boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{C}_k)\boldsymbol{\Omega}\right), \qquad \boldsymbol{A} = \boldsymbol{I} + \alpha\boldsymbol{S}. \end{aligned} \tag{23}$$

The last equality uses cyclic invariance of trace. Put $\boldsymbol{M} = \boldsymbol{C}_k\boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{C}_k$. Since $\boldsymbol{C}_k$ and $\boldsymbol{A}^{-1}$ are symmetric, $\boldsymbol{M}^\top = -\boldsymbol{M}$. We may therefore choose $\boldsymbol{\Omega} = \boldsymbol{M}$, obtaining

$$0 = \operatorname{tr}(\boldsymbol{M}^2) = -\operatorname{tr}(\boldsymbol{M}^\top\boldsymbol{M}) = -\|\boldsymbol{M}\|_F^2. \tag{24}$$

Hence $\boldsymbol{M} = 0$, or $\boldsymbol{C}_k\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{C}_k$. Multiplication on the left and right by $\boldsymbol{A}$ yields $\boldsymbol{A}\boldsymbol{C}_k = \boldsymbol{C}_k\boldsymbol{A}$. Substituting $\boldsymbol{A} = \boldsymbol{I} + \alpha\boldsymbol{S}$ and using $\alpha > 0$ proves $\boldsymbol{S}\boldsymbol{C}_k = \boldsymbol{C}_k\boldsymbol{S}$. □

Two real symmetric commuting matrices admit an orthonormal basis of common eigenvectors. We will use this for $\boldsymbol{S}$ and one chosen $\boldsymbol{C}_k$ at a time. We do not assert that all the class matrices commute with one another.

A.3 Step 2: why nonorthogonality leaves an unused direction

The assumption $\sum_k r_k \leq d$ alone does not imply that a nonorthogonal configuration has an unused direction. For example, two oblique, linearly independent vectors span $\mathbb{R}^2$. The first-order condition from Step 1 is essential to the following argument.

Lemma 2 (Rank equality forces orthogonality at rotational stationarity)**.** *Suppose $\boldsymbol{C}_k\boldsymbol{S} = \boldsymbol{S}\boldsymbol{C}_k$ for every k. If $\sum_k \operatorname{rank}(\boldsymbol{C}_k) = \operatorname{rank}(\boldsymbol{S})$, then $\boldsymbol{C}_k\boldsymbol{C}_\ell = 0$ for every $k \neq \ell$. Consequently, any nonorthogonal rotationally stationary point has $\operatorname{rank}(\boldsymbol{S}) < \sum_k r_k$.*

Proof. Let $V = \operatorname{range}(\boldsymbol{S})$ and $r = \dim V$. To see that every class matrix is supported on V, take $\boldsymbol{x} \in \ker \boldsymbol{S}$ and note

$$0 = \boldsymbol{x}^\top \boldsymbol{S}\boldsymbol{x} = \sum_k \boldsymbol{x}^\top \boldsymbol{C}_k \boldsymbol{x} = \sum_k \|\boldsymbol{Z}_k^\top \boldsymbol{x}\|_2^2. \tag{25}$$

Every summand is nonnegative, so $\boldsymbol{Z}_k^\top \boldsymbol{x} = 0$ and $\boldsymbol{C}_k\boldsymbol{x} = 0$ for every k. On V, the matrix $\boldsymbol{S}$ is positive definite. Its inverse square root is therefore well defined on that subspace. Define, on V only,

$$\boldsymbol{T}_k = \boldsymbol{S}^{-1/2}\boldsymbol{C}_k\boldsymbol{S}^{-1/2}. \tag{26}$$

These matrices express each class's contribution after scaling the total second moment to the identity:

$$\sum_k \boldsymbol{T}_k = \boldsymbol{I}_V, \qquad 0 \preceq \boldsymbol{T}_k \preceq \boldsymbol{I}_V, \qquad \operatorname{rank}(\boldsymbol{T}_k) = \operatorname{rank}(\boldsymbol{C}_k) = r_k. \tag{27}$$

Here the upper bound because $\boldsymbol{I}_V - \boldsymbol{T}_k = \sum_{\ell \neq k} \boldsymbol{T}_\ell \succeq 0$, and the rank equality follows from invertibility of $\boldsymbol{S}^{-1/2}$ on V.

Every positive eigenvalue of $\boldsymbol{T}_k$ lies in $(0, 1]$. Its trace is thus at most its rank. Under the rank-equality assumption,

$$r = \operatorname{tr}(\boldsymbol{I}_V) = \sum_k \operatorname{tr}(\boldsymbol{T}_k) \leq \sum_k \operatorname{rank}(\boldsymbol{T}_k) = \sum_k r_k = r. \tag{28}$$

All inequalities must be equalities. Since each difference $\operatorname{rank}(\boldsymbol{T}_k) - \operatorname{tr}(\boldsymbol{T}_k)$ is nonnegative, every such difference is zero. Hence every positive eigenvalue of each $\boldsymbol{T}_k$ equals one. Each $\boldsymbol{T}_k$ is consequently an orthogonal projector.

We next verify that their ranges are mutually perpendicular. If $\boldsymbol{x}$ belongs to the range of $\boldsymbol{T}_j$, then $\boldsymbol{T}_j\boldsymbol{x} = \boldsymbol{x}$ and

$$\|\boldsymbol{x}\|_2^2 = \boldsymbol{x}^\top \Big(\sum_k \boldsymbol{T}_k\Big)\boldsymbol{x} = \|\boldsymbol{x}\|_2^2 + \sum_{k \neq j} \|\boldsymbol{T}_k\boldsymbol{x}\|_2^2. \tag{29}$$

Thus $\boldsymbol{T}_k\boldsymbol{x} = 0$ for every $k \neq j$, proving $\boldsymbol{T}_k\boldsymbol{T}_j = 0$.

Finally, the commutation from Step 1 allows us to return to the original geometry. Each $\boldsymbol{C}_k$ commutes with $\boldsymbol{S}^{-1/2}$, so $\boldsymbol{T}_k = \boldsymbol{S}^{-1}\boldsymbol{C}_k$ and $\boldsymbol{C}_k = \boldsymbol{S}\boldsymbol{T}_k$ on V. Each $\boldsymbol{T}_k$ also commutes with $\boldsymbol{S}$. Therefore

$$\boldsymbol{C}_k\boldsymbol{C}_\ell = \boldsymbol{S}\boldsymbol{T}_k\boldsymbol{S}\boldsymbol{T}_\ell = \boldsymbol{S}^2\boldsymbol{T}_k\boldsymbol{T}_\ell = 0 \qquad (k \neq \ell). \tag{30}$$

The same product vanishes on $V^\perp$, where all class matrices vanish. In particular,

$$\|\boldsymbol{Z}_k^\top \boldsymbol{Z}_\ell\|_F^2 = \operatorname{tr}(\boldsymbol{C}_k\boldsymbol{C}_\ell) = 0, \tag{31}$$

which is exactly class orthogonality. Since always $\operatorname{rank}(\boldsymbol{S}) \leq \sum_k r_k$, taking the contrapositive proves the final assertion. □

Combining the lemma with the dimension budget gives

$$\text{nonorthogonal rotationally stationary point} \quad \Longrightarrow \quad \operatorname{rank}(\boldsymbol{S}) < \sum_k r_k \leq d. \tag{32}$$

Thus a unit vector $\boldsymbol{v} \in \ker \boldsymbol{S}$ exists. The calculation at the start of the lemma shows that $\boldsymbol{Z}_k^\top \boldsymbol{v} = 0$ for every class: $\boldsymbol{v}$ is perpendicular to every current sample. This is the unused direction needed below, including when the budget is met with equality.

A.4 Step 3: locating a direction shared by two or more classes

Proof of Theorem 1. The theorem assumes $\operatorname{rank}(\boldsymbol{S}) < d$ directly, so an unused unit vector $\boldsymbol{v}$ exists even without invoking the rank-budget argument in Step 2. Step 1 applies at the critical point.

Nonorthogonality means that $\boldsymbol{Z}_j^\top \boldsymbol{Z}_k \neq 0$ for some pair of distinct classes. For this class j,

$$\operatorname{tr}\left(\boldsymbol{C}_j(\boldsymbol{S}-\boldsymbol{C}_j)\right) = \sum_{k\neq j} \operatorname{tr}(\boldsymbol{C}_j\boldsymbol{C}_k) = \sum_{k\neq j} \|\boldsymbol{Z}_j^\top \boldsymbol{Z}_k\|_F^2 > 0. \tag{33}$$

Choose an orthonormal common eigenbasis $\boldsymbol{u}_i$ of $\boldsymbol{C}_j$ and $\boldsymbol{S}$, and write

$$\boldsymbol{C}_j\boldsymbol{u}_i = \mu_i\boldsymbol{u}_i, \qquad \boldsymbol{S}\boldsymbol{u}_i = \lambda_i\boldsymbol{u}_i. \tag{34}$$

The matrices $\boldsymbol{C}_j$ and $\boldsymbol{S}-\boldsymbol{C}_j = \sum_{k\neq j}\boldsymbol{C}_k$ are positive semidefinite. Thus $0 \leq \mu_i \leq \lambda_i$ for every i. Equation (33) becomes

$$\sum_i \mu_i(\lambda_i - \mu_i) > 0. \tag{35}$$

At least one summand is positive. For the corresponding unit vector $\boldsymbol{u}$, we therefore have

$$\boldsymbol{C}_j\boldsymbol{u} = \mu\boldsymbol{u}, \qquad \boldsymbol{S}\boldsymbol{u} = \lambda\boldsymbol{u}, \qquad 0 < \mu < \lambda. \tag{36}$$

The interpretation is concrete: class j has directional energy $\boldsymbol{u}^\top\boldsymbol{C}_j\boldsymbol{u} = \mu$, whereas the other classes together have energy $\boldsymbol{u}^\top(\boldsymbol{S}-\boldsymbol{C}_j)\boldsymbol{u} = \lambda - \mu > 0$. Both occupy this direction. Also $\boldsymbol{u} \perp \boldsymbol{v}$, since they are eigenvectors of the symmetric matrix $\boldsymbol{S}$ with distinct eigenvalues $\lambda > 0$ and zero.

Step 4: the exact improvement from a two-dimensional rotation

Define the orthogonal matrix $\boldsymbol{Q}(t)$ by

$$\boldsymbol{Q}(t)\boldsymbol{u} = \cos t\,\boldsymbol{u} + \sin t\,\boldsymbol{v}, \qquad \boldsymbol{Q}(t)\boldsymbol{v} = -\sin t\,\boldsymbol{u} + \cos t\,\boldsymbol{v}, \tag{37}$$

and let it act as the identity on $\operatorname{span}\{\boldsymbol{u},\boldsymbol{v}\}^\perp$. Rotate only class j: $\boldsymbol{Z}_j(t) = \boldsymbol{Q}(t)\boldsymbol{Z}_j$. This preserves every sample norm, class rank, and class Gram matrix by (20).

In the ordered basis $(\boldsymbol{u},\boldsymbol{v})$, the initial matrices for class j and for all the other classes together are

$$\boldsymbol{C}_j\big|_{(\boldsymbol{u},\boldsymbol{v})} = \begin{pmatrix}\mu & 0\\ 0 & 0\end{pmatrix}, \qquad (\boldsymbol{S}-\boldsymbol{C}_j)\big|_{(\boldsymbol{u},\boldsymbol{v})} = \begin{pmatrix}\lambda-\mu & 0\\ 0 & 0\end{pmatrix}. \tag{38}$$

Both matrices have no coupling to the orthogonal complement because $\boldsymbol{u},\boldsymbol{v}$ are their eigenvectors. Rotation replaces the first block by

$$\boldsymbol{C}_j(t)\big|_{(\boldsymbol{u},\boldsymbol{v})} = \mu\begin{pmatrix}\cos^2 t & \sin t\cos t\\ \sin t\cos t & \sin^2 t\end{pmatrix}, \tag{39}$$

whereas the second block stays fixed. Adding gives

$$\boldsymbol{S}(t)\big|_{(\boldsymbol{u},\boldsymbol{v})} = \begin{pmatrix}\lambda-\mu\sin^2 t & \mu\sin t\cos t\\ \mu\sin t\cos t & \mu\sin^2 t\end{pmatrix}. \tag{40}$$

All other blocks of $\boldsymbol{S}$ are unchanged. Put $s = \sin t$ and $c = \cos t$. The determinant of the changing block of $\boldsymbol{I} + \alpha\boldsymbol{S}(t)$ is

$$\begin{aligned} D(t) &= (1+\alpha\lambda-\alpha\mu s^2)(1+\alpha\mu s^2) - \alpha^2\mu^2 s^2 c^2\\ &= 1 + \alpha\lambda + \alpha^2\lambda\mu s^2 - \alpha^2\mu^2 s^4 - \alpha^2\mu^2 s^2 c^2\\ &= 1 + \alpha\lambda + \alpha^2\mu(\lambda-\mu)\sin^2 t, \end{aligned} \tag{41}$$

where the last step uses $s^2 + c^2 = 1$. At zero, $D(0) = 1 + \alpha\lambda$. The determinants on the orthogonal complement cancel in the ratio, and every within-class rate is fixed. Consequently

$$\begin{aligned} F(\boldsymbol{Z}(t)) - F(\boldsymbol{Z}) &= \frac{1}{2}\log\frac{D(t)}{D(0)}\\ &= \frac{1}{2}\log\left(1 + q\sin^2 t\right), \qquad q = \frac{\alpha^2\mu(\lambda-\mu)}{1+\alpha\lambda} > 0. \end{aligned} \tag{42}$$

This is strictly positive for all sufficiently small nonzero t. It also equals $\frac{q}{2}t^2 + O(t^4)$, so its second derivative at zero is $q > 0$.

The path has a nonzero tangent: its rotating block satisfies $\boldsymbol{Z}_j'(0) = \boldsymbol{v}\boldsymbol{u}^\top \boldsymbol{Z}_j$ because $\boldsymbol{v}^\top \boldsymbol{Z}_j = 0$, and hence $\|\boldsymbol{Z}_j'(0)\|_F^2 = \boldsymbol{u}^\top \boldsymbol{C}_j \boldsymbol{u} = \mu > 0$. At a Riemannian critical point, the second derivative along a smooth feasible path is the Riemannian Hessian quadratic form on the path's tangent. The gradient-dependent acceleration term vanishes. Thus the path provides a strictly positive Hessian direction for F, or a strictly negative one for the loss $L = -F$. □

A.5 Combining the steps: local-optimum orthogonality

Proof of Corollary 1. Suppose, for contradiction, that $\boldsymbol{Z}$ is a nonorthogonal local maximum of F satisfying $\sum_k r_k \le d$. A local maximum is stationary under all class rotations, so Step 1 applies. Step 2 gives $\operatorname{rank}(\boldsymbol{S}) < d$. Steps 3 and 4 then construct feasible matrices $\boldsymbol{Z}(t) \to \boldsymbol{Z}$ with $F(\boldsymbol{Z}(t)) > F(\boldsymbol{Z})$, contradicting local maximality. Hence all class subspaces are mutually orthogonal. The same reasoning at any nonorthogonal Riemannian critical point gives the strictly positive curvature computed above, proving the strict-saddle assertion.

Now suppose the feasible set includes prescribed constraints $\operatorname{rank}(\boldsymbol{Z}_k) \le d_k$, with $\sum_k d_k \le d$. This set need not be a smooth manifold at rank changes. Nevertheless, a local maximum is still stationary under every class rotation, because these paths preserve all class ranks for both signs of t. The rank lemma and the exact improving rotation therefore apply unchanged. The conclusion about local maxima holds also for this rank-constrained problem, including at its singular points. In loss-minimisation notation, every such local minimum of $L = -F$ has $\boldsymbol{Z}_k^\top \boldsymbol{Z}_\ell = 0$ for all $k \neq \ell$. □

B Proof of Theorem 2

We first record the determinant inequality used to separate classes.

Lemma 3. *For positive semidefinite $\boldsymbol{C}_1, \dots, \boldsymbol{C}_K$,*

$$\log\det\left(\boldsymbol{I} + \sum_k \boldsymbol{C}_k\right) \le \sum_k \log\det(\boldsymbol{I} + \boldsymbol{C}_k). \tag{43}$$

Equality holds if and only if $\boldsymbol{C}_k\boldsymbol{C}_\ell = 0$ for every $k \neq \ell$.

Proof. For two terms, the determinant identity and $0 \preceq (\boldsymbol{I} + \boldsymbol{C}_1)^{-1} \preceq \boldsymbol{I}$ give

$$\begin{aligned}\det(\boldsymbol{I} + \boldsymbol{C}_1 + \boldsymbol{C}_2) &= \det(\boldsymbol{I} + \boldsymbol{C}_1)\det\Big(\boldsymbol{I} + \boldsymbol{C}_2^{1/2}(\boldsymbol{I} + \boldsymbol{C}_1)^{-1}\boldsymbol{C}_2^{1/2}\Big)\\ &\le \det(\boldsymbol{I} + \boldsymbol{C}_1)\det(\boldsymbol{I} + \boldsymbol{C}_2).\end{aligned}$$

Equality requires $\boldsymbol{C}_2^{1/2}\big(\boldsymbol{I} - (\boldsymbol{I} + \boldsymbol{C}_1)^{-1}\big)\boldsymbol{C}_2^{1/2} = 0$, equivalently $\boldsymbol{C}_1\boldsymbol{C}_2 = 0$. Induction proves the assertion for K terms. □

Let $s_{k,1}, \dots, s_{k,r_k} > 0$ be the positive eigenvalues of $\boldsymbol{Z}_k\boldsymbol{Z}_k^\top$. They sum to m_k because every column has unit norm. Lemma 3 yields

$$F(\boldsymbol{Z}) \le \sum_{k=1}^{K}\sum_{j=1}^{r_k} \phi_{p_k}(s_{k,j}), \qquad \phi_p(s) = \frac{1}{2}\left[\log(1 + \alpha s) - p\log\left(1 + \frac{\alpha s}{p}\right)\right], \tag{44}$$

with equality precisely when the class spans are orthogonal. Since $d \ge m$, any individually feasible class Gram matrices can be embedded into mutually orthogonal subspaces. Therefore it remains to maximise the right-hand side class by class, subject to $s_j \ge 0$ and $\sum_j s_j = m_k$.

Every nonnegative spectrum with at most m_k entries and sum m_k is realised by an $m_k \times m_k$ positive semidefinite Gram matrix with unit diagonal: its eigenvalues majorise the all-ones diagonal, so the real symmetric Schur–Horn theorem applies. Such a Gram matrix can be factored in $\mathbb{R}^d$ because $d \ge m$. Hence the spectral maximisation is exact, rather than a relaxation.

For $x = \alpha s$,

$$\phi_p'(s) = \frac{\alpha(1-p)x}{2(1+x)(p+x)}, \qquad \phi_p''(s) = \frac{\alpha^2(1-p)(p-x^2)}{2(1+x)^2(p+x)^2}. \tag{45}$$

At a maximum on any fixed positive support, the derivatives of positive coordinates agree. If two positive coordinates $s < t$ differ, this equality implies $(\alpha s)(\alpha t) = p$. Writing $x = \alpha s < \sqrt{p}$ and $y = \alpha t = p/x$, direct substitution gives

$$\phi_p''(s) + \phi_p''(t) = \phi_p''(s)\left(1 - \frac{x^2}{p}\right) > 0. \tag{46}$$

Here $\phi_p''(s) > 0$ because $x < \sqrt{p}$, whereas $\phi_p''(t) = -(x^2/p)\phi_p''(s)$. In particular, this rules out the one-smaller-eigenvalue alternative allowed by the more general scalar lemma in Yu et al. (2020). Ruling out only two small eigenvalues would not suffice. The feasible perturbation $(s,t) \mapsto (s+u, t-u)$ then has positive second derivative, contradicting maximality. Thus all positive eigenvalues in a maximising class block are equal to m_k/r_k for some $1 \le r_k \le m_k$. The resulting value is $G_k(r_k)$. Comparing this finite list gives (8). Unit-norm tight frames with every such rank exist by the same Schur–Horn argument. The equality case of Lemma 3 gives (9) and completes the proof. □

C An alternative Gram-matrix proof of Corollary 1

Let $\boldsymbol{G} = \boldsymbol{Z}^\top \boldsymbol{Z}$ and $\boldsymbol{D} = \mathrm{blockdiag}(\boldsymbol{G}_{11}, \dots, \boldsymbol{G}_{KK})$, where $\boldsymbol{G}_{kk} = \boldsymbol{Z}_k^\top \boldsymbol{Z}_k$. Both are correlation matrices: they are positive semidefinite and have unit diagonal. Suppose $\boldsymbol{G} \neq \boldsymbol{D}$. By Lemma 3 and its equality condition,

$$\log\det(\boldsymbol{I}_m + \alpha\boldsymbol{D}) > \log\det(\boldsymbol{I}_m + \alpha\boldsymbol{G}). \tag{47}$$

For $0 < t \le 1$ put $\boldsymbol{G}(t) = (1-t)\boldsymbol{G} + t\boldsymbol{D}$. These are correlation matrices with exactly the same within-class blocks as $\boldsymbol{G}$. Strict improvement at $\boldsymbol{D}$ and concavity of $\log\det$ imply

$$\log\det(\boldsymbol{I}_m+\alpha\boldsymbol{G}(t)) \ge (1-t)\log\det(\boldsymbol{I}_m+\alpha\boldsymbol{G})+t\log\det(\boldsymbol{I}_m+\alpha\boldsymbol{D}) > \log\det(\boldsymbol{I}_m+\alpha\boldsymbol{G}). \tag{48}$$

The class terms of (1) are unchanged along this path by Sylvester's determinant identity.

Let $H = \mathrm{range}(\boldsymbol{D})$, whose dimension is $q = \sum_k r_k$. Since $\ker \boldsymbol{D} \subseteq \ker \boldsymbol{G}$, both $\boldsymbol{G}$ and $\boldsymbol{G}(t)$ are supported on H. Under $q \le d$, their Gram path has a continuous lift through $\boldsymbol{Z}$: extend the partial isometry in the polar decomposition of $\boldsymbol{Z}$ to an isometry $\boldsymbol{E} : H \to \mathbb{R}^d$, and put $\boldsymbol{Z}(t) = \boldsymbol{E}\boldsymbol{G}(t)^{1/2}$, with the square root viewed as mapping into H. Then $\boldsymbol{Z}(t) \to \boldsymbol{Z}$, $\boldsymbol{Z}(t)^\top \boldsymbol{Z}(t) = \boldsymbol{G}(t)$, and every column has unit norm. Every within-class Gram matrix is exactly fixed. Thus every neighbourhood of $\boldsymbol{Z}$ contains a feasible point of larger objective value. This directly excludes a nonorthogonal local maximum under $\sum_k r_k \le d$, even with fixed within-class Gram matrices. □

D Singleton classes with sufficient dimension

Theorem 6 (Singleton-class landscape)**.** *Suppose every class contains exactly one sample, $m = K \ge 2$, and $d \ge m$. The global maximisers of (1) are exactly the configurations with orthonormal columns,*

$$\boldsymbol{Z}^\top \boldsymbol{Z} = \boldsymbol{I}_m. \tag{49}$$

Every critical point that is not a global maximiser has a tangent direction of strictly positive Riemannian curvature.

For singleton classes, $p_k = 1/m$ and $\boldsymbol{Z}_k \boldsymbol{Z}_k^\top = \boldsymbol{z}_k \boldsymbol{z}_k^\top$ has one eigenvalue equal to one. Therefore

$$F(\boldsymbol{Z}) = \frac{1}{2}\log\det(\boldsymbol{I}_d + \alpha\boldsymbol{S}) - \frac{1}{2}\log(1+m\alpha), \qquad \boldsymbol{S} = \boldsymbol{Z}\boldsymbol{Z}^\top. \tag{50}$$

The matrix $\boldsymbol{S}$ is positive semidefinite, has trace m, and has rank at most $m \le d$. Pad its nonzero spectrum to m entries $s_1, \dots, s_m \ge 0$. Concavity gives

$$\sum_{i=1}^{m} \log(1+\alpha s_i) \le m\log(1+\alpha), \tag{51}$$

with equality exactly when all $s_i = 1$. This spectrum is feasible by taking orthonormal columns. Since $\boldsymbol{Z}^\top \boldsymbol{Z}$ has the same nonzero spectrum as $\boldsymbol{S}$, equality is equivalent to $\boldsymbol{Z}^\top \boldsymbol{Z} = \boldsymbol{I}_m$.

Each class has rank one, so $\sum_k r_k = m \le d$. Any critical point that is not global has nonorthogonal columns, and Corollary 1 supplies a strictly positive Hessian direction. This proves the claimed landscape theorem. □

E Proofs for insufficient representation dimension

E.1 Full sample-wise global optima under the balance condition

Proof of Theorem 3. All sample norms are one, so $\operatorname{tr} \boldsymbol{S} = m$ and $\operatorname{tr} \boldsymbol{C}_k = m_k$. If $\lambda_1, \dots, \lambda_d \ge 0$ are the eigenvalues of $\boldsymbol{S}$, strict concavity gives

$$\frac{1}{2} \log\det(\boldsymbol{I} + \alpha \boldsymbol{S}) = \frac{1}{2} \sum_{i=1}^{d} \log(1 + \alpha \lambda_i) \le \frac{d}{2} \log(1 + \alpha m/d), \tag{52}$$

with equality exactly when $\boldsymbol{S} = (m/d)\boldsymbol{I}_d$.

For a class matrix $\boldsymbol{C}_k$ with nonnegative eigenvalues σ_i, the product expansion gives

$$\det\left(\boldsymbol{I} + \frac{\alpha}{p_k} \boldsymbol{C}_k\right) = \prod_i \left(1 + \frac{\alpha}{p_k} \sigma_i\right) \ge 1 + \frac{\alpha}{p_k} \sum_i \sigma_i = 1 + \alpha m. \tag{53}$$

Equality holds precisely when at most one σ_i is positive. Since $m_k > 0$, this means $\operatorname{rank} \boldsymbol{C}_k = 1$. Weighting the logarithms by $p_k/2$ and summing, using $\sum_k p_k = 1$, yields $F(\boldsymbol{Z}) \le F^*$ as in (10). Equality requires exactly the two conditions in (11).

It remains to prove simultaneous attainability. Write $w_k = m_k$, sorted decreasingly, and $\tau = m/d$. The hypothesis says every $w_k \le \tau$, and implies $K \ge d$. The vector

$$(\underbrace{\tau, \dots, \tau}_{d \text{ entries}}, 0, \dots, 0) \in \mathbb{R}^K \tag{54}$$

majorises $(w_1, \dots, w_K)$: for $j \le d$, the sum of the largest j weights is at most $j\tau$, and for $j > d$ it is at most the common total m. The real symmetric Schur–Horn theorem therefore gives a positive semidefinite matrix with this spectrum and diagonal (w_k). Factor it as $\boldsymbol{G} = \boldsymbol{V}^\top \boldsymbol{V}$ with $\boldsymbol{V} \in \mathbb{R}^{d \times K}$. Its columns $\boldsymbol{v}_k$ have squared norm w_k, and $\boldsymbol{V}\boldsymbol{V}^\top = \tau \boldsymbol{I}_d$. Taking $\boldsymbol{u}_k = \boldsymbol{v}_k / \sqrt{w_k}$ and making the m_k unit samples of class k equal to $\boldsymbol{u}_k$ produces an admissible $\boldsymbol{Z}$ attaining both bounds. Arbitrary sample signs give the same class matrices. This proves the value and all equality cases. □

E.2 The complete rank-one landscape with arbitrary class masses

On the rank-one family, $\boldsymbol{C}_k = w_k \boldsymbol{u}_k \boldsymbol{u}_k^\top$ and $\det(\boldsymbol{I} + (\alpha/p_k)\boldsymbol{C}_k) = 1 + \alpha m$. Thus

$$F(\boldsymbol{u}_1, \dots, \boldsymbol{u}_K) = \frac{1}{2} \log\det(\boldsymbol{I} + \alpha \boldsymbol{S}) - \frac{1}{2} \log(1 + \alpha m), \qquad \boldsymbol{S} = \sum_k w_k \boldsymbol{u}_k \boldsymbol{u}_k^\top. \tag{55}$$

The class signs do not affect this objective. We prove Theorem 4 directly.

Stationary directions lie in eigenspaces of $\boldsymbol{S}$. For each class direction, the Euclidean gradient of the nonconstant term is $\alpha w_k (\boldsymbol{I} + \alpha \boldsymbol{S})^{-1} \boldsymbol{u}_k$. Its tangent projection vanishes precisely when $\boldsymbol{u}_k$ is an eigenvector of $\boldsymbol{S}$. Group all the $\boldsymbol{u}_k$ by their eigenvalue. In any positive-eigenvalue eigenspace E of dimension q, with eigenvalue λ, the corresponding index set J satisfies

$$\sum_{k \in J} w_k \boldsymbol{u}_k \boldsymbol{u}_k^\top = \lambda \boldsymbol{I}_E. \tag{56}$$

These vectors span E, so $|J| \ge q$. If $|J| = q$, the square matrix whose columns are $\sqrt{w_k}\boldsymbol{u}_k$ has product with its transpose equal to $\lambda \boldsymbol{I}_E$. Its transpose product is then also $\lambda \boldsymbol{I}_q$. Hence these $\boldsymbol{u}_k$ are orthonormal and all their weights equal λ.

A redundant higher-eigenvalue group gives strict positive curvature. Suppose an eigenspace E_h of eigenvalue λ_h contains more class directions than its dimension, and a lower eigenvalue $\lambda_\ell < \lambda_h$ exists, possibly zero. Linear dependence supplies nonzero real coefficients c_k, supported in E_h, such that

$$\sum_{k \in J_h} w_k c_k \boldsymbol{u}_k = 0. \tag{57}$$

Choose a unit vector $\boldsymbol{v}$ in the lower eigenspace and define unit-sphere paths

$$\boldsymbol{u}_k(t) = \cos(c_k t)\boldsymbol{u}_k + \sin(c_k t)\boldsymbol{v} \quad (k \in J_h), \tag{58}$$

leaving all other directions fixed. Because $\boldsymbol{u}_k \perp \boldsymbol{v}$, each norm is exactly one. The chosen dependence makes

$$\boldsymbol{S}'(0) = \sum_{k \in J_h} w_k c_k (\boldsymbol{v}\boldsymbol{u}_k^\top + \boldsymbol{u}_k \boldsymbol{v}^\top) = 0. \tag{59}$$

Also $\boldsymbol{S}''(0) = 2\sum_{k \in J_h} w_k c_k^2 (\boldsymbol{v}\boldsymbol{v}^\top - \boldsymbol{u}_k \boldsymbol{u}_k^\top)$. The second derivative of (55) therefore is

$$F''(0) = \alpha \sum_{k \in J_h} w_k c_k^2 \left(\frac{1}{1+\alpha\lambda_\ell} - \frac{1}{1+\alpha\lambda_h} \right) > 0. \tag{60}$$

The usual negative trace term in the second derivative of log determinant vanishes because $\boldsymbol{S}'(0) = 0$. The tangent is nonzero, and this proves strict positive Riemannian curvature at the critical point.

All other critical configurations have the stated optimal structure. Since $K > d$, a rank-deficient $\boldsymbol{S}$ necessarily has a redundant positive-eigenvalue group. Taking $\lambda_\ell = 0$ yields (60). A critical point without a strictly positive direction must therefore have full rank. Let $\tau > 0$ be its smallest eigenvalue. Every eigenspace of eigenvalue greater than τ must have exactly as many class directions as its dimension, by the preceding paragraph. These are orthogonal individual classes of weights strictly larger than τ. Let their total number be r.

All remaining class directions lie in the $(d-r)$-dimensional lowest eigenspace, where they satisfy $\sum_{k>r} w_k \boldsymbol{u}_k \boldsymbol{u}_k^\top = \tau \boldsymbol{I}$. Each of their weights is at most τ, since $w_k \boldsymbol{u}_k \boldsymbol{u}_k^\top \preceq \tau \boldsymbol{I}$ on that space. Taking the trace yields $(d-r)\tau = \sum_{k>r} w_k$. This proves (12) and the claimed structure. The threshold is unique: starting with the average m/d, remove the largest weight whenever it exceeds the average of the remaining weights per remaining dimension. The removed weights are exactly those strictly above the final τ. Equivalently,

$$r = \max\left(\{0\} \cup \left\{ j \in \{1, \dots, d-1\} : w_j > \frac{\sum_{k>j} w_k}{d-j} \right\} \right). \tag{61}$$

Finally, a global maximum exists by compactness. It is critical and cannot have the positive-curvature direction above, so it must have this structure. All configurations with the structure have the same spectrum and the same objective value,

$$F^*_{\text{rank 1}} = \frac{1}{2}\sum_{i=1}^{r} \log(1+\alpha w_i) + \frac{d-r}{2}\log(1+\alpha\tau) - \frac{1}{2}\log(1+\alpha m). \tag{62}$$

They are consequently all global. Every other critical point admits (60). This proves both absence of nonglobal local maxima and the strict-saddle assertion on the rank-one family. □

E.3 Overlap bounds for higher-rank classes

Proof of Proposition 1. Expanding the square of $\boldsymbol{S} = \sum_k \boldsymbol{C}_k$ gives

$$\sum_{k \neq \ell} \operatorname{tr}(\boldsymbol{C}_k \boldsymbol{C}_\ell) = \operatorname{tr}(\boldsymbol{S}^2) - \sum_k \operatorname{tr}(\boldsymbol{C}_k^2). \tag{63}$$

Cauchy–Schwarz for the d eigenvalues of $\boldsymbol{S}$ yields $\operatorname{tr}(\boldsymbol{S}^2) \geq (\operatorname{tr}\boldsymbol{S})^2/d = m^2/d$, with equality exactly when $\boldsymbol{S} = (m/d)\boldsymbol{I}_d$. This proves (14). If class Gram matrices are fixed, every class term of F is fixed as well. The total rate is bounded by (52), whose equality condition is the same isotropy

condition. If an isotropic orientation is feasible, this proves necessity and sufficiency for global orientation optimality.

For $\boldsymbol{C}_k = (n/r)\boldsymbol{P}_k$, substitute $m = Kn$ and $\operatorname{tr}(\boldsymbol{C}_k^2) = n^2/r$ into (14). Dividing by $(n/r)^2 K(K-1)r$ gives (15). Its equality condition becomes $\sum_k \boldsymbol{P}_k = (Kr/d)\boldsymbol{I}_d$. In the cyclic example with $K = d$, each coordinate axis belongs to exactly r projectors, proving the asserted feasibility. This result concerns fixed within-class geometry. It does not assert that these higher-rank configurations globally maximise the full objective when within-class geometry can change. □

F Proof of Proposition 2

Keep $p = 2/3$ and let ϕ_p be as in (44). For a nearby feasible $\boldsymbol{Z}$, the correlation $c = \boldsymbol{z}_1^\top \boldsymbol{z}_2$ is close to one, and the two eigenvalues of the first-class Gram matrix are $1 + c$ and $1 - c$. Lemma 3 gives

$$F(\boldsymbol{Z}) \le \phi_p(1+c) + \phi_p(1-c) + \phi_{1-p}(1), \tag{64}$$

with equality when the singleton-class feature is orthogonal to the first-class span. The derivative of the right-hand side with respect to c at $c = 1$ is $\phi_p'(2) - \phi_p'(0) = \phi_p'(2) > 0$. Since $c \le 1$, this upper bound is strictly smaller than its value at $c = 1$ for every sufficiently nearby $c < 1$. At $c = 1$, the determinant inequality makes $\boldsymbol{Z}_{\mathrm{col}}$ locally optimal among all placements of the singleton vector. Hence $\boldsymbol{Z}_{\mathrm{col}}$ is a local maximiser for every $\alpha > 0$.

At $\alpha = 2$, Theorem 2 applies. The first class has only two rank choices, and direct substitution gives

$$G_1(2) - G_1(1) = \log 3 - \frac{2}{3}\log 4 - \frac{1}{2}\log 5 + \frac{1}{3}\log 7 = \frac{1}{6}\log\frac{3^6 7^2}{4^4 5^3} = \frac{1}{6}\log\frac{35721}{32000} > 0. \tag{65}$$

The singleton contribution is the same in both configurations. Therefore $\boldsymbol{Z}_{\mathrm{orth}}$ is global and $\boldsymbol{Z}_{\mathrm{col}}$ is suboptimal. □

G Proof of Theorem 5

At $\boldsymbol{Z}_0 = [\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3]$, all three matrices $\boldsymbol{A}, \boldsymbol{B}_1, \boldsymbol{B}_2$ are diagonal in the standard basis. Equation (3) therefore makes every Euclidean column gradient parallel to its column, so the Riemannian gradient vanishes.

For nearby unit vectors let $c = \boldsymbol{z}_1^\top \boldsymbol{z}_2$, $a = \boldsymbol{z}_1^\top \boldsymbol{z}_3$, and $b = \boldsymbol{z}_2^\top \boldsymbol{z}_3$. The Gram matrix is

$$\boldsymbol{G} = \boldsymbol{Z}^\top \boldsymbol{Z} = \begin{pmatrix} 1 & c & a \\ c & 1 & b \\ a & b & 1 \end{pmatrix}. \tag{66}$$

By Sylvester's determinant identity, the varying part of F can be written exactly as

$$\begin{aligned} F(\boldsymbol{Z}) = \frac{1}{2}\log\bigl(h^3 - h\alpha^2(c^2 + a^2 + b^2) + 2\alpha^3 abc\bigr) \\ - \frac{p}{2}\log\left(k^2 - \frac{\alpha^2}{p^2}c^2\right) - \frac{1-p}{2}\log\left(1 + \frac{\alpha}{1-p}\right), \end{aligned} \tag{67}$$

where $h = 1 + \alpha$ and $k = 1 + \alpha/p$. At $\boldsymbol{Z}_0$, $a = b = c = 0$. The quadratic Taylor part of (67) is

$$\frac{\alpha^2}{2}\left[\frac{p}{(p+\alpha)^2} - \frac{1}{(1+\alpha)^2}\right]c^2 - \frac{\alpha^2}{2(1+\alpha)^2}(a^2 + b^2). \tag{68}$$

When $\alpha = \sqrt{p}$, the coefficient of c^2 is zero and the other two coefficients are negative. Since a, b, c all vanish at $\boldsymbol{Z}_0$, their first variations determine the full Riemannian Hessian there. Equation (68) therefore proves that the Hessian is negative semidefinite.

To show that $\boldsymbol{Z}_0$ is not a local maximum, keep $\boldsymbol{z}_1 = \boldsymbol{e}_1$ and $\boldsymbol{z}_3 = \boldsymbol{e}_3$, and set $\boldsymbol{z}_2(c) = c\boldsymbol{e}_1 + \sqrt{1-c^2}\,\boldsymbol{e}_2$ for $0 < |c| < 1$. This is a feasible path with $a = b = 0$. Put $\gamma = \alpha^2/(p+\alpha)^2$. At $\alpha = \sqrt{p}$, one checks $\alpha^2/(1+\alpha)^2 = p\gamma$. Hence

$$F(\boldsymbol{Z}(c)) - F(\boldsymbol{Z}_0) = \frac{1}{2}\left[\log(1 - p\gamma c^2) - p\log(1 - \gamma c^2)\right] = \frac{p(1-p)\gamma^2}{4}c^4 + O(c^6) > 0. \tag{69}$$

The strict inequality follows directly from strict concavity of the logarithm, applied to $1 - p\gamma c^2 = p(1 - \gamma c^2) + (1-p)\cdot 1$. Therefore $\boldsymbol{Z}_0$ is a nonmaximal stationary point with no positive Hessian direction. □

H EXPERIMENTAL PROTOCOL AND ADDITIONAL DIAGNOSTICS

Parameters and reproducibility. Except in the three-sample precision sweep, we parameterise precision by $\rho = \alpha m/d = \epsilon^{-2}$. The suite uses deterministic seeds $2026092300 + j$, where j is the zero-based task index in the saved manifest. Each task stores its configuration, optimiser status, objective, constraint and stationarity residuals, and initial and final matrices. All 6,440 tasks completed without an exception. Eleven optimiser success flags were false. These runs were retained, and their independently measured relative Riemannian gradients satisfied the same 10^{-6} reporting threshold as every other run. No restart was selected or discarded according to its objective value.

The main suite ran on a Linux x86-64 server using eight CPU worker processes, one numerical-library thread per worker, Python 3.10.13, NumPy 1.26.3, and SciPy 1.11.4. All objectives and gradients use float64. The saved environment records include platform information and the SHA-256 hash of the executed script. Code snapshots accompany the results. Figures and tables are generated from the saved records without rerunning optimisation.

Full sphere optimisation. We minimise $-F$ with L-BFGS-B through the parameterisation $\boldsymbol{z}_i = \boldsymbol{y}_i/\|\boldsymbol{y}_i\|$. The gradient includes the normalisation derivative. The objective and Euclidean feature gradient are evaluated by Cholesky solves, with smaller class Gram matrices used where appropriate. We set the iteration limit to 1,800, gradient tolerance to 10^{-10}, function tolerance to 10^{-15}, and maximum line-search steps to 40. If the independently measured terminal $\|\operatorname{grad} F\|_F/(1 + |F|)$ exceeds 10^{-6}, a predeclared second pass starts from the normalised endpoint. Every evaluated feature matrix has unit columns. Numerical class rank counts singular values above 10^{-5} times the largest singular value in that class.

The sufficient-dimension configurations are $(2, 1)$, $(3, 3)$, $(2, 2, 2)$, $(6, 2, 1)$, four classes of size four, and eight classes of size three, with $d = m$. Each uses $\rho \in \{0.1, 0.3, 1, 3, 10\}$ and 30 starts, totalling 900 runs. The exact comparison value comes from Theorem 2.

The balanced dimension-limited configurations are three classes of size two in $d = 2$, five classes of size three in $d = 3$, eight classes of size two in $d = 4$, twelve classes of size three in $d = 6$, and sizes $(4, 3, 3, 2, 2)$ in $d = 3$. Each uses $\rho \in \{0.1, 1, 10\}$ and 30 starts, totalling 450 runs. All satisfy $\max_k m_k \le m/d$, so (10) is an exact full-problem reference value. The relative isotropy error is $\|\boldsymbol{S} - (m/d)\boldsymbol{I}\|_F/(m/\sqrt{d})$.

Rank-one optimisation with unequal class masses. The weighted configurations include six, twelve, and twenty-four singleton classes in dimensions three, four, and eight, respectively. We also use class sizes $(12, 2, 2, 2, 2, 2)$ in $d = 3$, $(20, 12, 2, 2, 2, 2)$ in $d = 4$, and $(8, 7, 6, 5, 4, 3, 2, 1)$ in $d = 4$. Three precisions $\rho \in \{0.1, 1, 10\}$ and 30 starts give 540 runs. We optimise one unit direction per class in (55), with a 2,200-iteration limit and otherwise the same tolerances. The reference spectrum is the heavy-weight/flat-tail spectrum in Theorem 4. The spectrum residual is its relative Euclidean error. This exact reference is for the rank-one feasible set.

Fixed-Gram class orientations. For $\boldsymbol{Z}_k = \boldsymbol{U}_k\boldsymbol{H}_k$, the matrices $\boldsymbol{H}_k$ are fixed and have unit columns. We parameterise $\boldsymbol{U}_k$ by the reduced QR factor of a full-column-rank matrix and differentiate the objective through QR using PyTorch. L-BFGS-B uses a 1,800-iteration limit, memory 20, and the tolerances above. We measure stationarity on the product of Stiefel manifolds: the ambient gradient $\boldsymbol{G}_k = \alpha(\boldsymbol{I} + \alpha\boldsymbol{S})^{-1}\boldsymbol{U}_k\boldsymbol{H}_k\boldsymbol{H}_k^\top$ is projected as $\boldsymbol{G}_k - \boldsymbol{U}_k \operatorname{sym}(\boldsymbol{U}_k^\top\boldsymbol{G}_k)$. This projection removes any scaling effects from the QR parameterisation. The relative Gram residual is $\max_k \|\boldsymbol{Z}_k^\top\boldsymbol{Z}_k - \boldsymbol{H}_k^\top\boldsymbol{H}_k\|_F/\|\boldsymbol{H}_k^\top\boldsymbol{H}_k\|_F$.

Synthetic sufficient-dimension configurations (K, r, m_k, d) are $(3, 2, 30, 6)$, $(5, 3, 40, 15)$, $(10, 4, 60, 40)$, and $(10, 4, 60, 64)$. Gaussian coordinates are scaled along the class axes by geometrically spaced factors from 1 to 0.3 and then column-normalised. Ten starts and three precisions produce 120 runs. Since $Kr \le d$, the separated class determinant bound is attainable and supplies the exact reference value.

For digits, we load the 1,797 images supplied by `load_digits` (scikit-learn developers, 2026), divide the 64 pixel coordinates by 16, take the leading four left singular vectors separately within each labelled class without centering, and normalise each projected sample to unit length. We keep

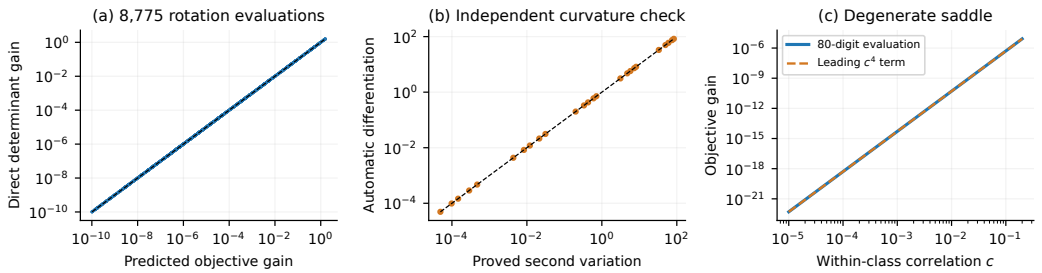


Figure 6: Independent identity checks. (a) Direct determinant differences agree with the exact rotation formula. (b) Automatic differentiation agrees with the proved positive curvature. (c) An 80-digit calculation resolves the positive fourth-order gain at the degenerate saddle down to correlations of 10^{-5}.

all samples. The projected class matrices, sample counts, and retained-energy fractions are saved. With $r_k = 4$, the dimension choices $d \in \{40, 64, 80\}$ satisfy the rank budget. Ten independent orientation starts and three precisions give 90 runs. The initial class orientations are random. The within-class geometry is data-derived. Neither test accuracy nor neural-network optimisation is evaluated.

For dimension-limited higher-rank orientations, $(K = d, r, m_k)$ equals $(3, 2, 30)$, $(5, 2, 40)$, $(8, 3, 60)$, or $(12, 4, 80)$. Each $\boldsymbol{H}_k$ consists of repeated columns of $\boldsymbol{I}_r$, so $\boldsymbol{H}_k \boldsymbol{H}_k^\top = (m_k/r)\boldsymbol{I}_r$. Twenty starts at each of three precisions give 240 runs. The cyclic construction in Proposition 1 guarantees that the isotropic reference value is attainable. The reported overlap is $[K(K-1)r]^{-1} \sum_{k \neq \ell} \|\boldsymbol{U}_k^\top \boldsymbol{U}_\ell\|_F^2$.

Rotation identities and Hessians. For $K \in \{3, 4, 5, 8, 12\}$, $r \in \{1, 2, 3\}$, and repetition count $n \in \{1, 4, 16\}$, construct the r orthogonal directions of class k as

$$\boldsymbol{a}_{k,h} = \cos(\pi k/K)\boldsymbol{e}_{2h-1} + \sin(\pi k/K)\boldsymbol{e}_{2h}, \qquad 1 \leq h \leq r, \tag{70}$$

and repeat each direction n times. Then $\boldsymbol{C}_k$ has nonzero eigenvalue n on its class space, while $\boldsymbol{S} = (Kn/2)\boldsymbol{I}$ on the occupied $2r$ dimensions. Thus the configuration is Riemannian critical although the class matrices generally do not commute pairwise. Embed into $d = Kr$ or $Kr + 3$ and apply a random common orthogonal transformation. Five precisions $\rho \in \{0.01, 0.1, 1, 10, 100\}$ and three random embeddings give 675 configurations. For each, 13 logarithmically spaced angles from 10^{-4} to 0.5 produce 8,775 direct evaluations, with $\mu = n$ and $\lambda = Kn/2$ in the prediction. The maximum initial gradient norm is 1.77×10^{-13}, and the maximum entrywise Gram change is 6.67×10^{-16}.

On every 27th configuration, we differentiate the actual determinant objective twice through the rotation path using automatic differentiation, giving 25 independent second-variation checks. Figure 6(a) omits gains below 10^{-10} only from the log-scale scatter to avoid subtraction-dominated markers. All 8,775 evaluations enter the reported absolute error and are retained in the raw table. For gains above 10^{-8}, the maximum relative error is 2.20×10^{-7}.

Three-sample phase sweep. For sizes $(2, 1)$ and $d = 3$, use 41 logarithmically spaced values of α from 0.1 to 10, with 100 starts each. An endpoint is assigned to the rank-one or rank-two class branch when $|\boldsymbol{z}_1^\top \boldsymbol{z}_2| > 0.99$ or < 0.01, respectively. The two objective branches exchange global optimality at $\alpha \simeq 1.254884445$, obtained from the rank-score equality. This differs from the rank-two Hessian transition $\alpha = \sqrt{2/3} \simeq 0.816497$. Figure 7 shows the observed endpoint fractions and 95% Wilson binomial intervals, conditional on the stated initialisation and optimiser. They are not estimates of an optimiser-independent geometric basin volume.

At $\boldsymbol{Z}_0 = \boldsymbol{I}_3$, the full six-dimensional tangent Hessian is computed at the 41 sweep precisions and at $\sqrt{2/3}$. Its predicted eigenvalues are three zeros from common rotations, two copies of $-2\alpha^2/(1+\alpha)^2$, and

$$2\alpha^2 \left[\frac{2/3}{(2/3+\alpha)^2} - \frac{1}{(1+\alpha)^2} \right]. \tag{71}$$

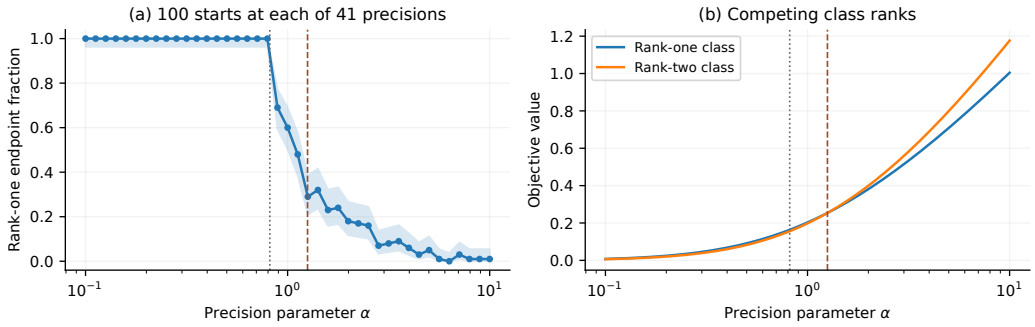


Figure 7: Precision-dependent rank selection in the three-sample example. Left: rank-one endpoint frequency and 95% Wilson intervals from 100 starts per precision. Right: exact objective values of the two class-rank branches. The dotted line marks the Hessian transition $\sqrt{2/3}$. The dashed line marks the distinct global-value crossing.

The maximum discrepancy is 8.89×10^{-16}. At the zero-curvature threshold, the exact gain is also evaluated with 80-digit arithmetic to avoid cancellation of its quadratic terms.

I CIFAR-10: Frobenius regularisation in the original training protocol

Question and scope. This experiment asks how the Frobenius-regularised objective behaves when inserted into the released MCR^2 CIFAR-10 training pipeline. We train only the regularised objective. In particular, this experiment does not independently retrain a sample-wise normalised arm, so its outcomes alone cannot establish a test-accuracy difference between the two formulations. The two completed regularisation values are $\lambda \in \{10^{-3}, 10^{-4}\}$. Earlier exploratory values 0.1 and 0.01 first produced nonfinite loss at minibatches 2 and 4 of epoch 1, respectively. Neither yielded valid test accuracy. The additional value 10^{-4} was chosen after observing those failures, without using test accuracy.

Data and encoder. We use all 50,000 CIFAR-10 training images and the official 10,000-image test set, with no validation split. The encoder is the released CIFAR ResNet-18 and 512-dimensional projection head from the original training code. The sole architectural change is to remove the final sample-wise ℓ_2 normalisation, so the loss receives raw features $\boldsymbol{h}_i$. We add the minibatch penalty $(\lambda/2)\sum_i \|\boldsymbol{h}_i\|_2^2$ to the original loss based on negative coding rate reduction. This is the sum-form Frobenius penalty rather than a mean over samples. The original implementation's precision argument is set to `--eps 0.5`. Its loss code uses this value directly in the denominator $d/(m\,\mathtt{eps})$.

Training. Both completed values use the same single random seed, 42. We retain the original CIFAR-10 script's training setup: batch size 1,000, SGD with learning rate 0.01, momentum 0.9, weight decay 5×10^{-4}, and 500 epochs. The original phase scheduler reduces the learning rate at epochs 200, 400, and 500. Training images use a random 32×32 crop with eight pixels of padding, random horizontal flip, and tensor conversion. Test images use tensor conversion. There is no label corruption. Apart from removal of output normalisation and addition of the stated Frobenius term, the optimiser, rate-reduction implementation, data pipeline, and other code defaults are retained.

Evaluation and reporting. For the Frobenius-regularised result in Table 1, we use nearest-subspace classification, following the uncentred SVD branch of the released `evaluate.py` routine `nearsub`. We extract features from the full training set and official test set using the deterministic test transform. For each class, we fit a 30-dimensional SVD subspace to its training features, then assign each test feature to the subspace with the smallest Euclidean residual. The randomised SVD uses seed 42. We evaluate the saved epoch-500 checkpoint for each completed run. The training script's logged classifiers do not include this nearest-subspace measure, so the table uses a separate evaluation of those checkpoints.

Results. The two completed runs both used seed 42 and all 500 epochs. Nearest-subspace test accuracy was 27.14% for $\lambda = 10^{-3}$ and 45.57% for $\lambda = 10^{-4}$. The two exploratory larger penalties, 0.1 and 0.01, first produced nonfinite losses at minibatches 2 and 4 of epoch 1, respectively, and yielded no valid test accuracy. A preliminary $\lambda = 10^{-3}$ launch was stopped after epoch 1 while correcting the evaluation output from training to test accuracy. Its incomplete artefacts are retained separately and the reported completed run began from initialisation with the same seed and training settings. These are single-seed outcomes, not estimates of variability. The improvement at the smaller completed penalty motivates testing still smaller values. Because no sample-wise normalised arm was independently retrained under this protocol, these data do not establish a test-accuracy difference between exact and regularised MCR^2.