ELASTIC CODING: ROBUST MAXIMAL CODING RATE REDUCTION UNDER LABEL NOISE

Anonymous authors
Paper under double-blind review

ABSTRACT

Maximal Coding Rate Reduction (MCR^2) exhibits striking empirical robustness to corrupted labels, yet neither the source nor the limit of this robustness is understood. We answer three connected questions. First, why is MCR^2 robust? Label corruption leaves its global expansion term unchanged, sample-wise normalisation bounds the influence of every training example, and uniformly random label flips distribute corrupted examples across distinct off-class directions in the class-level spectrum. Under a balanced union-of-orthogonal-subspaces model, the clean semantic subspace remains the unique leading eigenspace of every observed class when flipped labels are chosen uniformly from the other classes and the flip rate satisfies $\eta < (K-1)/K$. Thus, with ten classes and 50% corruption, the clean spectral direction is nine times stronger than any individual corrupted direction. Second, when does robustness fail? We show that failure is governed not by the noise rate alone, but by a clean-dominance gap: concentrated corruption can form an off-class subspace whose spectral strength rivals or exceeds the clean one, causing competing class-compression directions. Third, how can this failure be corrected? A gradient decomposition identifies the component of class compression that directly opposes label-independent global expansion. This leads to Elastic Coding, which relaxes class compression when this opposing component becomes excessive. Experiments on multiple datasets with structured and real-world label noise show that Elastic Coding achieves the best performance among the compared baselines in most evaluated settings.

1 INTRODUCTION

Over the past decade, deep learning has transformed natural language processing (Katz et al., 2024; Gemini Team et al., 2024; DeepSeek-AI et al., 2025), biology (Abramson et al., 2024; Oh et al., 2023), game playing (Wurman et al., 2022; Fukaya et al., 2024), and robotics (Kaufmann et al., 2023; DeSouza & Kak, 2002). Much of this progress, however, has come from empirical design. Networks are often developed through iterative experimentation, trained from random initialisation by backpropagation (Rumelhart et al., 1986; Groenendijk et al., 2023), and deployed as effective but difficult-to-interpret models (Chan et al., 2022). They can also be sensitive to simple transformations (Azulay & Weiss, 2018; Engstrom et al., 2019), vulnerable to adversarial perturbations (Szegedy et al., 2013), prone to catastrophic forgetting (McCloskey & Cohen, 1989; De Lange et al., 2021; Wu et al., 2021), and capable of memorising arbitrary labels (Zhang et al., 2017). These limitations expose a gap between empirical success and principled understanding.

Theoretical research has illuminated 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 analyses 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), while others explain individual mechanisms such as dropout (Cavazza et al., 2018). These insights offer limited guidance on what representations practical networks should learn. Model development therefore continues to rely heavily on empirical search, including neural architecture search (Zoph & Le, 2017; Baker et al., 2017) and automated machine learning (Hutter et al., 2019). This gap motivates a principle for designing the structure of learned representations.

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.

The *Maximal Coding Rate Reduction* (MCR^2) principle (Yu et al., 2020) shifts attention from *why* a network can fit data to *what* structure its representation should learn. It maximises the gap between the coding rate of the full dataset and the average coding rate of its classes, encouraging diverse features across classes and compact structure within each class (Figure 1). Its original formulation encompasses supervised, self-supervised, and unsupervised representation learning. Subsequent work has applied rate reduction to unsupervised representation learning and image clustering (Tong et al., 2022; Chu et al., 2024), incremental learning (Wu et al., 2021), and joint discriminative and generative modelling (Dai et al., 2022). The principle has also guided interpretable architectures, from ReduNet's unrolled optimisation (Chan et al., 2022) to sparse-rate-reduction transformers (Yu et al., 2023; 2024). More recently, Buchanan et al. (2026) further develop this line of research toward a mathematical theory of memory based on MCR^2.

We argue that one promising practical property of MCR^2 is its remarkable empirical robustness to label noise: it retains strong performance even when 50% of the training labels are corrupted (Yu et al., 2020). Yet MCR^2 was not designed specifically for noisy supervision, nor does it identify or correct mislabeled examples. Its robustness therefore appears to emerge from the coding-rate objective itself. This observation raises three questions: what gives rise to this robustness, when does it break down, and can these insights guide a better method for learning with noisy labels?

To the best of our knowledge, this is the first systematic investigation of MCR^2's robustness to label noise. Our main contributions are summarised as follows:

1. **MCR^2 remains robust when label errors are dispersed.** We trace this behaviour to label-independent global expansion, sample-wise normalisation, and the distribution of corrupted examples across different off-class directions. Under a balanced orthogonal-subspace model, the clean class subspace remains dominant even at substantial levels of uniform corruption.

2. **Concentrated errors reveal the limit of this robustness.** We characterise the competition between the clean subspace and the strongest off-class subspace through a clean-dominance gap. This explains why the structure of label corruption, rather than the noise rate alone, determines when class compression becomes unreliable.

3. **We propose Elastic Coding, a new method for robust learning with noisy labels.** Motivated by the identified failure mechanism, Elastic Coding uses a gradient decomposition to detect when class compression opposes global expansion and adjusts their balance accordingly. It requires neither clean labels nor a specified noise model. Extensive experiments with structured and real-world label noise show that Elastic Coding achieves the best performance among the compared methods in most evaluated settings.

2 Theory and Elastic Coding

The MCR^2 objective. MCR^2 learns features that separate classes while preserving structure within each class. Rather than requiring every class to collapse to a single point, it encourages each class to occupy a compact, low-dimensional subspace. Its objective combines global expansion, which spreads the features across the representation space, with class compression, which brings features assigned to the same class into a smaller subspace (Figure 1). This distinction matters under label noise: expansion is independent of class assignments, whereas compression uses the observed labels.

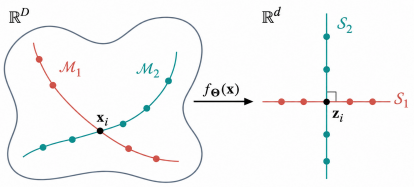


Figure 1: Illustration of MCR^2.

Let $\boldsymbol{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m] \in \mathbb{R}^{d \times m}$ be the features of m training examples from K classes. The class selector $\boldsymbol{\Pi}_k$ is diagonal, with a one for each example assigned to class k. We write $m_k = \operatorname{tr}(\boldsymbol{\Pi}_k)$ and $\boldsymbol{Z}_k \boldsymbol{Z}_k^\top = \boldsymbol{Z} \boldsymbol{\Pi}_k \boldsymbol{Z}^\top$. The coding-rate reduction

objective is

$$\Delta\mathcal{R}(\boldsymbol{Z},\boldsymbol{\Pi}) = \underbrace{\frac{1}{2}\log\det(\boldsymbol{I}_d + \alpha\boldsymbol{Z}\boldsymbol{Z}^\top)}_{\mathcal{R}(\boldsymbol{Z})} - \underbrace{\sum_{k=1}^{K}\frac{m_k}{2m}\log\det(\boldsymbol{I}_d + \alpha_k\boldsymbol{Z}_k\boldsymbol{Z}_k^\top)}_{\mathcal{R}_c(\boldsymbol{Z}|\boldsymbol{\Pi})}, \tag{1}$$

where $\alpha = d/(m\epsilon^2)$, $\alpha_k = d/(m_k\epsilon^2)$, and ϵ is the coding precision (Yu et al., 2020; Chan et al., 2022). The first term expands the representation as a whole. The second compresses each observed class. Their difference favours separated, low-dimensional class structure (Figure 1). The encoder output $\boldsymbol{h}_i = f_\theta(\boldsymbol{x}_i)$ is normalised sample by sample:

$$\boldsymbol{z}_i = \boldsymbol{h}_i/\|\boldsymbol{h}_i\|_2, \qquad \|\boldsymbol{z}_i\|_2 = 1. \tag{2}$$

With noisy labels, the model uses $\widetilde{y}_i$ and the corresponding selector $\widetilde{\boldsymbol{\Pi}}$ in Eq. (1). For fixed features, this changes class compression but leaves global expansion unchanged. An observed class may therefore contain features from several true classes.

2.1 Why dispersed label errors can be tolerated

We first examine why dispersed label errors need not erase the clean class signal. We begin with sample-wise normalisation, which prevents an individual mislabeled example from dominating through its norm. The matrix determinant lemma gives the following bound. Its proof is in Appendix A.1.

Proposition 1 (Bounded single-example influence)**.** *If $\boldsymbol{A} \succeq \boldsymbol{I}$, $\beta \geq 0$, and $\|\boldsymbol{z}\|_2 \leq 1$, the increment from adding $\boldsymbol{z}$ satisfies*

$$0 \leq \delta\mathcal{R}(\boldsymbol{z},\boldsymbol{A}) = \frac{1}{2}\log(1+\beta\boldsymbol{z}^\top\boldsymbol{A}^{-1}\boldsymbol{z}) \leq \frac{1}{2}\log(1+\beta). \tag{3}$$

The bound concerns one example. To study many errors, let $T_{kj} = \Pr(\widetilde{y} = j \mid y = k)$ be the label-transition matrix and let $\boldsymbol{\Sigma}_k = \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top \mid y = k]$ be the uncentered feature second moment of true class k. The second moment of observed class j is a mixture:

$$\widetilde{\boldsymbol{\Sigma}}_j = \sum_{k=1}^{K} q_{k|j}\boldsymbol{\Sigma}_k, \qquad q_{k|j} = \Pr(y = k \mid \widetilde{y} = j). \tag{4}$$

Thus, the question is how strongly each true-class direction contributes to an observed class. Under a tractable fixed-feature model, the rate-reduction signal has a simple form.

Proposition 2 (Transition-spectrum attenuation)**.** *Suppose classes are balanced, T is doubly stochastic, and $\boldsymbol{\Sigma}_k = \boldsymbol{P}_k/r$ for pairwise orthogonal rank-r projectors $\boldsymbol{P}_k$. Holding these moments fixed, the second-order expansion for small $\gamma = d/\epsilon^2$ is*

$$\Delta\mathcal{R}_T = \frac{\gamma^2}{4Kr}\left\|T - \frac{1}{K}\mathbf{1}\mathbf{1}^\top\right\|_F^2 + O(\gamma^3). \tag{5}$$

For forced symmetric flips, a label remains correct with probability $1-\eta$ and otherwise changes uniformly to one of the other $K-1$ classes.

Corollary 1 (Uniform-flip threshold)**.** *The second-order class signal is scaled by $\big(1-K\eta/(K-1)\big)^2$ and vanishes at $\eta = (K-1)/K$.*

At that threshold, an observed label is independent of the true class. Below it, each particular wrong class receives probability $\eta/(K-1)$, while the correct class receives $1-\eta$. Dispersed errors can therefore leave the clean direction stronger. These statements describe fixed features under the stated approximation. They do not guarantee that training recovers those features. Proofs and direct checks are in Appendices A.2 and B.

2.2 When an off-class subspace takes over

We next examine when concentrated errors can overturn the clean class direction. Consider an observed class with clean-aligned features $\boldsymbol{C}_j$ and contaminating features $\boldsymbol{N}_j$. Set $\alpha_j = d/(\widetilde{m}_j\epsilon^2)$ and $\boldsymbol{A}_j = \boldsymbol{I} + \alpha_j \boldsymbol{C}_j\boldsymbol{C}_j^\top$. The additional class coding rate is

$$\delta\mathcal{R}_c(\boldsymbol{N}_j) = \frac{1}{2}\log\det(\boldsymbol{I} + \alpha_j \boldsymbol{N}_j^\top \boldsymbol{A}_j^{-1}\boldsymbol{N}_j). \tag{6}$$

We call this increment *coding strain*. The inverse $\boldsymbol{A}_j^{-1}$ discounts directions already populated by $\boldsymbol{C}_j$. A new orthogonal direction is not discounted. Several mislabeled examples aligned in the same off-class direction can therefore create a competing subspace. Appendix D gives the directional decomposition.

Under the same orthogonal, isotropic model as Proposition 2, the competition is exact.

Proposition 3 (Observed-class subspace competition)**.** *For observed class j, $\widetilde{\boldsymbol{\Sigma}}_j$ has eigenvalue $q_{k|j}/r$ on the true-class subspace $\mathcal{S}_k = \mathrm{range}(\boldsymbol{P}_k)$, each with multiplicity r. Its class rate is*

$$\frac{1}{2}\log\det(\boldsymbol{I} + \gamma\widetilde{\boldsymbol{\Sigma}}_j) = \frac{r}{2}\sum_{k=1}^{K}\log\Big(1 + \frac{\gamma q_{k|j}}{r}\Big). \tag{7}$$

The matching subspace $\mathcal{S}_j$ is uniquely leading exactly when

$$g_j = \frac{1}{r}\left(q_{j|j} - \max_{k\neq j} q_{k|j}\right) > 0. \tag{8}$$

A positive gap means that the clean direction exceeds every individual competitor. A negative gap means that some off-class direction is stronger. For cyclic span-s noise, each flipped label is assigned to one of the next s classes.

Corollary 2 (Span-noise threshold)**.** *The gap is $g^{(s)} = (1 - \eta - \eta/s)/r$. The clean subspace is uniquely leading if and only if $\eta < s/(s+1)$.*

Spreading a fixed amount of noise over more classes weakens each competitor. The condition identifies the leading direction of a fixed representation, not the outcome of nonconvex training. Appendix B tests this prediction.

2.3 Elastic Coding: relaxing conflicting compression

The preceding analysis suggests a way to respond when class compression works against the clean structure. The clean-dominance gap uses true labels and cannot generally be measured during training. We instead compare the two gradients already present in Eq. (1). For an example i with observed label j, let $\boldsymbol{G} = \boldsymbol{I} + \alpha\boldsymbol{Z}\boldsymbol{Z}^\top$ and $\boldsymbol{C}_j = \boldsymbol{I} + \alpha_j\widetilde{\boldsymbol{Z}}_j\widetilde{\boldsymbol{Z}}_j^\top$. Then

$$\nabla_{\boldsymbol{z}_i}\Delta\mathcal{R} = \alpha(\boldsymbol{G}^{-1} - \boldsymbol{C}_j^{-1})\boldsymbol{z}_i. \tag{9}$$

The first term expands features independently of labels. The second compresses examples grouped by their observed label. Normalisation determines which part of these updates reaches the encoder. With $\boldsymbol{z}_i = \boldsymbol{h}_i/\|\boldsymbol{h}_i\|_2$, the chain rule gives

$$\nabla_{\boldsymbol{h}_i}\Delta\mathcal{R} = \frac{1}{\|\boldsymbol{h}_i\|_2}(\boldsymbol{I} - \boldsymbol{z}_i\boldsymbol{z}_i^\top)\nabla_{\boldsymbol{z}_i}\Delta\mathcal{R}. \tag{10}$$

The projection removes the radial component, which would change the feature norm rather than its direction. We therefore compare expansion and compression after the same tangent projection. Writing $\mathcal{P}_{\boldsymbol{z}_i}(\boldsymbol{g}) = (\boldsymbol{I} - \boldsymbol{z}_i\boldsymbol{z}_i^\top)\boldsymbol{g}$, their feature-level ascent directions are

$$\boldsymbol{g}_{E,i}^{\tan} = \mathcal{P}_{\boldsymbol{z}_i}(\alpha\boldsymbol{G}^{-1}\boldsymbol{z}_i), \qquad \boldsymbol{g}_{C,i}^{\tan} = -\mathcal{P}_{\boldsymbol{z}_i}(\alpha\boldsymbol{C}_j^{-1}\boldsymbol{z}_i). \tag{11}$$

The minus sign comes from subtracting the class rate in the objective. Both vectors are computed from the same minibatch and require only observed labels. Figure 2 illustrates the projection and their possible conflict.

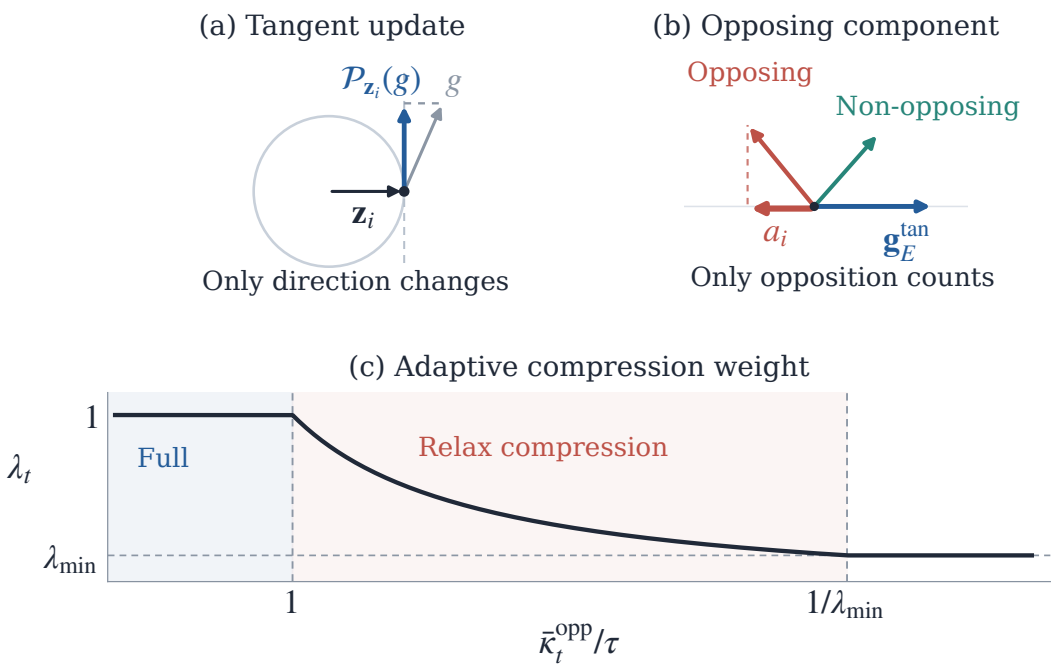


Figure 2: Elastic Coding mechanism. (a) Tangent update. (b) Opposing compression. (c) Adaptive compression weight.

We measure only compression that points against expansion. With a small numerical constant $\delta > 0$, define

$$s_{i,t} = \frac{\langle \boldsymbol{g}_{E,i}^{\tan}, \boldsymbol{g}_{C,i}^{\tan} \rangle}{\|\boldsymbol{g}_{E,i}^{\tan}\|_2 \|\boldsymbol{g}_{C,i}^{\tan}\|_2 + \delta}, \tag{12}$$

$$a_{i,t} = \|\boldsymbol{g}_{C,i}^{\tan}\|_2 [-s_{i,t}]_+, \tag{13}$$

$$\kappa_t^{\text{opp}} = \frac{\|(a_{1,t}, \ldots, a_{m,t})\|_2}{\|\boldsymbol{G}_E^{\tan}\|_F + \delta}. \tag{14}$$

Here $\boldsymbol{G}_E^{\tan}$ collects the tangent expansion gradients for the minibatch, $\boldsymbol{a}_t = (a_{1,t}, \ldots, a_{m,t})^\top$, and $[x]_+ = \max(x, 0)$. Compatible and orthogonal compression do not increase κ_t^{opp}. This distinction avoids weakening a large compression update simply because of its magnitude: the update must also oppose expansion. We smooth minibatch fluctuations using

$$\bar{\kappa}_t^{\text{opp}} = \beta \bar{\kappa}_{t-1}^{\text{opp}} + (1 - \beta)\kappa_t^{\text{opp}}, \tag{15}$$

where $\beta \in [0, 1)$ sets the amount of smoothing. The compression weight and training objective are

$$\lambda_t = \text{clip}\left(\frac{\tau}{\bar{\kappa}_t^{\text{opp}}}, \lambda_{\min}, 1\right), \qquad \Delta\mathcal{R}_{\text{E}} = \mathcal{R} - \lambda_t \mathcal{R}_c. \tag{16}$$

The weight is held constant during the encoder update. When opposition is weak, $\lambda_t = 1$ and the objective is ordinary MCR^2. Stronger opposition reduces class compression while retaining global expansion. The floor $\lambda_{\min}$ keeps compression active.

Proposition 4 (Control of opposing compression)**.** *If the moving-average lag is ignored and* $\lambda_{\min} = 0$*, then*

$$\lambda_t \|(a_{1,t}, \ldots, a_{m,t})\|_2 \le \tau \|\boldsymbol{G}_E^{\tan}\|_F + \tau\delta. \tag{17}$$

The proof is in Appendix A.8. The bound describes the simplified rule. With a moving average, the weight responds to recent conflict rather than only the current minibatch. A positive floor also prevents compression from being reduced arbitrarily. The proposition therefore does not guarantee the same bound for every update of the full training procedure.

Training procedure. Each iteration first computes normalised features and the two coding rates. It then obtains their tangent gradients, measures their alignment, and updates the smoothed conflict ratio. The resulting weight is held fixed while differentiating the weighted objective with respect to the encoder parameters. This prevents the encoder update from also differentiating through the rule that selects λ_t. The threshold τ determines how much opposing compression is tolerated, while $\lambda_{\min}$ preserves a minimum amount of class compression. If the smoothed conflict is zero, the weight is set to one, consistent with the limit of Eq. (16).

Elastic Coding does not identify which individual labels are wrong. Its signal measures a conflict between two objective terms, so it can also respond to difficult clean examples. The intended effect is to preserve global expansion while limiting compression that works against it. Appendix D gives the gradient derivation and a class-adaptive extension. The experiments below evaluate the global weighting rule.

3 Experiments

We examine the effects of per-sample norm constraints and label-transition structure on MCR^2 robustness, and evaluate Elastic Coding under structured and real-world label noise. A random label flip assigns a selected sample a label drawn uniformly from the other classes.

3.1 Evaluation metric

We first describe how we assess whether representations trained with noisy labels retain true class structure. We use *clean-centroid probe accuracy*. After training, we freeze the encoder and use clean training labels only for evaluation to compute each true-class centroid:

$$\boldsymbol{\mu}_k = \frac{1}{|\{i : y_i = k\}|} \sum_{i:y_i=k} \boldsymbol{z}_i, \qquad \widehat{y}(\boldsymbol{x}) = \arg\max_k \frac{f_\theta(\boldsymbol{x})^\top \boldsymbol{\mu}_k}{\|f_\theta(\boldsymbol{x})\|_2 \|\boldsymbol{\mu}_k\|_2}. \tag{18}$$

We then measure accuracy on the cleanly labeled test set. Clean labels do not enter MCR^2 training. This post-training diagnostic measures whether true semantic classes remain separable in feature space. It does not measure accuracy in predicting the original class names directly from noisy training labels.

For the CIFAR-10 study below, we also form centroids using only the observed training labels $\widetilde{y}_i$. We call the resulting readout *observed-centroid accuracy*. It is measured against observed labels on the training set and clean labels on the test set. We report it alongside clean-centroid probe accuracy because the two can differ sharply when noisy labels associate a class name with the wrong feature cluster.

Checks of the proposed mechanism. Fixed-feature MNIST experiments support the predicted decay of class signal under random flips and the change in the leading subspace as errors become more concentrated. At 80% cyclic noise, the empirical clean-dominance gap crosses near the predicted span $s = 4$. A low-rate CIFAR-10 control also shows that concentrated errors do not by themselves cause failure. The settings, measurements, and limitations of these checks are in Appendix B.

3.2 Elastic Coding results on CIFAR-10

We now test whether Elastic Coding mitigates representation degradation under structured label noise on CIFAR-10. Experiments use ResNet-18 with $d = 512$. We compare original MCR^2 and Elastic Coding at noise rates from 10% to 90% in steps of 10%. A selected label is changed to one of the next two classes in cyclic order, with equal probability. Both methods use the same ResNet-18 encoder, batch size 1000, seed 42, and 100 epochs. We use Adam with initial learning rate 10^{-3}, weight decay 5×10^{-4}, and a loss-based learning-rate reduction with factor 0.3 and patience 5. The test set is not used for parameter selection. Elastic Coding sets the compression weight using the opposing compression component measured during training.

Table 1: CIFAR-10 observed-centroid and clean-centroid probe accuracy (%) under span-2 label noise.

Noise rate	10%	20%	30%	40%	50%	60%	70%	80%	90%
MCR^2									
Final λ	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Obs. centroid train	82.24	68.90	59.36	49.63	38.84	31.85	29.13	34.47	36.91
Obs. centroid test	86.79	81.73	80.43	78.06	70.34	60.26	**14.27**	3.46	**3.94**
Clean-centroid probe	86.81	81.55	80.12	78.06	70.63	67.20	66.13	76.69	71.71
Elastic Coding									
Final λ	0.543	0.493	0.483	0.514	0.498	0.502	0.531	0.518	0.531
Obs. centroid train	**84.79**	**75.38**	**64.75**	**56.11**	**46.26**	**36.76**	**33.80**	**38.09**	**43.46**
Obs. centroid test	**89.22**	**88.23**	**86.37**	**85.53**	**82.78**	**75.79**	11.30	**4.46**	2.28
Clean-centroid probe	**89.71**	**88.19**	**86.34**	**86.00**	**83.61**	**83.02**	**83.32**	**83.47**	**85.48**

Table 1 shows that Elastic Coding improves both observed-centroid test accuracy and clean-centroid probe accuracy from 10% to 60% noise. At 50% noise, the respective gains over MCR^2 are 12.44 and 12.98 percentage points. They increase to 15.53 and 15.82 points at 60%. The agreement between the two readouts indicates that the improvement is accompanied by a clearer clean-class structure, rather than arising only from the choice of observed-label centroids.

The two readouts separate sharply at higher noise rates. At 70% to 90% noise, observed-centroid test accuracy is low for both methods, while Elastic Coding retains clean-centroid probe accuracy above 83%. In the idealised span-2 mixture, an observed class contains a matching component of weight $1-\eta$ and two off-class components of weight $\eta/2$ each. An individual wrong-class component overtakes the matching one when $\eta > 2/3$. The collapse of observed-centroid accuracy near 70% is consistent with this change in class-name alignment. It does not imply that clean semantic structure has disappeared. Indeed, Elastic Coding has higher clean-centroid probe accuracy than MCR^2 even at 70% and 90%, where its observed-centroid test accuracy is lower. The clean-centroid probe uses unavailable true training labels, so this remaining structure does not by itself provide a usable noisy-label classifier.

3.3 Representation geometry

We also examine how the learned features are organised under structured noise. Figure 3 shows the final training features at 50% span-2 noise. The same $2{,}000$ images are coloured by true and observed labels within each method. Elastic Coding shows clearer grouping by true class, while observed labels remain mixed within several groups. Recolouring the same projection therefore separates preservation of semantic structure from agreement with the corrupted class names. The two methods use separate t-SNE projections, so distances between their panels are not directly comparable. In the original 512-dimensional space, a clean-reference centroid recovers the true class of 86.00% of mislabeled training images for Elastic Coding, versus 72.52% for MCR^2. This higher-dimensional check supports the visual pattern without relying on t-SNE distances. It uses synthetic ground truth and is not a deployable classifier. The 80% visualisation and full protocol appear in Appendix C.

3.4 Broader evaluation with concentrated and real-world label noise

We next examine whether the CIFAR-10 findings extend to other datasets and naturally noisy labels. We report standard test accuracy using each method's usual readout and observed-centroid test accuracy using observed-label centroids for all methods. For the real-noise datasets, we also report linear-probe test accuracy.

For synthetic noise, we use the 20 coarse classes of CIFAR-100 (CIFAR-20) (Krizhevsky, 2009) and the ten-class Imagenette subset (Howard, 2019). At each noise rate, approximately 10% or 20% of training labels are changed. A selected label k is replaced by either $(k+1) \bmod K$ or $(k+2) \bmod K$, with equal probability. This span-2 protocol concentrates errors in two other class indices. The indices do not represent semantic similarity, so this is a controlled simulation of concentrated errors rather than a model of which categories people confuse.

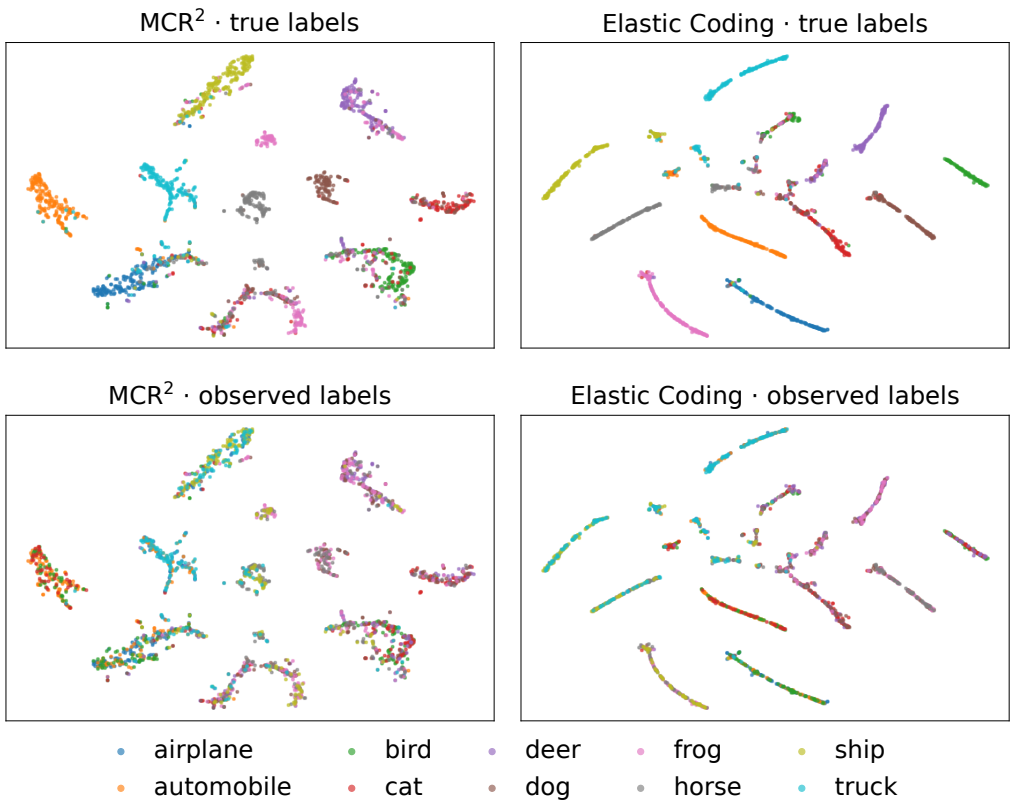


Figure 3: CIFAR-10 training representations at 50% span-2 label noise. Columns compare methods, and rows show true or observed labels.

Table 2: Test accuracy (%) under structured and real-world label noise using standard, observed-centroid, and linear-probe readouts. Gain compares Elastic Coding with the best baseline in each row.

Noise	Metric	CE	GCE	SCE	APL	AEL	ANL	JAL	VBL	MCR²	Elastic Coding	Gain
CIFAR-20 (span-2 noise)												
10%	Standard test	72.07	63.23	71.37	69.50	17.16	52.94	54.26	70.60	69.44	**73.47**	+1.40
	Obs. centroid	72.39	63.38	71.38	69.51	17.66	54.11	55.38	70.64	69.44	**73.47**	+1.08
20%	Standard test	65.16	61.31	65.69	66.66	12.35	31.77	50.36	67.75	67.68	**71.11**	+3.36
	Obs. centroid	67.12	61.60	65.99	66.64	14.09	33.36	51.59	67.70	67.68	**71.11**	+3.41
Imagenette (span-2 noise)												
10%	Standard test	**79.57**	75.90	73.56	68.31	45.23	35.50	68.87	69.33	75.57	79.24	-0.33
	Obs. centroid	**80.11**	76.18	73.64	68.57	46.91	42.35	69.36	69.79	75.57	79.24	-0.87
20%	Standard test	**77.92**	74.73	72.01	58.96	37.74	24.59	59.52	59.42	74.78	75.90	-2.02
	Obs. centroid	**78.68**	74.83	71.90	59.88	40.01	34.71	60.80	60.62	74.78	75.90	-2.78
ANIMAL-10N (original labels)												
–	Standard test	78.28	76.50	77.60	54.68	41.12	40.86	61.54	66.86	78.12	**81.88**	+3.60
	Obs. centroid	78.38	76.50	77.52	55.64	42.64	46.76	61.84	67.12	78.12	**81.88**	+3.50
	Linear probe	78.28	76.26	77.92	55.00	42.36	47.94	61.66	67.00	78.44	**82.12**	+3.68
Clothing1M-100K (original labels)												
–	Standard test	60.95	46.30	60.58	60.38	14.55	57.46	57.26	60.64	64.18	**65.38**	+1.20
	Obs. centroid	61.97	48.27	60.98	62.05	14.65	61.29	57.96	61.92	64.18	**65.38**	+1.20
	Linear probe	61.10	47.79	60.89	60.88	14.41	58.01	57.40	60.91	62.24	**65.68**	+3.44

We also evaluate on ANIMAL-10N (Song et al., 2019) and Clothing1M (Xiao et al., 2015) using their observed training labels without additional corruption. The ANIMAL-10N data come from an Activeloop mirror and contain 50,000 training and 5,000 filtered test images at 64×64 resolution. Their class counts match the authors' dataset description. For Clothing1M, we use a 64×64 Kaggle derivative, sample 100,000 training images proportionally by observed class label with seed 42, and retain its 10,526-image test split. This Clothing subset is a smaller, resized case study, so its accuracies should not be read as results on the original million-image benchmark.

All ten methods use the same ResNet-18 encoder and $d = 512$ features. We compare MCR^2 and Elastic Coding with eight classification-loss baselines: cross-entropy (CE), generalized cross-entropy (GCE) (Zhang & Sabuncu, 2018), symmetric cross-entropy (SCE) (Wang et al., 2019), active-passive loss (APL) (Ma et al., 2020), asymmetric exponential loss (AEL) (Zhou et al., 2021), active negative loss (ANL) (Ye et al., 2024), joint asymmetric loss (JAL) (Wang et al., 2025), and variation-bounded loss (VBL) (Wang et al., 2026). Each setting uses Adam for 100 full epochs, weight decay 5×10^{-4}, and seed 42. The learning rate follows cosine annealing to 10^{-5} over 100 epochs on the synthetic datasets and reaches that floor after 80 epochs on the real-noise datasets. Its initial value is 5×10^{-4} for the coding methods and 10^{-3} for the classification baselines. Batch size is 1000 on CIFAR-20 and 400 on the other datasets.

For standard test accuracy, MCR^2 and Elastic Coding use centroids fitted with observed training labels, while the eight classification-loss baselines use their trained classifiers. Observed-centroid accuracy uses the same centroid readout for every method's frozen features. On the real-noise datasets, linear-probe accuracy uses a ridge-regularised probe. None of these readouts uses true training labels.

On CIFAR-20, Elastic Coding leads all ten methods under both standard and observed-centroid readouts at 10% and 20% span-2 noise. Its standard test advantage over the strongest baseline grows from 1.40 to 3.36 percentage points as the noise rate increases. Imagenette gives a more qualified result. Elastic Coding still exceeds MCR^2 by 3.67 points at 10% noise and 1.12 points at 20%, but CE leads the standard test comparison at both rates. CE also leads under the common observed-centroid readout. Thus, the benefit over the original coding-rate objective does not imply a uniform advantage over classification losses. On ANIMAL-10N and Clothing1M-100K, Elastic Coding leads under all three reported readouts. Its standard test accuracy is 81.88% on ANIMAL-10N and 65.38% on Clothing1M-100K, exceeding the strongest baseline by 3.60 and 1.20 points. The corresponding linear-probe advantages are 3.68 and 3.44 points. On Clothing1M-100K, the larger linear-probe margin suggests that some of the representation advantage is more visible with a flexible readout than with observed-label centroids. These results extend the comparison beyond synthetic corruption, while the Clothing1M-100K result remains specific to the resized subset used here.

Across the six dataset–noise settings in Table 2, Elastic Coding leads in four under both standard and observed-centroid readouts. The agreement across these readouts supports an improvement in the learned features as well as in the method's usual prediction rule. The Imagenette results also show that the size and ranking of the gain depend on the dataset.

4 Conclusion

We studied why MCR^2 tolerates label noise and when this robustness weakens. Its label-independent global expansion continues to act on the full dataset, while sample-wise normalisation limits the influence of individual examples. When label errors are dispersed, no single off-class direction need overpower the clean subspace. Under our fixed-feature model, the clean-dominance gap makes this competition explicit and shows why the structure of corruption matters alongside its rate. Motivated by this analysis, we proposed Elastic Coding, which compares the tangent updates from expansion and class compression and relaxes compression when their conflict grows. It requires neither clean labels nor a specified noise transition. Across the evaluated datasets, it improves on MCR^2 in many settings and leads the compared methods in most evaluated settings. At severe structured noise, clean class structure can persist even when observed labels no longer identify those classes reliably. This distinction motivates evaluating both feature geometry and class-name prediction. Extending the analysis to learned, overlapping subspaces and testing on broader datasets are natural next steps.

AI USE STATEMENT

GPT-5.6 Sol and GPT-6 Astra were used to assist with language polishing, checking mathematical proofs, and writing experimental code. The authors reviewed these contributions and remain fully responsible for the accuracy of the proofs, code, experimental results, claims, citations, and final manuscript.

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.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 242–252. PMLR, 2019.

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.

Bowen Baker, Otkrist Gupta, Nikhil Naik, and Ramesh Raskar. Designing neural network architectures using reinforcement learning. In *International Conference on Learning Representations*, 2017.

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.

Jacopo Cavazza, Pietro Morerio, Benjamin Haeffele, Connor Lane, Vittorio Murino, and Rene Vidal. Dropout as a low-rank regularizer for matrix factorization. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84, 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.

Tianzhe Chu, Shengbang Tong, Tianjiao Ding, Xili Dai, Benjamin D. Haeffele, René Vidal, and Yi Ma. Image clustering via the principle of rate reduction in the age of pretrained models. In *International Conference on Learning Representations*, 2024.

Xili Dai, Shengbang Tong, Mingyang Li, Ziyang Wu, Michael Psenka, Kwan Ho Ryan Chan, Pengyuan Zhai, Yaodong Yu, Xiaojun Yuan, Heung-Yeung Shum, and Yi Ma. CTRL: Closed-loop transcription to an LDR via minimaxing rate reduction. *Entropy*, 24(4):456, 2022. doi: 10.3390/e24040456.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Guilherme N. DeSouza and Avinash C. Kak. Vision for mobile robot navigation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):237–267, 2002.

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, pp. 1802–1811, 2019.

Kaisei Fukaya, Damon Daylamani-Zad, and Harry Agius. Evaluation metrics for intelligent generation of graphical game assets: a systematic survey-based framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):7998–8017, 2024.

Gemini Team, 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.

Rick Groenendijk, Leo Dorst, and Theo Gevers. Geometric back-propagation in morphological neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):14045–14051, 2023.

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

Jeremy Howard. Imagenette: A smaller subset of 10 easily classified classes from ImageNet. Dataset, 2019.

Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren (eds.). *Automated Machine Learning: Methods, Systems, Challenges*. Springer, 2019.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. GPT-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270):20230254, 2024.

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.

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

Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 6543–6553, 2020.

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*, 2019.

Kwanseok Oh, Jee Seok Yoon, and Heung-Il Suk. Learn-explain-reinforce: Counterfactual reasoning and its guidance to reinforce an Alzheimer's disease diagnosis model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4843–4857, 2023.

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

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, October 1986.

Hwanjun Song, Minseok Kim, and Jae-Gil Lee. SELFIE: Refurbishing unclean samples for robust deep learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 5907–5915, 2019.

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.

Shengbang Tong, Xili Dai, Yubei Chen, Mingyang Li, Zengyi Li, Brent Yi, Yann LeCun, and Yi Ma. Unsupervised learning of structured representations via closed-loop transcription. *arXiv preprint arXiv:2210.16782*, 2022.

Jialiang Wang, Xianming Liu, Xiong Zhou, Gangfeng Hu, Deming Zhai, Junjun Jiang, and Xiangyang Ji. Joint asymmetric loss for learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1947–1956, 2025.

Jialiang Wang, Xiong Zhou, Xianming Liu, Gangfeng Hu, Deming Zhai, Junjun Jiang, and Haoliang Li. Variation-bounded loss for noise-tolerant learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26251–26259, 2026.

Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 322–330, 2019.

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

Peter R. Wurman, Samuel Barrett, Kenta Kawamoto, James MacGlashan, Kaushik Subramanian, Thomas J. Walsh, Roberto Capobianco, Alisa Devlic, Franziska Eckert, Florian Fuchs, et al. Outracing champion Gran Turismo drivers with deep reinforcement learning. *Nature*, 602(7896): 223–228, 2022.

Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2691–2699, 2015.

Xichen Ye, Yifan Wu, Yiqi Wang, Xiaoqiang Li, Weizhong Zhang, and Yifan Chen. Active negative loss: A robust framework for learning with noisy labels. *arXiv preprint arXiv:2412.02373*, 2024.

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, pp. 9422–9434, 2020.

Yaodong Yu, Sam Buchanan, Druv Pai, Tianzhe Chu, Ziyang Wu, Shengbang Tong, Benjamin Haeffele, and Yi Ma. White-box transformers via sparse rate reduction. In *Advances in Neural Information Processing Systems*, volume 36, pp. 9422–9457, 2023.

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.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Xiao Zhang, Yaodong Yu, Lingxiao Wang, and Quanquan Gu. Learning one-hidden-layer ReLU networks via gradient descent. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89, pp. 1524–1534. PMLR, 2019.

Zhilu Zhang and Mert R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Xiong Zhou, Xianming Liu, Junjun Jiang, Xin Gao, and Xiangyang Ji. Asymmetric loss functions for learning with noisy labels. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 12846–12856, 2021.

Barret Zoph and Quoc V. Le. Neural architecture search with reinforcement learning. In *International Conference on Learning Representations*, 2017.

A Detailed proofs of propositions and corollaries

A.1 Proof of the single-sample coding-rate influence bound

We begin by bounding the change in coding rate caused by one feature.

Proof of Proposition 1. Since $\boldsymbol{A} \succeq \boldsymbol{I}$, every eigenvalue of $\boldsymbol{A}$ is at least 1. Thus, $\boldsymbol{A}$ is positive definite and invertible. First, using $\log a - \log b = \log(a/b)$ and $\det(\boldsymbol{A}^{-1}\boldsymbol{B}) = \det(\boldsymbol{B})/\det(\boldsymbol{A})$, we obtain

$$\begin{aligned}\delta\mathcal{R}(\boldsymbol{z}, \boldsymbol{A}) &= \frac{1}{2}\log\frac{\det(\boldsymbol{A} + \beta\boldsymbol{z}\boldsymbol{z}^\top)}{\det \boldsymbol{A}} \\ &= \frac{1}{2}\log\det\big[\boldsymbol{A}^{-1}(\boldsymbol{A} + \beta\boldsymbol{z}\boldsymbol{z}^\top)\big] \\ &= \frac{1}{2}\log\det\big(\boldsymbol{I} + \beta\boldsymbol{A}^{-1}\boldsymbol{z}\boldsymbol{z}^\top\big)\,.\end{aligned}$$

The second line holds because $\det[\boldsymbol{A}^{-1}(\boldsymbol{A} + \beta\boldsymbol{z}\boldsymbol{z}^\top)] = \det(\boldsymbol{A}^{-1})\det(\boldsymbol{A} + \beta\boldsymbol{z}\boldsymbol{z}^\top)$. Next, use the matrix determinant lemma $\det(\boldsymbol{I} + \boldsymbol{u}\boldsymbol{v}^\top) = 1 + \boldsymbol{v}^\top\boldsymbol{u}$. Setting $\boldsymbol{u} = \beta\boldsymbol{A}^{-1}\boldsymbol{z}$ and $\boldsymbol{v} = \boldsymbol{z}$ gives

$$\det\big(\boldsymbol{I} + \beta\boldsymbol{A}^{-1}\boldsymbol{z}\boldsymbol{z}^\top\big) = 1 + \beta\boldsymbol{z}^\top\boldsymbol{A}^{-1}\boldsymbol{z},$$

which proves the equality in Eq. (3).

Finally, let $\boldsymbol{A} = \boldsymbol{U}\,\mathrm{diag}(a_1, \ldots, a_d)\boldsymbol{U}^\top$ be an eigendecomposition. Since $a_\ell \geq 1$, we have $0 < a_\ell^{-1} \leq 1$, hence $\boldsymbol{0} \prec \boldsymbol{A}^{-1} \preceq \boldsymbol{I}$. Therefore, $0 \leq \boldsymbol{z}^\top\boldsymbol{A}^{-1}\boldsymbol{z} \leq \boldsymbol{z}^\top\boldsymbol{z} = \|\boldsymbol{z}\|_2^2 \leq 1$. As $\beta \geq 0$, we further have $1 \leq 1 + \beta\boldsymbol{z}^\top\boldsymbol{A}^{-1}\boldsymbol{z} \leq 1 + \beta$. Monotonicity of $\log(x)$ on the positive reals then gives $0 \leq \delta\mathcal{R}(\boldsymbol{z}, \boldsymbol{A}) \leq \frac{1}{2}\log(1 + \beta)$. □

A.2 Proof of transition-spectrum decay under the second-order approximation

We next derive how label transitions attenuate the fixed-feature rate-reduction signal.

Proof of Proposition 2. The proof has four steps.

Step 1: Express posterior mixture weights through T. Balanced true classes imply $\Pr(y = k) = 1/K$. By Bayes' rule,

$$\begin{aligned}q_{k|j} &= \Pr(y = k \mid \widetilde{y} = j) \\ &= \frac{\Pr(\widetilde{y} = j \mid y = k)\Pr(y = k)}{\sum_{\ell=1}^{K}\Pr(\widetilde{y} = j \mid y = \ell)\Pr(y = \ell)} \\ &= \frac{T_{kj}/K}{K^{-1}\sum_{\ell=1}^{K} T_{\ell j}} = T_{kj},\end{aligned}$$

The last step uses the fact that every column of T sums to 1. The observed classes are also balanced: $\Pr(\widetilde{y} = j) = K^{-1}\sum_k T_{kj} = 1/K$. Thus,

$$\widetilde{\boldsymbol{\Sigma}}_j = \sum_{k=1}^{K} T_{kj}\boldsymbol{\Sigma}_k, \qquad \boldsymbol{\Sigma} = \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\Sigma}_k = \frac{1}{K}\sum_{j=1}^{K}\widetilde{\boldsymbol{\Sigma}}_j.$$

Step 2: Expand the log determinant. Let $F_\gamma(\boldsymbol{\Sigma}) = \frac{1}{2}\log\det(\boldsymbol{I} + \gamma\boldsymbol{\Sigma})$. Using $\log\det \boldsymbol{B} = \operatorname{tr}(\log \boldsymbol{B})$ and $\log(\boldsymbol{I} + \gamma\boldsymbol{\Sigma}) = \gamma\boldsymbol{\Sigma} - \gamma^2\boldsymbol{\Sigma}^2/2 + O(\gamma^3)$, we have

$$F_\gamma(\boldsymbol{\Sigma}) = \frac{\gamma}{2}\operatorname{tr}(\boldsymbol{\Sigma}) - \frac{\gamma^2}{4}\operatorname{tr}(\boldsymbol{\Sigma}^2) + O(\gamma^3).$$

At the population level, the noisy MCR^2 objective is

$$\Delta\mathcal{R}_T = F_\gamma(\boldsymbol{\Sigma}) - \frac{1}{K}\sum_{j=1}^{K} F_\gamma(\widetilde{\boldsymbol{\Sigma}}_j).$$

Because $\boldsymbol{\Sigma} = K^{-1}\sum_j \widetilde{\boldsymbol{\Sigma}}_j$, the first-order terms cancel exactly, giving

$$\Delta\mathcal{R}_T = \frac{\gamma^2}{4}\left[\frac{1}{K}\sum_{j=1}^{K}\operatorname{tr}(\widetilde{\boldsymbol{\Sigma}}_j^2) - \operatorname{tr}(\boldsymbol{\Sigma}^2)\right] + O(\gamma^3).$$

Step 3: Compute the two quadratic traces using subspace orthogonality. From $\boldsymbol{\Sigma}_k = \boldsymbol{P}_k/r$, $\boldsymbol{P}_k^2 = \boldsymbol{P}_k$, $\operatorname{tr}(\boldsymbol{P}_k) = r$, and $\boldsymbol{P}_k\boldsymbol{P}_\ell = \boldsymbol{0}$ for $k \neq \ell$, we obtain

$$\operatorname{tr}(\boldsymbol{\Sigma}_k\boldsymbol{\Sigma}_\ell) = \begin{cases} 1/r, & k = \ell, \\ 0, & k \neq \ell. \end{cases}$$

Hence, for every observed class j,

$$\begin{aligned}\operatorname{tr}(\widetilde{\boldsymbol{\Sigma}}_j^2) &= \operatorname{tr}\left[\left(\sum_k T_{kj}\boldsymbol{\Sigma}_k\right)\left(\sum_\ell T_{\ell j}\boldsymbol{\Sigma}_\ell\right)\right] \\ &= \sum_{k,\ell} T_{kj}T_{\ell j}\operatorname{tr}(\boldsymbol{\Sigma}_k\boldsymbol{\Sigma}_\ell) = \frac{1}{r}\sum_k T_{kj}^2.\end{aligned}$$

Therefore,

$$\frac{1}{K}\sum_j \operatorname{tr}(\widetilde{\boldsymbol{\Sigma}}_j^2) = \frac{\|T\|_F^2}{Kr}.$$

Similarly,

$$\operatorname{tr}(\boldsymbol{\Sigma}^2) = \frac{1}{K^2}\sum_{k,\ell}\operatorname{tr}(\boldsymbol{\Sigma}_k\boldsymbol{\Sigma}_\ell) = \frac{1}{Kr}.$$

Substituting into Step 2 yields

$$\Delta\mathcal{R}_T = \frac{\gamma^2}{4Kr}(\|T\|_F^2 - 1) + O(\gamma^3).$$

Step 4: Rewrite the result as distance from the fully mixed matrix. Let $\boldsymbol{J} = K^{-1}\boldsymbol{1}\boldsymbol{1}^\top$. Because T is doubly stochastic, $\sum_{k,j} T_{kj} = K$, so

$$\|\boldsymbol{J}\|_F^2 = 1, \qquad \langle T, \boldsymbol{J}\rangle_F = \frac{1}{K}\sum_{k,j} T_{kj} = 1.$$

Consequently,

$$\begin{aligned}\|T - \boldsymbol{J}\|_F^2 &= \|T\|_F^2 + \|\boldsymbol{J}\|_F^2 - 2\langle T, \boldsymbol{J}\rangle_F \\ &= \|T\|_F^2 - 1.\end{aligned}$$

Substitution gives Eq. (5). □

A.3 Proof of the critical point for symmetric noise

We now specialise the transition-spectrum result to uniformly chosen wrong labels.

Proof of Corollary 1. Let $\boldsymbol{J} = K^{-1}\mathbf{1}\mathbf{1}^\top$,and denote $c = 1 - K\eta/(K-1)$. First, rearrange the transition matrix:

$$\begin{aligned} T &= (1-\eta)\boldsymbol{I} + \frac{\eta}{K-1}(\mathbf{1}\mathbf{1}^\top - \boldsymbol{I}) \\ &= \left(1 - \frac{K}{K-1}\eta\right)\boldsymbol{I} + \frac{\eta}{K-1}\mathbf{1}\mathbf{1}^\top. \end{aligned}$$

On the other hand, $-c/K = \eta/(K-1) - 1/K$, so

$$T - \boldsymbol{J} = c\boldsymbol{I} + \left(\frac{\eta}{K-1} - \frac{1}{K}\right)\mathbf{1}\mathbf{1}^\top = c(\boldsymbol{I} - \boldsymbol{J}).$$

The matrix $\boldsymbol{I} - \boldsymbol{J}$ is an orthogonal projector of rank $K-1$, so

$$\|\boldsymbol{I} - \boldsymbol{J}\|_F^2 = \operatorname{tr}[(\boldsymbol{I} - \boldsymbol{J})^2] = \operatorname{tr}(\boldsymbol{I} - \boldsymbol{J}) = K - 1.$$

Therefore,

$$\|T - \boldsymbol{J}\|_F^2 = c^2(K-1).$$

Clean labels correspond to $\eta = 0$ and have spectral energy $K-1$. Thus, relative to clean labels, the second-order signal in Proposition 2 is multiplied by $c^2 = (1 - K\eta/(K-1))^2$. Setting $c = 0$ gives $\eta = (K-1)/K$. For $K = 10$, the critical point is 0.9 for MNIST. □

A.4 Derivation of the coding-strain expression

We derive the additional class coding rate introduced by contaminating features.

For observed class j, hold $\widetilde{m}_j$ fixed and let $\boldsymbol{A}_j = \boldsymbol{I}_d + \alpha_j \boldsymbol{C}_j \boldsymbol{C}_j^\top$. Since $\widetilde{\boldsymbol{Z}}_j = [\boldsymbol{C}_j, \boldsymbol{N}_j]$, the increase in its coding rate is

$$\begin{aligned} \delta\mathcal{R}_c(\boldsymbol{N}_j) &= \frac{1}{2}\log\det\left(\boldsymbol{I}_d + \alpha_j \widetilde{\boldsymbol{Z}}_j \widetilde{\boldsymbol{Z}}_j^\top\right) - \frac{1}{2}\log\det \boldsymbol{A}_j \\ &= \frac{1}{2}\log\frac{\det(\boldsymbol{A}_j + \alpha_j \boldsymbol{N}_j \boldsymbol{N}_j^\top)}{\det \boldsymbol{A}_j} \\ &= \frac{1}{2}\log\det\left(\boldsymbol{I}_d + \alpha_j \boldsymbol{A}_j^{-1}\boldsymbol{N}_j \boldsymbol{N}_j^\top\right) \\ &= \frac{1}{2}\log\det\left(\boldsymbol{I} + \alpha_j \boldsymbol{N}_j^\top \boldsymbol{A}_j^{-1}\boldsymbol{N}_j\right). \end{aligned}$$

The final step uses Sylvester's identity, $\det(\boldsymbol{I} + \boldsymbol{U}\boldsymbol{V}) = \det(\boldsymbol{I} + \boldsymbol{V}\boldsymbol{U})$. Because $\boldsymbol{A}_j$ is positive definite, its inverse and logarithmic determinant are well defined. The identity matrix in the last line has dimension equal to the number of columns of $\boldsymbol{N}_j$.

A.5 Proof of the observed-class subspace competition spectrum

We use the observed-class second moment to compare clean and off-class directions.

Proof of Proposition 3. From Eq. (4) and $\boldsymbol{\Sigma}_k = \boldsymbol{P}_k/r$,

$$\widetilde{\boldsymbol{\Sigma}}_j = \frac{1}{r}\sum_{k=1}^{K} q_{k|j}\boldsymbol{P}_k.$$

Take any $\boldsymbol{v} \in \mathcal{S}_k = \operatorname{range}(\boldsymbol{P}_k)$. We have $\boldsymbol{P}_k\boldsymbol{v} = \boldsymbol{v}$, while orthogonality gives $\boldsymbol{P}_\ell\boldsymbol{v} = \mathbf{0}$ for $\ell \neq k$. Hence

$$\widetilde{\boldsymbol{\Sigma}}_j\boldsymbol{v} = \frac{1}{r}\sum_{\ell=1}^{K} q_{\ell|j}\boldsymbol{P}_\ell\boldsymbol{v} = \frac{q_{k|j}}{r}\boldsymbol{v}.$$

Thus, every vector in $\mathcal{S}_k$ is an eigenvector of $\widetilde{\boldsymbol{\Sigma}}_j$ with eigenvalue $q_{k|j}/r$. Because $\dim(\mathcal{S}_k) = r$, this eigenvalue has multiplicity r. Eigenvalues on the orthogonal complement of all semantic subspaces are zero.

Consequently, $\boldsymbol{I} + \gamma\widetilde{\boldsymbol{\Sigma}}_j$ has eigenvalue $1 + \gamma q_{k|j}/r$ on $\mathcal{S}_k$, with multiplicity r. Its determinant is the product of all eigenvalues, so

$$\begin{aligned} F_\gamma(\widetilde{\boldsymbol{\Sigma}}_j) &= \frac{1}{2}\log\det(\boldsymbol{I} + \gamma\widetilde{\boldsymbol{\Sigma}}_j) \\ &= \frac{1}{2}\sum_{k=1}^{K} r\log\Big(1 + \frac{\gamma q_{k|j}}{r}\Big). \end{aligned}$$

This proves Eq. (7).

The eigenvalue on the matching true-class subspace $\mathcal{S}_j$ is $q_{j|j}/r$, whereas the largest eigenvalue on an outside-class subspace is $\max_{k\neq j} q_{k|j}/r$. Their difference is g_j in Eq. (8). Therefore, $\mathcal{S}_j$ is the unique leading r-dimensional eigensubspace exactly when $g_j > 0$. □

A.6 Proof of the clean-dominance threshold for span noise

We apply the subspace comparison to cyclic span noise.

Proof of Corollary 2. Under cyclic span noise, each true class k retains label k with probability $1-\eta$ and changes to each of $k+1, \ldots, k+s$ (modulo K) with probability η/s. Cyclic shifts give every observed class the same incoming probability mass, so the transition matrix is doubly stochastic. By class balance and Step 1 of Appendix A.2, the posterior weights satisfy $q_{k|j} = T_{kj}$.

For any observed class j, the matching true class has posterior weight $q_{j|j} = 1 - \eta$. Exactly s other true classes can transition to j, each with weight η/s. All remaining weights are zero. Thus

$$\max_{k\neq j} q_{k|j} = \frac{\eta}{s}.$$

Substitution into Eq. (8) gives

$$g_j = g^{(s)} = \frac{1}{r}\left(1 - \eta - \frac{\eta}{s}\right).$$

Finally, $g^{(s)} > 0$ is equivalent to $1 - \eta > \eta/s$, or $\eta < s/(s+1)$, equivalently $s > \eta/(1-\eta)$. □

A.7 The boundary case of deterministic class permutation

We also examine a boundary case in which every class name is permuted consistently. The clean-dominance gap compares directions by class name. Under such a permutation, however, the grouping of samples by true class does not change. Suppose every class is renamed by the same one-to-one permutation. This deterministic, globally consistent, reversible noise structure is uncommon in real data. We use it only as a theoretical boundary check.

Corollary 3 (A deterministic class permutation preserves class partitions)**.** *If $T = \boldsymbol{P}$ is a permutation matrix, then*

$$\left\|\boldsymbol{P} - \frac{1}{K}\mathbf{1}\mathbf{1}^\top\right\|_F^2 = K - 1, \tag{19}$$

exactly as with clean labels. A deterministic class permutation therefore renames entire classes without changing the sample partition.

This corollary shows why the nominal mislabeling rate alone cannot determine whether class partitions are damaged. Even if every class name is permuted, MCR^2 sees the same sample groups, with only their names changed.

Proof of Corollary 3. Let $\boldsymbol{J} = K^{-1}\mathbf{1}\mathbf{1}^\top$. A permutation matrix satisfies $\boldsymbol{P}^\top\boldsymbol{P} = \boldsymbol{I}$ and $\boldsymbol{P}\mathbf{1} = \mathbf{1}$, so

$$\|\boldsymbol{P}\|_F^2 = \mathrm{tr}(\boldsymbol{P}^\top\boldsymbol{P}) = K,$$
$$\|\boldsymbol{J}\|_F^2 = 1,$$
$$\langle \boldsymbol{P}, \boldsymbol{J}\rangle_F = \mathrm{tr}(\boldsymbol{P}^\top\boldsymbol{J}) = \frac{1}{K}\mathbf{1}^\top\boldsymbol{P}^\top\mathbf{1} = 1.$$

Hence

$$\|\boldsymbol{P} - \boldsymbol{J}\|_F^2 = K + 1 - 2 = K - 1.$$

The clean-label transition matrix is $\boldsymbol{I}$, which also satisfies $\|\boldsymbol{I} - \boldsymbol{J}\|_F^2 = K - 1$. Thus, both have the same second-order class signal in Proposition 2. More directly, $T = \boldsymbol{P}$ merely renames each true class: the collection of class-conditional second moments is permuted, and the class-conditional coding rate sums over classes, so the objective itself is unchanged. □

As a related synthetic-noise control, pair-flip keeps the original label of each true class k with probability $1-\eta$ and changes it to a designated unique target class $P(k)$ with probability η. Its transition matrix is $T_\eta = (1-\eta)\boldsymbol{I} + \eta\boldsymbol{P}$. Only at $\eta = 1$ does it become a deterministic class permutation. At 80% pair-flip noise with $d = 128$, clean-centroid probe accuracy remains about 98.2%. If centroids are instead built from noisy training labels and evaluated against original clean test labels, accuracy is only about 0.57%. Thus, the representation can preserve true digit structure even when class names have nearly followed a fixed permutation. This control illustrates why nominal noise rate does not describe noise structure. It does not suggest that deterministic permutations are a common real-world noise model.

A.8 Proof of control of the opposing compression component under elastic adjustment

Finally, we show how the adaptive weight limits the opposing compression component under the simplified rule.

Proof of Proposition 4. Ignoring the moving-average lag means $\bar{\kappa}_t^{\mathrm{opp}} = \kappa_t^{\mathrm{opp}}$. For brevity, let

$$E = \|\boldsymbol{G}_E^{\tan}\|_F, \qquad A = \|\boldsymbol{a}_t\|_2.$$

By Eq. (14), $\kappa_t^{\mathrm{opp}} = A/(E+\delta)$. When $\lambda_{\min} = 0$, Eq. (16) becomes

$$\lambda_t = \min\left\{\frac{\tau}{\kappa_t^{\mathrm{opp}}}, 1\right\},$$

where the continuous limit gives $\lambda_t = 1$ if $\kappa_t^{\mathrm{opp}} = 0$. Consider two cases.

If $\kappa_t^{\mathrm{opp}} \le \tau$, then $\tau/\kappa_t^{\mathrm{opp}} \ge 1$ and $\lambda_t = 1$. Since $A = \kappa_t^{\mathrm{opp}}(E+\delta)$,

$$\lambda_t A = A = \kappa_t^{\mathrm{opp}}(E+\delta) \le \tau(E+\delta) = \tau E + \tau\delta.$$

If $\kappa_t^{\mathrm{opp}} > \tau$, then $\tau/\kappa_t^{\mathrm{opp}} < 1$ and $\lambda_t = \tau/\kappa_t^{\mathrm{opp}}$. Hence

$$\lambda_t A = \frac{\tau}{\kappa_t^{\mathrm{opp}}}A = \frac{\tau}{A/(E+\delta)}A = \tau(E+\delta) = \tau E + \tau\delta.$$

Both cases satisfy Eq. (17), completing the proof. □

B Empirical checks of the theory

B.1 Sample-wise normalisation under random label noise

We first check how sample-wise normalisation affects robustness under random label noise. We compare standard MCR^2 with MCR^2 without sample-wise normalisation on MNIST. Both use the same encoder: two 3×3 convolutional layers with 32 and 64 output channels, each followed by

Table 3: MNIST clean-centroid probe accuracy (%) under uniform label noise after 30 epochs.

Noise η	Clean-centroid probe accuracy (%)	
	MCR^2	Without normalisation
0.1	**98.40**	32.23
0.3	**98.46**	33.35
0.5	**98.08**	44.40
0.7	**96.62**	26.45
0.9	**77.38**	48.37

ReLU, a 2×2 max-pooling layer, and fully connected layers of widths 1024 and 128. The network has no dropout. Standard MCR^2 normalises each 128-dimensional output to unit length. The ablation uses the raw output.

At each noise rate $\eta \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, we randomly select exactly an η fraction of the training examples and replace each selected label uniformly with one of the other nine classes. Thus, every selected label is changed. Both methods share the same network initialisation, data order, and corrupted-label realisation. We train for 30 epochs with batch size 1000, SGD learning rate 10^{-3}, weight decay 5×10^{-4}, model seed 42, and noise seed 10. We evaluate each frozen encoder using the clean-centroid probe of Section 3.1.

Standard MCR^2 retains high clean-centroid probe accuracy through 70% corruption and declines near 90%. Without sample-wise normalisation, accuracy is much lower at every tested rate. This is consistent with Proposition 1: unit-length features give a uniform bound on each example's coding-rate contribution. Because removing normalisation also changes feature scale and optimisation, the accuracy gap alone does not isolate a label-noise-specific mechanism. The decline near $\eta = 0.9$ is also consistent with Corollary 1, which predicts that the second-order class signal vanishes when the observed label becomes independent of the true class.

B.2 Direct check of the random-label signal

We next test the predicted decline in the class signal under random label corruption. Corollary 1 makes this prediction for fixed features, so we do not retrain the encoder. We freeze a clean-label MCR^2 model with $d = 128$ and select 5,421 MNIST training examples per class, yielding a balanced set of 54,210 examples. At each noise rate, we randomly relabel the specified fraction within every true class, choosing each wrong label uniformly from the other nine classes.

Let $\widehat{\boldsymbol{\Sigma}}$ be the second moment of all fixed features, and let $\widehat{\boldsymbol{\Sigma}}_k^{\text{true}}$ and $\widehat{\boldsymbol{\Sigma}}_j^{\text{obs}}(\eta)$ be the moments grouped by true and observed labels. We measure the class-moment signal relative to clean labels:

$$S(\eta) = \frac{\frac{1}{K}\sum_{j=1}^{K} \|\widehat{\boldsymbol{\Sigma}}_j^{\text{obs}}(\eta) - \widehat{\boldsymbol{\Sigma}}\|_F^2}{\frac{1}{K}\sum_{k=1}^{K} \|\widehat{\boldsymbol{\Sigma}}_k^{\text{true}} - \widehat{\boldsymbol{\Sigma}}\|_F^2}.$$

For $K = 10$, the predicted population scaling is $S(\eta) = (1 - 10\eta/9)^2$. Figure 4 compares this curve with the measured signal.

The measured signal closely follows the predicted decline. At $\eta = 0.5$, the predicted and measured ratios are 0.1975 and 0.1996. At $\eta = 0.9$, they are 0 and 0.0022. This diagnostic directly examines label-induced mixing of fixed feature moments. It does not measure the exact log-determinant rate reduction or the outcome of noisy-label training.

B.3 Structured noise and subspace competition

We then examine how the concentration of label errors changes the leading class direction. To test Proposition 3 and Corollary 2, we fix $\eta = 0.8$ on MNIST and vary the span of cyclic noise. The corollary predicts a zero clean-dominance gap at $s = 4$ and a positive gap when $s > 4$.

We freeze the clean-label MNIST encoder with $d = 128$ used above and take a balanced subset of 5,421 training examples per class. For each true class k, the top $r = 4$ eigenvectors of its clean second moment define a reference subspace with orthonormal basis $\boldsymbol{U}_k$. The mean pairwise overlap

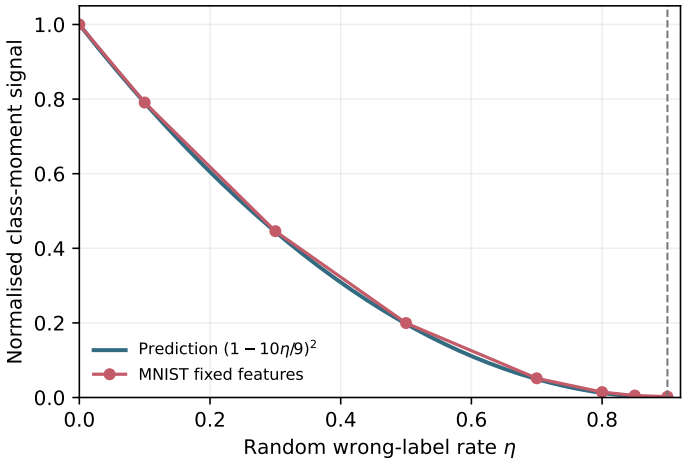


Figure 4: Predicted and measured fixed-feature class signal under random label noise on balanced MNIST.

Table 4: Fixed-feature MNIST clean-dominance gap under 80% cyclic label noise. The empirical gap is normalised by its clean-label value.

Noise span s	2	3	4	5	9
Idealised $rg^{(s)}$	-0.200	-0.067	0	0.040	0.111
Empirical mean gap / clean gap	-0.224	-0.095	-0.026	0.017	0.092
Classes with positive empirical gap	$0/10$	$0/10$	$1/10$	$8/10$	$10/10$

between these ten subspaces is below 10^{-5}. We then change 80% of labels within each true class using cyclic noise with span $s \in \{2, 3, 4, 5, 9\}$, keeping the features fixed.

For observed class j, let $\widehat{\boldsymbol{\Sigma}}_j^{\text{obs}}$ be its feature second moment. We compare the energy along the matching true-class subspace with the strongest off-class competitor:

$$\widehat{g}_j = \frac{1}{r}\operatorname{tr}(\boldsymbol{U}_j^\top \widehat{\boldsymbol{\Sigma}}_j^{\text{obs}} \boldsymbol{U}_j) - \max_{k \neq j} \frac{1}{r}\operatorname{tr}(\boldsymbol{U}_k^\top \widehat{\boldsymbol{\Sigma}}_j^{\text{obs}} \boldsymbol{U}_k).$$

We average $\widehat{g}_j$ over the ten observed classes and divide by the corresponding mean gap with clean labels. Under the idealised model, the comparable prediction is $rg^{(s)} = 1 - \eta - \eta/s = 0.2 - 0.8/s$. Figure 5 also compares the leading four-dimensional eigenspace of each observed-class moment with the matching and strongest off-class reference subspaces.

The measured mean gap changes from -0.224 at $s = 2$ and -0.095 at $s = 3$ to 0.017 at $s = 5$ and 0.092 at $s = 9$, with a near-zero value at $s = 4$. The matching subspace has a positive gap in 0, 8, and 10 of the ten observed classes at $s = 2$, $s = 5$, and $s = 9$, respectively. At $s = 2$, the leading observed-class eigenspace aligns mainly with an off-class subspace. By $s = 9$, it aligns mainly with the matching one. This supports the predicted change in which class direction dominates as errors become more dispersed. The real MNIST class moments are not isotropic rank-four matrices, so the experiment checks the qualitative transition rather than the exact eigenvalues or threshold in Proposition 3.

B.4 Control experiment with a small cyclic label-flip rate

Finally, we check whether concentrating a small number of errors in one wrong class necessarily undermines MCR^2 robustness. For each true CIFAR-10 class, we select exactly 500 training images (10% of the class) and cyclically change their labels to the next class, wrapping the last class to the first. For the random-flip control, we give the same selected images labels drawn uniformly from the

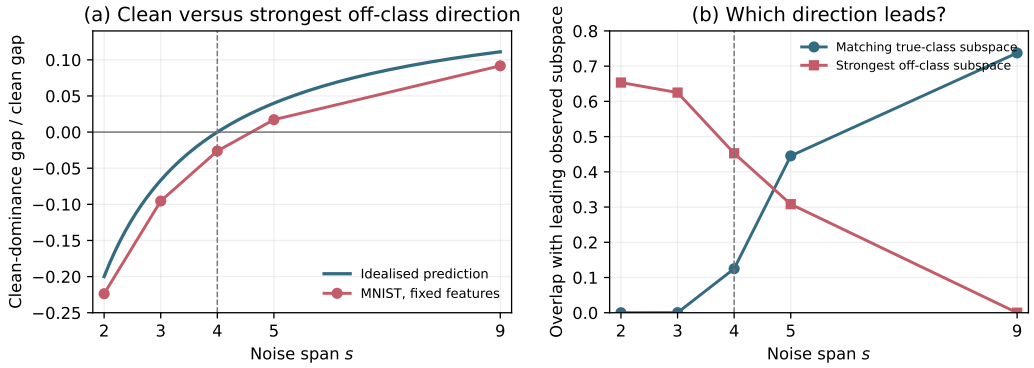


Figure 5: Fixed-feature MNIST subspace competition at 80% cyclic label noise. Left: normalised clean-dominance gap. Right: overlap with matching and strongest off-class subspaces.

Table 5: CIFAR-10 clean-centroid probe accuracy (%) under 10% label flips. Cyclic and random flips use the same images and seed 42.

Training labels	Clean-centroid probe accuracy (%)
Clean labels	**88.13**
10% cyclic flips to the next class	87.21
10% random flips to another class	84.61

other nine classes. Clean-label training provides a baseline. All three runs use ResNet-18, $d = 512$, batch size 1000, the same model seed, and 100 epochs. We evaluate by the clean-centroid probe of Section 3.1.

Table 5 shows that cyclic flipping is only 0.92 percentage points below the clean baseline and 2.60 points above random flipping at the same rate. It does not produce the anticipated pronounced failure. In the idealized model of Corollary 2, each observed class contains 90% from its matching true class and 10% from the preceding class, giving $rg^{(1)} = 0.8 > 0$. This experiment examines the low-noise side of the condition and shows that concentrated errors alone need not remove clean-direction dominance. The idealized spectral gap does not guarantee final accuracy ordering across noise structures in nonconvex training.

C Additional representation geometry

We inspect the final training features at 50% and 80% span-2 noise. For each noise rate, the same 2,000 CIFAR-10 training images are projected for MCR^2 and Elastic Coding. Within each method, identical t-SNE coordinates are coloured by true or observed labels. The methods use separate projections. The 50% plot is in Figure 3, and the 80% plot is below.

To quantify structure in the original 512-dimensional space, we form a reference centroid for each true class using only training images whose observed labels are correct. We then assign every incorrectly labelled training image to its nearest reference centroid. This uses synthetic ground truth to identify corrupted labels, which is unavailable in ordinary noisy-label training.

At 80% noise, observed-centroid test accuracy remains poor for both methods (Table 1). Preserving feature structure therefore does not by itself recover the mapping from clusters to the original class names.

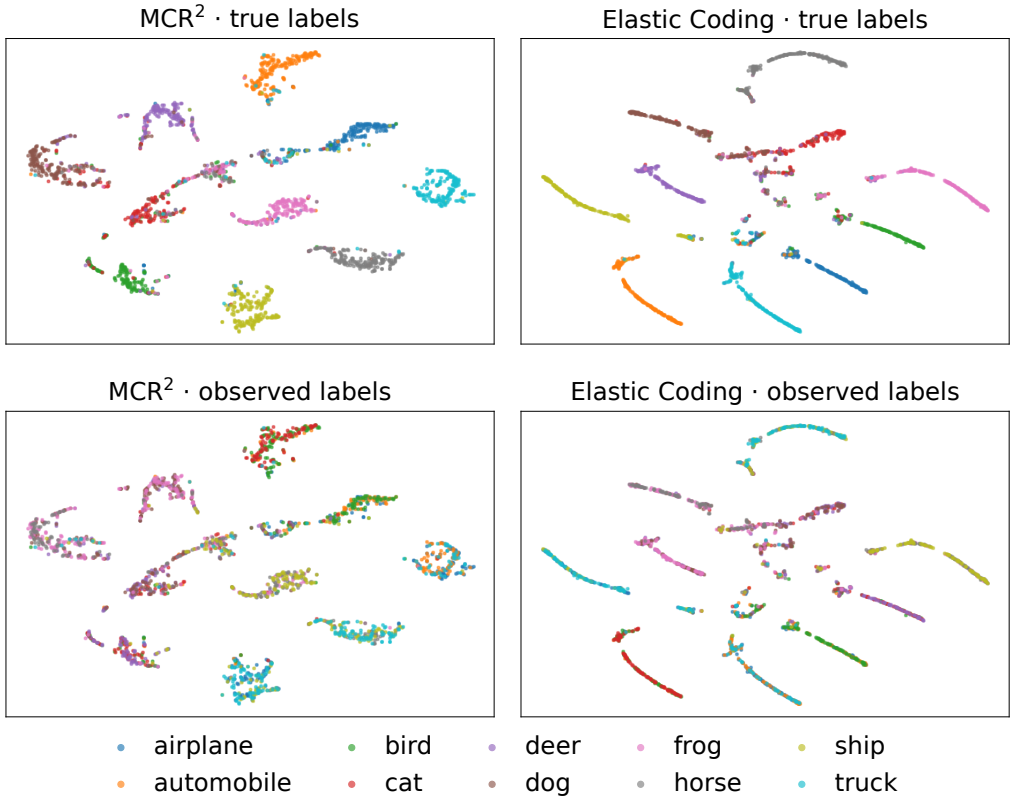


Figure 6: CIFAR-10 training representations at 80% span-2 label noise. Columns compare methods, and rows show true or observed labels.

D Additional method details

D.1 Directional coding strain

We first separate the contribution of a contaminating feature along an existing class direction from its contribution along a new direction.

Let $\boldsymbol{C}_j\boldsymbol{C}_j^\top = \sum_{\ell=1}^{r_j} \rho_{j\ell}\boldsymbol{u}_{j\ell}\boldsymbol{u}_{j\ell}^\top$ with positive eigenvalues $\rho_{j\ell}$. The inverse in Eq. (6) has weight $1/(1+\alpha_j\rho_{j\ell})$ along $\boldsymbol{u}_{j\ell}$ and weight one on the orthogonal complement. Writing one contaminating feature as $\boldsymbol{n} = \sum_{\ell=1}^{r_j} a_\ell \boldsymbol{u}_{j\ell} + \boldsymbol{n}_\perp$ gives

$$\delta\mathcal{R}_c(\boldsymbol{n}) = \frac{1}{2}\log\left[1 + \alpha_j\left(\sum_{\ell=1}^{r_j}\frac{a_\ell^2}{1+\alpha_j\rho_{j\ell}} + \|\boldsymbol{n}_\perp\|_2^2\right)\right]. \tag{20}$$

An already populated direction contributes less to the marginal class rate than a new orthogonal direction. The group expression in Eq. (6) also accounts for correlations among several contaminating features.

D.2 Gradients on the unit sphere

We next give the gradients used to compare expansion and compression after sample-wise normalisation.

The matrix identity $\mathrm{d}\log\det\boldsymbol{A} = \mathrm{tr}(\boldsymbol{A}^{-1}\mathrm{d}\boldsymbol{A})$ gives $\partial\mathcal{R}/\partial\boldsymbol{Z} = \alpha\boldsymbol{G}^{-1}\boldsymbol{Z}$. For an example in observed class j, the product of the class-rate weight and its inner coefficient is $(\widetilde{m}_j/m)\alpha_j = \alpha$. Hence $\partial\mathcal{R}_c/\partial\boldsymbol{z}_i = \alpha\boldsymbol{C}_j^{-1}\boldsymbol{z}_i$, which yields Eq. (9). The normalisation map has Jacobian $(\boldsymbol{I} -$

$\boldsymbol{z}_i\boldsymbol{z}_i^\top)/\|\boldsymbol{h}_i\|_2$. Applying the chain rule yields Eq. (10). For a matrix $\boldsymbol{H} = [\boldsymbol{g}_1, \ldots, \boldsymbol{g}_m]$, let $\mathcal{P}_{\boldsymbol{Z}}(\boldsymbol{H})$ apply $(\boldsymbol{I} - \boldsymbol{z}_i\boldsymbol{z}_i^\top)$ to column i. The two ascent directions used by Elastic Coding are $\boldsymbol{G}_E^{\tan} = \mathcal{P}_{\boldsymbol{Z}}(\partial\mathcal{R}/\partial\boldsymbol{Z})$ and $\boldsymbol{G}_C^{\tan} = -\mathcal{P}_{\boldsymbol{Z}}(\partial\mathcal{R}_c/\partial\boldsymbol{Z})$. Their columns define the vectors in Eqs. (12)–(13). The smoothed conflict ratio obeys $\bar{\kappa}_t^{\mathrm{opp}} = \beta\bar{\kappa}_{t-1}^{\mathrm{opp}} + (1-\beta)\kappa_t^{\mathrm{opp}}$. The weight is treated as constant during the encoder update.

D.3 Class-adaptive extension

We finally describe how the same weighting rule can be applied separately to each observed class.

For observed class j, let $\mathcal{I}_j$ index its minibatch examples, $\boldsymbol{a}_{j,t} = (a_{i,t})_{i\in\mathcal{I}_j}$, and $\boldsymbol{G}_{E,j}^{\tan}$ contain the corresponding expansion-gradient columns. A class-specific version is

$$\kappa_{j,t}^{\mathrm{opp}} = \frac{\|\boldsymbol{a}_{j,t}\|_2}{\|\boldsymbol{G}_{E,j}^{\tan}\|_F + \delta}, \qquad \lambda_{j,t} = \mathrm{clip}\left(\frac{\tau}{\bar{\kappa}_{j,t}^{\mathrm{opp}}}, \lambda_{\min}, 1\right). \tag{21}$$

It maximises $\mathcal{R}(\boldsymbol{Z}) - \sum_{j=1}^{K} \lambda_{j,t}(\widetilde{m}_j/2m)\log\det(\boldsymbol{I}_d + \alpha_j\widetilde{\boldsymbol{Z}}_j\widetilde{\boldsymbol{Z}}_j^\top)$. This extension can relax compression differently across observed classes. The experiments in the main text use the global weight in Eq. (16).