ORIENTATION CAN REVERSE THE HARM OF REPRESENTATION COLLAPSE

Anonymous authors
Paper under double-blind review

ABSTRACT

Representation collapse is blamed for brittle features, yet neural collapse accompanies good generalisation, and no label-free statistic says which of the two a model is in. We show that collapse magnitude alone does not decide it: at a matched contraction, the orientation of that contraction relative to semantic and nuisance directions changes how much task-relevant information survives, in a construction where every spectral statistic of the Jacobian is identical. We formalise this with a representation's linear semantic Fisher information and test the formal claim where its hypotheses hold exactly, on a frozen encoder: the predicted change in semantic information is reproduced to better than $4 \cdot 10^{-10}$, while the companion statement about nuisance contraction has no exact realisation on these encoders, which is consistent with contracting a nuisance response costing information rather than saving it. Holding one regulariser fixed and rotating only the factor it acts on, the representation-level out-of-distribution error is 65.7% for a semantic contraction against 42.4% for a neutral nuisance, and, at effective ranks within 4% of each other, 6.4% for a semantic protection against 38.5% for a neutral one; six of our eight pre-registered predictions nonetheless failed and one of two kill criteria fired. Over unregularised networks swept through two training regimes our index keeps one sign in all thirty strata where each of eight collapse statistics reverses, but an information-matched control shows most of that advantage is the counterfactual pool rather than our summary of it. The one failure the index provably cannot see is a fold, which we exhibit and measure.

1 INTRODUCTION

Representation collapse has two reputations. In supervised training, the collapse of last-layer features onto their class means accompanies good in-distribution accuracy (Papyan et al., 2020), and features of inputs that play the same role in a task collapse precisely when networks generalise (Laurent et al., 2024). In self-supervised learning, dimensional collapse destroys information that downstream tasks need (Jing et al., 2022); stronger collapse on a source task lowers transferability (Kornblith et al., 2021), and classifiers that are invariant to task-relevant changes are easy to fool (Jacobsen et al., 2019). Diagnostics inherit the ambiguity: spectral statistics such as effective rank and RankMe reward less collapse (Garrido et al., 2023), neural-collapse statistics reward more (Galanti et al., 2022), and neither says which regime a model is in. This raises a basic question: *what property of a collapsed representation decides whether the collapse helps or hurts?*

We find that the amount of collapse does not settle the question; its orientation can reverse the answer. Discarding nuisance variation, such as a colour that does not determine the class, loses nothing the task needs; discarding an equally large task-relevant variation leaves the representation unable to tell semantic states apart. Two distinct things can go wrong and we keep them apart throughout: an **overlap degeneracy**, a task-relevant response that has become indistinguishable from nuisance variation, which our index measures; and a **fold**, an actual loss of injectivity where distinct semantic states map to one value (Berry & Upstill, 1980), which it cannot see. Conflating them is the easiest way to overclaim, so Section 3 defines and relates both. Figure 1 shows the idea and the evidence: folding a representation along the semantic axis rather than a nuisance axis leaves every statistic of its distribution unchanged while the Bayes accuracy moves from chance to perfect (panel b), and in trained networks one regulariser applied to different factors produces similar effective ranks with very different out-of-distribution error (panel c).

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.

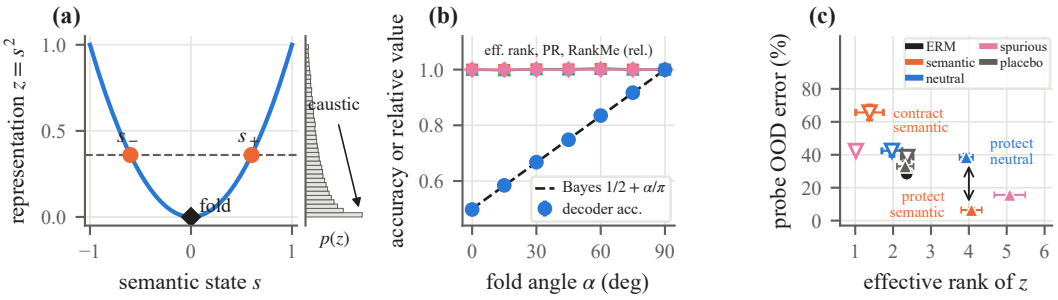


Figure 1: **Matched collapse can have opposite consequences depending on orientation.** (a) A fold maps two semantic states to the same representation value and concentrates the density as $|z - z_c|^{-1/2}$, the signature of an optical caustic. (b) Folding the direction at angle α from the semantic axis of a 2D Gaussian (Proposition 2): over 7 angles and 5 seeds the decoder plotted here tracks the accuracy implied by the proposition, $\frac{1}{2} + \frac{\alpha}{\pi}$, to within 0.006, while the seed-averaged effective rank, participation ratio and RankMe vary by under 1%. (c) Trained 3D Shapes encoders under one regulariser applied to different factors (§4.2; colour: factor, the neutral and spurious ones being nuisances; open: contract, filled: protect; mean ± sd over 5 seeds). The arrow marks the two protection arms, which differ by 3.4% in effective rank, 4.07 against 3.93, and sixfold in probe error, 6.4% against 38.5%.

We make this precise with the *semantic Fisher information* of a representation, the information about the semantic coordinates that survives when nuisance variation is unknown, and we test the resulting theorem where its hypotheses hold exactly rather than only by analogy (Sections 3 and 4.1). Contracting nuisance responses cannot lower it, contracting semantic responses by the same factor lowers its volume by a computable amount, and in a matched construction every spectral statistic of the Jacobian is identical in the two cases (Theorem 1). That theorem is about information, not error: the step to downstream failure is a prediction we test, not something we prove. Where a representation folds, the first-order picture gives out and singularity theory supplies the signatures instead (Theorem 2).

We test these predictions with experiments that could have come out the other way, on data whose semantic and nuisance factors we control.

- **At matched collapse, orientation reverses the effect.** Holding one regulariser and its weight fixed and rotating only the factor it acts on, the representation-level OOD error is 65.7% contracting the semantic factor against 42.4% contracting a neutral nuisance, and 6.4% protecting it against 38.5% protecting a neutral nuisance, the two protection arms within 4% of each other in effective rank (5 seeds, paired intervals in Appendix D.19). Controls separate this from the generic cost of a penalty, label supervision, counterfactual exposure and shrinkage without alignment; six of eight pre-registered predictions failed and one kill criterion fired, and Section 4.2 says why before what held.

- **Counterfactual direction information, more than our summary of it, is what makes a diagnostic survive a change of regime.** Our index keeps one sign in all thirty strata of four sweeps where each of eight collapse statistics reverses in nine to fourteen; but given the same counterfactual pool a published baseline reaches thirteen of fourteen and a plain contraction index nearly matches ours, so what we claim is the pairing of a declared counterfactual pool with a same-operator rotation, not a better summary statistic. What the log-determinant adds is sensitivity to a uniform semantic contraction, to which an effective rank is exactly invariant (Section 4.3).

- **Folds obey the fold normal form, and the index cannot see them.** When topology forces a network to fold a periodic semantic variable the folds have local order near 2, a density exponent of 0.50 to 0.52 (predicted $\frac{1}{2}$), and 80% to 96% of errors that are collisions between preimages of one code value against 50% to 65% for a distance-matched null; on our own fold construction the index orders the angles backwards, which fixes its scope (Appendices D.9, D.15).

2 RELATED WORK

Collapse as a virtue, and as a failure. Neural collapse concentrates last-layer features on their class means (Papyan et al., 2020); it is provably optimal in several regularised models (Jacot et al., 2025) though not every deep one (Súkeník et al., 2024), tracks generalisation with scale (Wu & Papyan, 2024), and appears for tokens in identical roles (Laurent et al., 2024). Information-theoretic accounts ask for representations sufficient for the task and invariant to nuisances (Tishby & Zaslavsky, 2015; Achille & Soatto, 2018). Against this, self-supervised objectives suffer dimensional collapse (Jing et al., 2022), which variance and decorrelation terms prevent (Bardes et al., 2022); augmentations decide which factors are nuisances (Balestriero et al., 2022); source-task collapse lowers transferability (Kornblith et al., 2021; Hui et al., 2022; Harun et al., 2025; Xue et al., 2024), while Galanti et al. (2022) argue the opposite for few-shot transfer; classifiers are excessively invariant to task-relevant changes (Jacobsen et al., 2019; Feather et al., 2023); and invariance and sensitivity trade off task-dependently (Tramèr et al., 2020; Xiao et al., 2021). That an invariance is harmful exactly when it discards a task-relevant factor is the premise of this line, not our claim. What we add is one thing: a same-operator rotation, holding the regulariser and its weight fixed and changing only which factor it reads, at matched collapse magnitude. The semantic Fisher information is how we read that intervention; the counterfactual pool it needs is borrowed, and Appendix D.13 quantifies how much of our observational result the pool alone buys.

Diagnostics of representation quality. Label-free spectral measures (Roy & Vetterli, 2007; Garrido et al., 2023; Agrawal et al., 2022) predict linear-probe accuracy from the spectrum alone; class-distance normalised variance (Galanti et al., 2022) and its directional form (Luthra et al., 2026) add labels; Chou et al. (2026) forecast out-of-distribution failure from class-manifold geometry, after Chung et al. (2018); intrinsic dimension (Ansuini et al., 2019) and contrastive alignment and uniformity (Wang & Isola, 2020) are read globally too. Proposition 2 shows statistics of P_z are blind to orientation by construction. The variability collapse index of Xu & Liu (2023) is label-aware and, unlike ours (Proposition 3), invariant to every invertible linear map: the right choice when the readout may rescale, the wrong one when a stated readout resolution is what makes a contraction costly. Where ground-truth factors exist, the disentanglement literature scores representations factor by factor, most directly the informativeness term of DCI (Eastwood & Williams, 2018) and the interventional robustness of Suter et al. (2019); those ask how well each factor can be predicted, where we ask how much semantic information survives *after* conditioning on nuisance variation. LiDAR (Thilak et al., 2024) is the closest relative: it whitens a between-group scatter by the within-group scatter augmentations generate, the pairing Eq. (1) makes. Two things differ and Section 4.3 measures each: its groups are augmentations rather than interventions on identified factors, and it summarises by an effective rank, exactly invariant to a uniform contraction of the semantic response. Giving LiDAR our pool recovers most of the cross-regime sign stability. The closest geometric quantity outside this line is the signal-noise overlap of Sorscher et al. (2022), of which our index is the local, counterfactual analogue.

Sensitivity penalties, and folds. Invariance to chosen transformations has been measured since Goodfellow et al. (2009); small Jacobian norms correlate with generalisation (Novak et al., 2018) and Jacobian penalties improve robustness (Hoffman et al., 2019). Tangent propagation penalises sensitivity along chosen transformations (Simard et al., 1991) and contractive auto-encoders penalise the whole input-output Jacobian (Rifai et al., 2011), explanation and gradient penalties do so with human or counterfactual supervision (Ross et al., 2017; Teney et al., 2020); and Rajput (2026) orders a matched, an isotropic and a wrong-axis penalty by how well each covers the deployment shift. The closest penalties to our contract arm are the conditional-variance penalty of Heinze-Deml & Meinshausen (2021) and the counterfactual-invariance regularisers of Veitch et al. (2021); both aim to *remove* a nuisance, where our arms rotate one fixed operation across factors to measure what each removal costs. Our Q1 design shares that scaffold and asks a different question of it: the axes are generative factors, the arms raise a floor on a factor as well as shrink it, a placebo matches the penalty's norm, and we read off what the collapse did to the semantic information; this also explains why one sensitivity number can carry either sign. In population coding, noise correlations aligned with the signal cap the Fisher information of arbitrarily large populations (Moreno-Bote et al., 2014) and arise when input nuisance propagates through the computation (Kanitscheider et al., 2015); Eq. (1) is that quantity, and new here are the link to collapse statistics and the test by intervention.

Piecewise-linear networks fold their input space (Montúfar et al., 2014) and deep layers reduce representation rank (Feng et al., 2022; Masarczyk et al., 2023). Shortcut learning (Geirhos et al., 2020; Sagawa et al., 2020; Shah et al., 2020; Pezeshki et al., 2021) is reduced by SGD's implicit regularisation (Mirzaie et al., 2026), detectable from embedding geometry (Păduraru et al., 2026), and often repaired by last-layer retraining (Kirichenko et al., 2023), which works only when the representation still separates the core feature from the nuisance. Caustics and their normal forms are classical (Berry & Upstill, 1980; Arnold et al., 2012) and Theorem 2 is theirs, not ours; that a low-dimensional latent cannot encode a topologically nontrivial variable is likewise known (Batson et al., 2021); new here is testing those signatures in learned representations and keeping them separate from the overlap degeneracy a first-order index can detect.

3 Collapse orientation, semantic information, and folds

Setup and notation. Locally the data manifold is $x = \psi(s, n)$ with *semantic* coordinates $s \in \mathbb{R}^r$, on which the label depends, and *nuisance* coordinates $n \in \mathbb{R}^q$, on which it does not. An encoder $f : \mathbb{R}^D \to \mathbb{R}^d$ induces $F = f \circ \psi$ with semantic and nuisance Jacobians $J_S = \partial_s F$ and $J_N = \partial_n F$, and $J = [J_S \; J_N]$; the encoder *collapses* a tangent direction when J maps it to (nearly) zero. Label-free collapse statistics are *orientation-agnostic*: functions of $P_{f(X)}$ (covariance spectrum, effective rank, RankMe, participation ratio) or of the singular values of J. Label-aware ones such as NC1 and CDNV compare within-class with between-class variation, but never ask which input directions that variation comes from. Table 3 (Appendix A) collects the notation.

Semantic Fisher information. What matters is how well s can be read from $z = f(x)$ when n is unknown. Linearising F at (s_0, n_0) with $n \sim \mathcal{N}(n_0, \Sigma_n)$ and a readout noise floor $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ gives $z \mid s \sim \mathcal{N}(F_0 + J_S(s - s_0), \Sigma_z)$ with $\Sigma_z = J_N \Sigma_n J_N^\top + \sigma^2 I$, and the *linear* Fisher information about s is

$$I_S \;=\; J_S^\top \, \Sigma_z^{-1} \, J_S. \tag{1}$$

This is the quantity studied for population codes whose noise correlations are driven by input nuisance (Moreno-Bote et al., 2014; Kanitscheider et al., 2015), and it already contains the orientation argument: nuisance responses enter only through the noise covariance, semantic responses as the signal. Two caveats fix its status: Σ_z depends on s through J_N, so Eq. (1) is the mean-dependent part of the Fisher information and a lower bound on it, and $\mathrm{Cov}(\hat{s}) \succeq I_S^{-1}$ (Lehmann & Casella, 1998) is the Cramér–Rao statement for locally unbiased readouts of the linearised model. At an exact fold the mean channel vanishes along the folded direction, so Eq. (1) is zero there; the full Fisher information need not be, because Σ_z can itself vary with s, and we do not claim otherwise. We summarise I_S by the **semantic information deficit** $\mathrm{SID}_\sigma(x) = -\frac{1}{2r} \log\det\big(\sigma^2 I_S(x)\big)$, which grows as the representation loses semantic volume relative to nuisance variation. For comparison we also use the plain contraction indices $C_S = -\frac{1}{2r} \log\det(J_S^\top J_S)$ and $C_N = -\frac{1}{2q} \log\det(J_N^\top J_N)$, which ignore the interaction between the two.

Two distinct objects, and the index measures only one. Two different failures are easy to conflate, and keeping them apart is what makes the claims falsifiable. A **semantic overlap degeneracy** is a point where some semantic direction has become indistinguishable from nuisance variation, $J_S u \in \mathrm{range}(J_N)$ for some $u \neq 0$, with $J_S u = 0$ the extreme case; there I_S stays bounded along u as $\sigma \to 0$ while it diverges along every direction whose response escapes the nuisance span, so $\mathrm{SID}_0 = +\infty$ exactly there (Appendix B). A **fold caustic** is a loss of injectivity of F along a semantic coordinate: a nondegenerate critical point $F'(s_c) = 0$, $F''(s_c) = \kappa \neq 0$, at which two semantic states map to one representation value.

Proposition 1 (Relation between the two)**.** *A fold caustic at s_c is an extreme overlap degeneracy* at that point*, since $J_S u = 0$ there. Neither converse holds: responses can overlap with F still locally injective, and a global two-to-one identification can exist while F is a local diffeomorphism off a measure-zero set, where no first-order quantity detects it.*

Proposition 1 fixes the scope of the index: SID_σ measures overlap degeneracy, *not* folding, and we state the negative result before using it anywhere. On the construction of Proposition 2 it orders the seven fold angles *backwards* (Spearman -0.54 with the Bayes error, -0.75 with secants long enough to cross the fold; Appendix D.15). Folds are what Theorem 2 is for; we never claim the index detects them.

As a binary predicate, overlap degeneracy is informative only when the representation has room for both spans: for *generic* Jacobians it holds exactly when $\min(r, d) + \min(q, d) > d$. With our $r = 3$ and $q = 5$, dimension alone therefore forces it for every width $d \leq 7$, hence for the $d = 2$ and $d = 4$ networks in our sweeps; for $d = 16$ or 64 it does not, though learned, non-generic Jacobians may still overlap, and in our networks they do, with about 99% of the squared semantic secant lying in the nuisance span (Appendix D.7). What we measure everywhere is the graded index at a fixed σ, not the predicate.

The same contraction has opposite consequences.

Theorem 1 (Orientation changes semantic information at matched contraction magnitude)**.** *Let $\Sigma_n = I$ (nuisance coordinates whitened), $\sigma > 0$, and $\mathrm{rank}(J_S) = r$, so that SID_σ is finite.* ***(i)*** *If $J'_N = J_N Q$ with $\|Q\|_2 \leq 1$ and J_S is unchanged, then $I'_S \succeq I_S$ and $\mathrm{SID}'_\sigma \leq \mathrm{SID}_\sigma$: contracting nuisance responses cannot decrease the semantic Fisher information.* ***(ii)*** *If $J'_S = J_S B$ with $B \in \mathbb{R}^{r\times r}$ nonsingular and J_N is unchanged, then $\mathrm{SID}'_\sigma = \mathrm{SID}_\sigma - \frac{1}{r}\log|\det B|$; contracting one semantic direction by $\beta \in (0, 1]$ raises the index by $\frac{1}{r}\log\frac{1}{\beta}$. For singular B the identity holds in the extended-real sense with $\mathrm{SID}'_\sigma = +\infty$.* ***(iii)*** *If J has orthonormal columns, contracting one semantic column by β and contracting one nuisance column by β produce Jacobians with identical singular values $\{\beta, 1, \ldots, 1\}$, so every function of the Jacobian spectrum coincides, whereas SID_σ rises by $\frac{1}{r}\log\frac{1}{\beta}$ in the first case and is unchanged in the second.*

What Theorem 1 does not say. It concerns a local linear-Gaussian model with a fixed readout metric and bounds no classification or out-of-distribution error. Part (iii) is a matched *construction*: spectral statistics are blind to orientation there, not in trained networks, where they are often informative. Section 4.1 tests it where its hypotheses hold exactly and says which of them a trained encoder can meet; the rest of Section 4 tests the prediction it licenses, that rotating the factor moves semantic information and downstream failure with it.

Proposition 2 (An exact orientation law)**.** *Let $(s, n) \sim \mathcal{N}(0, I_2)$, $y = \mathbf{1}[s > 0]$, and fold the direction at angle α from the semantic axis: $z_\alpha = \big((s\cos\alpha + n\sin\alpha)^2,\ -s\sin\alpha + n\cos\alpha\big)$. For every $\alpha \in [0, \pi/2]$, z_α has the same distribution, so every statistic of P_z (spectrum, effective rank, RankMe) is constant in α; yet the Bayes error of predicting y from z_α is exactly $\frac{1}{2} - \frac{\alpha}{\pi}$.*

What the index is invariant to. The floor $\sigma^2 I$ is the precision with which anything downstream can resolve z, not noise inside a deterministic encoder, so SID_σ is a property of a representation *together with that metric*.

Proposition 3 (Invariance class)**.** *Under $z' = Uz$ with $U^\top U = I$, SID_σ is unchanged. Under $z' = az$ with the floor rescaled to $\sigma' = |a|\sigma$, it is unchanged. Under a general invertible A with the floor held at $\sigma^2 I$, it is not invariant.*

The first two are exact and the third is not, confirmed numerically without retraining (Appendix D.17). This is the intended invariance class, not a defect: the index answers how much semantic information survives *at a stated readout resolution*, so we hold σ^2 and the global scale fixed in every comparison.

Fold caustics and their signatures. In geometric optics a caustic is where a family of rays loses local invertibility and the ray density diverges as the inverse square root of the distance to it (Berry & Upstill, 1980; Arnold et al., 2012); the same normal form governs a representation that folds a semantic coordinate.

Theorem 2 (Fold caustics; classical, stated in full in Appendix B)**.** *Let F be C^3 along a semantic coordinate with a nondegenerate critical point at s_c, and let s have a Lipschitz positive density there. Then in a neighbourhood of s_c the representation density diverges as $|z - z_c|^{-1/2}$ (CDF exponent $\frac{1}{2}$); each attainable z near z_c has exactly two local preimages, so a decoder errs by $\Omega(|z - z_c|^{1/2})$ on one branch and noise of size δ becomes a semantic error of order $\delta^{1/2}$; and if the semantic variable is periodic with $d = 1$, every continuous F identifies an antipodal pair and has at least two critical points, while for $d \geq 2$ a fold-free embedding exists. At a non-smooth* kink *the CDF exponent is* 1 *instead.*

Theorem 2 is classical singularity theory and is not ours; its value here is to turn "semantic collapse" into four measurable predictions that Section D.9 tests: local order 2, CDF exponent $\frac{1}{2}$, preimage-collision errors, and exponent 1 for a kink. Table 2 in Appendix A gives each result's assumptions and its test; the line to carry forward is that **no result here bounds a classification or out-of-distribution error**, so every step from semantic information to downstream failure is a prediction we test, not a theorem.

3.1 Estimators

The theory above is exact for continuous semantic coordinates. Two of our three experiments have a *categorical* semantic factor, where no J_S exists; there the estimator measures a counterfactual Mahalanobis discriminability between semantic states, motivated by Eq. (1) and predicting the sign of its effects, though the Cramér–Rao reading does not transfer literally. On data with known factors we compute exact *secants*: for a base input we change one semantic value (a shape) or one nuisance value (a hue), keep everything else fixed, and record δz, normalised by the global scale of z. Stacking the r semantic secants $S \in \mathbb{R}^{d\times r}$ and the q nuisance secants $N \in \mathbb{R}^{d\times q}$ gives $\hat{C}_S = -\frac{1}{2r}\log\det(S^\top S)$, $\hat{C}_N = -\frac{1}{2q}\log\det(N^\top N)$ (both corrected for the input-space step length), and

$$\hat{C}_F = -\tfrac{1}{2r}\log\det\big(S^\top(NN^\top+\sigma^2 I)^{-1}S\big), \tag{2}$$

the secant version of SID_σ (up to the additive constant $\frac{1}{2}\log\sigma^2$) with $\sigma^2 = 10^{-2}$ (sensitivity in Appendix D.6). When $d < r$ the semantic secants cannot be independent and $\hat{C}_F$ is $+\infty$; such a network is censored at the top of the scale rather than measured, affecting the $d = 2$ arm of one sweep only (Appendix D.14). The index is defined relative to a *declared set of interventions*: the columns of N are one-step changes of each nuisance factor, which sets $\Sigma_n = I$ in those units and, as Section 4.1 shows, decides which part of Theorem 1 an encoder can satisfy exactly. Generative factors are not orthogonal pixel directions, so we call these *factor-defined secants* and measure the overlap (Appendix D.16). A secant is a *counterfactual*, so the estimator needs a generator of single-factor counterfactuals but no out-of-distribution data, test labels or retrained probe; the baselines see training samples only, so Section 4.3 measures how much that access explains.

4 Experiments

We first test Theorem 1 where its hypotheses can be met exactly, then three predictions of Section 3: that holding the operation fixed and rotating only the factor it acts on moves failure in opposite directions (**Q1**, Theorem 1); that a direction-resolved index keeps its sign across training regimes where orientation-agnostic statistics do not, and how much of that comes from the extra information rather than the summary (**Q2**); and that a network forced to fold a semantic variable shows the signatures of Theorem 2 (**Q3**). We use data whose semantic and nuisance factors are known exactly: 3D Shapes (Burgess & Kim, 2018; Kim & Mnih, 2018) and dSprites (Matthey et al., 2017), where the label is the object's shape and every other factor is a nuisance, and CIFAR-10 (Krizhevsky, 2009) with a planted coloured border. Training data carry a spurious correlation of strength ρ between the label and one nuisance (object hue, horizontal position, border colour) and out-of-distribution (OOD) test data have $\rho = 0$; besides the model's own OOD error we report the *representation-level* OOD error of a probe retrained on frozen features of decorrelated data (Kirichenko et al., 2023), which separates what the representation lost from what the head learned (Appendix C). A multinomial logistic decoder and a two-layer MLP decoder both reproduce the law of Proposition 2 to within 0.008, the MLP plotted in Figure 1b to within 0.006, while every label-free statistic stays flat (Figure 1b).

4.1 An exact test of Theorem 1

The factorial below moves the representation with a soft penalty, so neither hypothesis of Theorem 1 holds there. We first ask when they can be met exactly: freeze a trained encoder and post-multiply its representation by $A = I-(1-\beta)P$, P an orthogonal projector of rank k. With $V = \mathrm{range}(J_S)$ and $W = \mathrm{range}(J_N)$, part (ii) needs $P \subseteq V \cap W^\perp$ and part (i) needs $P \subseteq W \cap V^\perp$, of generic dimension $\max(0, r-q)$ and $\max(0, q-r)$. So with the declared 3D Shapes set ($r = 3$, $q = 5$) part (ii) has *no* nontrivial exact operator and part (i) has a two-dimensional one, while a smaller

declared nuisance set ($q = 2$) makes (ii) testable instead (Appendix D.2; 400 base points, 3 seeds). Where the hypothesis holds, the theorem holds exactly: with the nuisance block fixed to 10^{-16}, the measured change in the deficit matches the predicted $(k/r)\log(1/\beta)$ to $3 \cdot 10^{-10}$ at every base point and every β. Part (i) is the instructive case. Contracting nuisance directions *in representation space* gives $J'_N = J_N Q$ with $\|Q\|_2$ up to 174 rather than ≤ 1, so the hypothesis fails and the deficit rises at all 4800 point-β pairs. That is the theory's own account of something Section 4.2 observes and we previously could not explain: contracting a nuisance response is not the operation part (i) protects, and it can cost semantic information. Part (iii) is the one that does not come out as predicted. At identical operator spectra the deficit rises 0.82 ± 0.05 nats for a semantic subspace against 0.09 ± 0.02 for a nuisance one, a ninefold separation in the predicted direction but short of the predicted 1.535 and above the predicted 0, because J is not orthonormal here ($a_S \approx 0.99$). Nor are the spectral statistics blind here: effective rank reads 1.79 against 2.35 and orders the two correctly; what the deficit adds is the size of the gap, not its existence.

4.2 Q1: THE SAME REGULARISER ALONG DIFFERENT FACTORS

We train the same network on the same data with one extra loss term whose only varying ingredient is the factor-defined secant k it acts on. *Contract*$[k]$ shrinks the Fisher-normalised length $\ell_k^2 = \delta z^\top \hat{\Sigma}^{-1} \delta z / d$ of the representation change caused by changing factor k alone, where $\hat{\Sigma}$ is the covariance of z over the batch; *protect*$[k]$ keeps ℓ_k above a floor. Unlike a total-variance normalisation it cannot be satisfied by rescaling an irrelevant direction (Appendix D.1). The factors are the semantic one (shape, or a different-class image on Colored-CIFAR), the spurious nuisance (object hue, horizontal position, border colour), a neutral nuisance (floor hue, vertical position), and a random pixel direction of matched norm (placebo); all arms share $\lambda = 1$, $\tau = 0.5$, architecture, data and seeds. We wrote the predictions down before running and report the outcome first: **six of the eight predicted comparisons with ERM did not hold, and one of the two pre-registered kill criteria fires as written**. Criterion 1 asked contract[spurious] to beat contract[semantic] on both error measures; at five seeds it wins on the probe (45.0% against 65.7%) and loses on the model's OOD error (69.2% against 68.3%), because the penalty running from the first step stopped four of five contract[spurious] networks from fitting at all. We report it as fired and do not let the post hoc warm-up overturn it. The plan said the orientation claim would then be abandoned; we depart from that deliberately rather than quietly, judging the criterion mis-specified because it scores an arm that never optimised, and withdraw the narrower claim it tested, that contracting a spurious nuisance is beneficial (Appendix D.3). Separately, every penalty on an irrelevant factor moves the error by up to fourteen points, so the "no change" predictions were wrong. What the pre-registered design does establish, and what the theory licenses, is the contrast *between* factors under one fixed operation. One of the five contrasts we report does involve that arm (contract semantic against spurious, +20.7 points) and inherits its optimisation failure; the other four do not.

The same operation, rotated. On 3D Shapes (Table 1, Figure 1c) the two same-operator contrasts are large and opposite. Contracting the semantic factor raises the probe OOD error to 65.7%, against 42.4% for a neutral nuisance and 38.7% for a matched placebo (ERM 28.6%); protecting it lowers the error to 6.4%, against 38.5% for a neutral nuisance. A probe on an untrained encoder scores 60.4%: the nuisance and placebo arms stay inside that floor, the semantic contraction is past it, worse than no training at all. The two protection arms are the learned analogue of the construction Section 4.1 runs exactly: effective ranks 4.07 and 3.93, within 4% and ordered the right way, RankMe ordering them the wrong way (8.3 against 10.7), probe errors differing sixfold, 6.4% against 38.5%. Representations of nearly the same collapse magnitude therefore reach sharply different error; that comparison alone does not show the remaining rank difference is causally irrelevant, since a small change in a scalar can accompany a large change in an outcome. We do not claim identical rank, and Figure 1c plots every arm against its rank. The sharpest form of the point needs no matching at all: under the post hoc warm-up the semantic contraction sits at a *higher* effective rank than ERM (2.94 against 2.34) with twice its probe error (72.3% against 30.4%), so less collapse with more harm, which no monotone reading of collapse amount can produce (Table 8, three seeds). Because the arms share seeds the contrasts are paired: protecting the semantic factor rather than a neutral one is worth -32.0 points of probe error (95% bootstrap $[-35.3, -29.4]$), contracting it rather than a neutral one $+23.3$ ($[+17.0, +28.4]$; Appendix D.19). The semantic-versus-neutral contrast repeats on dSprites, where contracting the semantic factor is the worst arm (65.2% against 59.2% for a neutral

Table 1: **Same regulariser, different direction.** Test error (%) of the model (OOD) and of a probe retrained on decorrelated data (probe). The three data sets use different protocols and are not pooled: 3D Shapes is the pre-registered design (5 seeds, penalty from step 1) and dSprites needs a 6-epoch warm-up, without which no contract arm fits (3 seeds), both at $\rho = 0.99$; Colored-CIFAR is three seeds and five arms at $\rho = 0.9$, with C10-C the mean error over 15 CIFAR-10-C corruptions and C10.1 the error on CIFAR-10.1 (Appendix C). Starred rows are placebos of matched norm; $^{\dagger k}$: k seeds did not fit (ID error above 50%) and are included; last row: the floor of each measure, a probe on an untrained encoder. Chance: 75, 66.7 and 90%. Bold: best per column.

	3D Shapes		dSprites		Colored-CIFAR			
Arm	OOD	probe	OOD	probe	OOD	C10-C	C10.1	probe
ERM	35.2±1.7	28.6±1.5	63.0±1.0	54.8±0.5	44.3±0.4	64.9±0.4	56.6±0.8	28.7±0.5
contract semantic	68.3±4.9	65.7±5.1	65.3±0.2	65.2±0.7	73.1±3.8	79.3±2.7	78.2±3.1	64.4±5.7
contract spurious	69.2±11.1$^{\dagger 4}$	45.0±10.2$^{\dagger 4}$	**46.6±7.0**	41.2±9.3	52.2±7.7	68.3±3.4	62.7±4.8	46.3±8.3
contract neutral	48.8±3.9	42.4±3.7	64.0±0.1	59.2±2.7	–	–	–	–
contract random*	44.6±4.8	38.7±3.7	62.6±1.0	62.4±1.2	–	–	–	–
protect semantic	**22.8±3.2**	**6.4±1.8**	50.5±7.9	38.6±11.5	**40.1±1.7**	**63.2±1.5**	**52.1±1.9**	**27.9±1.3**
protect spurious	31.6±2.2	15.6±1.4	52.1±5.1	**34.6±8.2**	40.8±0.8	64.2±0.6	52.6±0.2	28.6±0.3
protect neutral	45.0±1.7	38.5±2.6	64.5±0.2	64.0±0.3	–	–	–	–
protect random*	47.2±4.7	33.0±3.5	63.7±1.5	52.9±1.3	–	–	–	–
untrained encoder	74.6±0.7	60.4±3.8	–	–	–	–	–	–

one, ERM 54.8%) and protecting it beats protecting a neutral factor (38.6% against 64.0%), though at three seeds its semantic and spurious arms are not separable (38.6 ± 11.5 against 34.6 ± 8.2). Colored-CIFAR has no neutral arm, so it tests the semantic factor against the spurious one only: contracting the semantic factor raises the probe error from 28.7% to 64.4% and the mean corruption error from 64.9% to 79.3%, while protecting it lowers them to 27.9% and 63.2%, the best of the five arms on every measure but by a margin inside one standard deviation. Only the contraction half is decisive there.

What is doing the work? The plan scored the model's own OOD error; we report the probe because it separates the representation from the head, and each contrast holds on both. Across the 41 networks that fit, $\hat{C}_F$ ranks the probe OOD error at $\rho = 0.94$, but this is a *manipulation check*, not a validation: the penalty acts on the secants the index reads, the networks form nine arm-clusters (cluster bootstrap $[0.75, 0.96]$), $\hat{C}_S$ does as well ($\rho = 0.94$ for both, and they agree at $\rho = 1.00$) and the baselines are informative too (-0.69 for effective rank). Four controls, with the penalty annealed over two epochs so that every arm fits, separate that contrast from cheaper explanations; that warm-up is *post hoc*, declared after the pre-registered run failed to fit, and marked wherever it supplies a result (Appendix D.7). Two change the reading. *Collapse in the right direction can be useful* (post hoc): contracting the spurious nuisance gives the most collapsed representation of that factorial (effective rank 1.35 against 2.34) at less than half ERM's error (14.9% against 30.4%), and our index ranks it sixth of nine rather than last, our clearest inversion here. *And it is alignment, not only length*: a term holding the semantic response inside the nuisance span, on the same Fisher-length floor, raises excess alignment from 0.00 to 0.24 and costs 9 to 16 OOD points in every seed, the one result the counterfactual pool cannot produce alone. The other two rule out label supervision and counterfactual exposure. The arms also differ in the size of the input change they ask for (Appendix D.16).

4.3 Q2: WHICH STATISTIC TRACKS FAILURE ACROSS REGIMES, AND WHY?

A diagnostic must also work on networks nobody intervened on. We train unregularised networks in two regimes: *spurious*, where a planted correlation lets a network replace the semantic factor by a nuisance and failure is the representation-level OOD error, and *label-noise*, where the network memorises random labels and failure is the clean-label test error. Each is swept over its strength, the bottleneck width and the weight decay (24 and 12 configurations; 3 seeds on the 3D Shapes spurious sweep, 2 on its label-noise sweep, 1 on dSprites), every statistic computed on training-

distribution inputs at the final checkpoint. Because a correlation pooled over a grid can take a sign no part of the grid has, we read each statistic pooled, inside every stratum of every knob, and after removing what the knobs explain (Table 15, Figure 3); two of the 32 strata are dropped because their outcome does not vary. The stratified reading separates the statistics: over the 30 strata $\hat{C}_F$ is positively associated with failure in every one and every baseline points the wrong way in 9 to 14. Effective rank, RankMe and the participation ratio are positive inside each bottleneck width but negative inside each correlation strength on both spurious sweeps, so the positive pooled value on 3D Shapes ($+0.55$) comes from the grid, not the statistic; NC1 and CDNV are negative in the spurious regime and strongly positive under label noise; total input sensitivity flips hardest, -0.85 and -0.65 to $+0.83$ and $+0.91$. A raw sign count would overstate this: the strata hold 4 to 36 networks, each network sits in two or three of them, and the weakest of ours is $+0.03$. On the same resolvable criterion applied to both sides, our index is positive in 19 of 30 and negative in none, whereas every baseline is both, in different strata. Removing all three knobs at once rather than one at a time leaves our index at -0.10 on the 3D Shapes spurious sweep, its one negative reading anywhere, and not resolvable ($[-0.33, +0.15]$) where NC1 and directional CDNV are, so that adjustment favours them; we therefore claim sign stability *under stratified comparison*, not invariance to every adjustment.

Which ingredient does the work? A direction-resolved diagnostic has two parts, the information given and the summary applied, so we gave the closest published baseline our information. LiDAR (Thilak et al., 2024) whitens a between-group scatter by an augmentation scatter, the pairing of Eq. (1); on our pool it keeps one sign in 13 of the 14 strata of the two 3D Shapes sweeps, against 11 for effective rank and 14 for ours, all three read on the same re-run networks. **Most of the observational advantage is the pool, not the summary**, which we report as a finding rather than a caveat. What the pool does not buy is sensitivity to a uniform semantic contraction, to which an effective rank is exactly invariant and a log-determinant is not: on the Q1 arms LiDAR moves 0.09 across a 39-point change in probe error where $\hat{C}_F$ moves 2.66 nats. On arm means the two are close (0.88 against 0.82) and, with the in-distribution filter applied consistently, LiDAR also picks the best arm, so we claim the sensitivity difference and not that LiDAR chooses wrongly (Appendices D.12, D.13).

Two further readings go the same way: pooling all checkpoints rescues total input sensitivity but not its cross-regime sign, and used for model selection without out-of-distribution data $\hat{C}_F$ picks the best network of three sweeps and the 35th percentile of the fourth, where effective rank never beats the 50th (Appendices D.11, D.8).

5 Discussion, limitations and conclusion

Discussion and limitations. Collapse statistics measure how much a representation compresses; whether that is harmful depends on where, and Eq. (1) makes the two one quantity. Read that way, neural collapse is nuisance collapse with respect to the training labels and the harms reported for self-supervised representations are semantic collapse. We motivate that reinterpretation but do not test it; what the experiments support is narrower: report a collapse statistic with the task it refers to (Appendix D.16). The estimators need single-factor counterfactuals, exactly what is missing in the label-free setting where spectral statistics are used (Locatello et al., 2019), and since those counterfactuals are labelled and decorrelated, wherever the index can be computed the failure measure can be too. The analysis assumes continuous semantic coordinates while two of three experiments have a categorical one; no result bounds a classification error; and the index is relative to a declared intervention set and readout metric and cannot see a fold, a real and measurable failure that it orders backwards (Appendix D.9). Empirically: six of eight pre-registered predictions failed; the useful-collapse case rests on a post hoc control; one kill criterion fired; the matched-collapse evidence is one arm pair on one data set, matched on effective rank only, which NC1 and directional CDNV do separate; most of the cross-regime advantage is the pool rather than our summary, and the nuisance term separating our index from a plain contraction index moves little ($\rho = 1.00$ on the factorial); and all experiments use small networks on 32×32 images, leaving the self-supervised pre-training that motivated the question untested.

Conclusion. Collapse is neither good nor bad in itself. What matters is not how much a representation compressed, but which directions it can no longer tell apart.

AI USE STATEMENT

In this work we used generative AI tools as follows. Among the tasks for which disclosure is required, AI coding agents were used to write, edit and refactor the experiment, analysis and plotting code, and to help interpret intermediate results while the analysis was being built. Among the tasks for which disclosure is recommended, we used generative AI to draft and edit prose and improve readability, to review drafts of this paper and check its sentences against the generated tables, to discuss framing and presentation, to produce and restyle the figures, to search for and format references, and to prepare slides summarising the work for internal meetings. The research question, the theoretical development and the proofs are the authors' own; AI assistance there was limited to checking drafts, suggesting presentation and typesetting. We have reviewed all AI-assisted work. Every table of results in this paper is generated directly from the raw experiment logs by a script rather than transcribed, every AI-written code path was read by the authors, and the numbers quoted in the prose were checked against those tables. We take responsibility for the final content of this work, including all text, claims, figures and code.

ETHICS STATEMENT

This work uses only public data sets (3D Shapes, dSprites, CIFAR-10, CIFAR-10.1 and CIFAR-10-C) and procedurally generated images, and involves no human subjects or personal data. We are not aware of ethical concerns specific to this work.

REPRODUCIBILITY STATEMENT

Section 3 states the assumptions of every formal result and Table 2 summarises what each does and does not claim; Appendix B contains complete proofs. Appendix C specifies every data set, split, planted correlation, architecture, optimiser, schedule, regulariser, estimator and baseline, together with the hardware used. All data are public or generated by the procedure described there, and all reported numbers are computed from the protocol of Section 4 with seeds 0 to 4 on the 3D Shapes factorial and on the Proposition 2 check, seeds 0 to 2 on the 3D Shapes spurious sweep, seeds 0 and 1 on its label-noise sweep, and seed 0 on dSprites. The accelerator's kernels are not bit-deterministic, so a re-run at fixed seeds is an independent replicate rather than a copy; Appendix D.13 reports one for the whole 3D Shapes factorial. Against the five-seed run of Table 1, six of its nine arm means fall within three points and the largest gap is 8.2 points, on the contract[spurious] arm whose networks do not fit; against the matched three seeds, which is the comparison Appendix D.13 makes, every arm mean agrees to within 2.8 points.

REFERENCES

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020. doi: 10.1073/pnas.2015509117. URL https://www.pnas.org/doi/10.1073/pnas.2015509117.

Thomas Laurent, James von Brecht, and Xavier Bresson. Feature collapse. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gctmyMiPHH.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=YevsQ05DEN7.

Simon Kornblith, Ting Chen, Honglak Lee, and Mohammad Norouzi. Why do better loss functions lead to less transferable features? In *Advances in Neural Information Processing Systems*, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/f0bf4a2da952528910047c31b6c2e951-Abstract.html.

Jörn-Henrik Jacobsen, Jens Behrmann, Richard Zemel, and Matthias Bethge. Excessive invariance causes adversarial vulnerability. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BkfbpsAcF7.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann LeCun. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *International Conference on Machine Learning*, pp. 10929–10974, 2023. URL https://proceedings.mlr.press/v202/garrido23a.html.

Tomer Galanti, András György, and Marcus Hutter. On the role of neural collapse in transfer learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=SwIp410B6aQ.

Michael V. Berry and C. Upstill. Catastrophe optics: Morphologies of caustics and their diffraction patterns. In *Progress in Optics*, volume 18, pp. 257–346. Elsevier, 1980. doi: 10.1016/S0079-6638(08)70215-4. URL https://www.sciencedirect.com/science/article/pii/S0079663808702154.

Arthur Jacot, Peter Súkeník, Zihan Wang, and Marco Mondelli. Wide neural networks trained with weight decay provably exhibit neural collapse. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=1HCN4pjTb4.

Peter Súkeník, Christoph Lampert, and Marco Mondelli. Neural collapse vs. low-rank bias: Is deep neural collapse really optimal? In *Advances in Neural Information Processing Systems*, volume 37, pp. 138250–138288, 2024. doi: 10.52202/079017-4388. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f9c2ab8d429044e0c35bcece2ff6d123-Abstract-Conference.html.

Robert Wu and Vardan Papyan. Linguistic collapse: Neural collapse in (large) language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 137432–137473, 2024. doi: 10.52202/079017-4366. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f88cc8930b47a45ec4733123bf3039b9-Abstract-Conference.html.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *IEEE Information Theory Workshop*, pp. 1–5, 2015. doi: 10.1109/ITW.2015.7133169. URL https://ieeexplore.ieee.org/document/7133169.

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(50):1–34, 2018. URL https://jmlr.org/papers/v19/17-646.html.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xm6YD62D1Ub.

Randall Balestriero, Léon Bottou, and Yann LeCun. The effects of regularization and data augmentation are class dependent. In *Advances in Neural Information Processing Systems*, pp. 37878–37891, 2022. doi: 10.52202/068431-2745. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/f73c04538a5e1cad40ba5586b4b517d3-Abstract-Conference.html.

Like Hui, Mikhail Belkin, and Preetum Nakkiran. Limitations of neural collapse for understanding generalization in deep learning. *arXiv preprint arXiv:2202.08384*, 2022. URL https://arxiv.org/abs/2202.08384.

Md Yousuf Harun, Jhair Gallardo, and Christopher Kanan. Controlling neural collapse enhances out-of-distribution detection and transfer learning. In *International Conference on Machine Learning*, volume 267, pp. 22149–22176, 2025. URL https://proceedings.mlr.press/v267/harun25a.html.

Yihao Xue, Eric Gan, Jiayi Ni, Siddharth Joshi, and Baharan Mirzasoleiman. Investigating the benefits of projection head for representation learning. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=GgEAdqYPNA.

Jenelle Feather, Guillaume Leclerc, Aleksander Madry, and Josh H. McDermott. Model metamers reveal divergent invariances between biological and artificial neural networks. *Nature Neuroscience*, 26(11):2017–2034, 2023. doi: 10.1038/s41593-023-01442-0. URL https://www.nature.com/articles/s41593-023-01442-0.

Florian Tramèr, Jens Behrmann, Nicholas Carlini, Nicolas Papernot, and Jörn-Henrik Jacobsen. Fundamental tradeoffs between invariance and sensitivity to adversarial perturbations. In *International Conference on Machine Learning*, pp. 9561–9571, 2020. URL https://proceedings.mlr.press/v119/tramer20a.html.

Tete Xiao, Xiaolong Wang, Alexei A. Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=CZ8Y3NzuVzO.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *European Signal Processing Conference*, pp. 606–610, 2007. URL https://www.eurasip.org/Proceedings/Eusipco/Eusipco2007/Papers/a5p-h05.pdf.

Kumar Krishna Agrawal, Arnab Kumar Mondal, Arna Ghosh, and Blake Richards. α-ReQ: Assessing representation quality in self-supervised learning by measuring eigenspectrum decay. In *Advances in Neural Information Processing Systems*, pp. 17626–17638, 2022. doi: 10.52202/068431-1281. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/70596d70542c51c8d9b4e423f4bf2736-Abstract-Conference.html.

Achleshwar Luthra, Yash Salunkhe, and Tomer Galanti. Directional neural collapse explains few-shot transfer in self-supervised learning. *arXiv preprint arXiv:2603.03530*, 2026. doi: 10.48550/arXiv.2603.03530. URL https://arxiv.org/abs/2603.03530.

Chi-Ning Chou, Artem Kirsanov, Yao-Yuan Yang, and SueYeon Chung. Diagnosing generalization failures from representational geometry markers. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=c2fQBcoKhU.

SueYeon Chung, Daniel D. Lee, and Haim Sompolinsky. Classification and geometry of general perceptual manifolds. *Physical Review X*, 8(3):031003, 2018. doi: 10.1103/PhysRevX.8.031003. URL https://journals.aps.org/prx/abstract/10.1103/PhysRevX.8.031003.

Alessio Ansuini, Alessandro Laio, Jakob H. Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems*, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/cfcce0621b49c983991ead4c3d4d3b6b-Abstract.html.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pp. 9929–9939, 2020. URL https://proceedings.mlr.press/v119/wang20k.html.

Jing Xu and Haoxiong Liu. Quantifying the variability collapse of neural networks. In *International Conference on Machine Learning*, volume 202, pp. 38535–38550, 2023. URL https://proceedings.mlr.press/v202/xu23k.html.

Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=By-7dz-AZ.

Raphael Suter, Djordje Miladinovic, Bernhard Schölkopf, and Stefan Bauer. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 6056–6065. PMLR, 2019. URL https://proceedings.mlr.press/v97/suter19a.html.

Vimal Thilak, Chen Huang, Omid Saremi, Laurent Dinh, Hanlin Goh, Preetum Nakkiran, Joshua M. Susskind, and Etai Littwin. LiDAR: Sensing linear probing performance in joint embedding SSL architectures. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=f3g5XpL9Kb.

Ben Sorscher, Surya Ganguli, and Haim Sompolinsky. Neural representational geometry underlies few-shot concept learning. *Proceedings of the National Academy of Sciences*, 119(43): e2200800119, 2022. doi: 10.1073/pnas.2200800119. URL https://www.pnas.org/doi/10.1073/pnas.2200800119.

Ian J. Goodfellow, Quoc V. Le, Andrew M. Saxe, Honglak Lee, and Andrew Y. Ng. Measuring invariances in deep networks. In *Advances in Neural Information Processing Systems*, volume 22, pp. 646–654, 2009. URL https://proceedings.neurips.cc/paper_files/paper/2009/hash/428fca9bc1921c25c5121f9da7815cde-Abstract.html.

Roman Novak, Yasaman Bahri, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HJC2SzZCW.

Judy Hoffman, Daniel A. Roberts, and Sho Yaida. Robust learning with jacobian regularization. *arXiv preprint arXiv:1908.02729*, 2019. doi: 10.48550/arXiv.1908.02729. URL https://arxiv.org/abs/1908.02729.

Patrice Simard, Bernard Victorri, Yann LeCun, and John Denker. Tangent prop - a formalism for specifying selected invariances in an adaptive network. In *Advances in Neural Information Processing Systems*, volume 4, pp. 895–903, 1991. URL https://proceedings.neurips.cc/paper_files/paper/1991/hash/65658fde58ab3c2b6e5132a39fae7cb9-Abstract.html.

Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In *International Conference on Machine Learning*, pp. 833–840, 2011. URL https://icml.cc/2011/papers/455_icmlpaper.pdf.

Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 2662–2670, 2017. doi: 10.24963/ijcai.2017/371. URL https://www.ijcai.org/proceedings/2017/371.

Damien Teney, Ehsan Abbasnedjad, and Anton van den Hengel. Learning what makes a difference from counterfactual examples and gradient supervision. In *Computer Vision – ECCV 2020*, volume 12355 of *Lecture Notes in Computer Science*, pp. 580–599. Springer, 2020. doi: 10.1007/978-3-030-58607-2_34. URL https://link.springer.com/chapter/10.1007/978-3-030-58607-2_34.

Vishal Rajput. The matching principle: When does a training penalty cover deployment shift? *arXiv preprint arXiv:2605.22800*, 2026. doi: 10.48550/arXiv.2605.22800. URL https://arxiv.org/abs/2605.22800.

Christina Heinze-Deml and Nicolai Meinshausen. Conditional variance penalties and domain shift robustness. *Machine Learning*, 110(2):303–348, 2021. doi: 10.1007/s10994-020-05924-1. URL https://link.springer.com/article/10.1007/s10994-020-05924-1.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations in text classification. In *Advances in Neural Information Processing Systems*, volume 34, pp. 16196–16208, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/8710ef761bbb29a6f9d12e4ef8e4379c-Abstract.html.

Rubén Moreno-Bote, Jeffrey Beck, Ingmar Kanitscheider, Xaq Pitkow, Peter Latham, and Alexandre Pouget. Information-limiting correlations. *Nature Neuroscience*, 17(10):1410–1417, 2014. doi: 10.1038/nn.3807. URL https://www.nature.com/articles/nn.3807.

Ingmar Kanitscheider, Ruben Coen-Cagli, and Alexandre Pouget. Origin of information-limiting noise correlations. *Proceedings of the National Academy of Sciences*, 112(50):E6973–E6982, 2015. doi: 10.1073/pnas.1508738112. URL https://www.pnas.org/doi/10.1073/pnas.1508738112.

Guido Montúfar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *Advances in Neural Information Processing Systems*, 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/hash/fa6f2a469cc4d61a92d96e74617c3d2a-Abstract.html.

Ruili Feng, Kecheng Zheng, Yukun Huang, Deli Zhao, Michael Jordan, and Zheng-Jun Zha. Rank diminishing in deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 33054–33065, 2022. doi: 10.52202/068431-2395. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/d5cd70b708f726737e2ebace18c3f71b-Abstract-Conference.html.

Wojciech Masarczyk, Mateusz Ostaszewski, Ehsan Imani, Razvan Pascanu, Piotr Miłoś, and Tomasz Trzcinski. The tunnel effect: Building data representations in deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 76772–76805, 2023. doi: 10.52202/075280-3355. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f249db9ab5975586f36df46f8958c008-Abstract-Conference.html.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z. URL https://www.nature.com/articles/s42256-020-00257-z.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In *Advances in Neural Information Processing Systems*, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/6cfe0e6127fa25df2a0ef2ae1067d915-Abstract.html.

Mohammad Pezeshki, Sékou-Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. In *Advances in Neural Information Processing Systems*, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/0987b8b338d6c90bbedd8631bc499221-Abstract.html.

Nahal Mirzaie, Alireza Alipanah, Ali Abbasi, Amirmahdi Farzane, Hossein Jafarinia, Erfan Sobhaei, Mahdi Ghaznavi, Amir Najafi, Mahdieh Soleymani Baghshah, and Mohammad Hossein Rohban. Implicit regularization of SGD reduces shortcut learning. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=CPdAB7H8mU.

Cristian Daniel Păduraru, Antonio Bărbălau, Radu Filipescu, Andrei Liviu Nicolicioiu, and Elena Burceanu. Bridging explainability and embeddings: BEE aware of spuriousness. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=9jYpHmI8ot.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Zb6c8A-Fghk.

Vladimir I. Arnold, Sabir M. Gusein-Zade, and Alexander N. Varchenko. *Singularities of Differentiable Maps, Volume 1*. Birkhäuser, 2012. doi: 10.1007/978-0-8176-8340-5. URL https://link.springer.com/book/10.1007/978-0-8176-8340-5.

Joshua Batson, C. Grace Haaf, Yonatan Kahn, and Daniel A. Roberts. Topological obstructions to autoencoding. *Journal of High Energy Physics*, 2021(4):280, 2021. doi: 10.1007/JHEP04(2021)280. URL https://link.springer.com/article/10.1007/JHEP04(2021)280.

Erich L. Lehmann and George Casella. *Theory of Point Estimation*. Springer, 2 edition, 1998. doi: 10.1007/b98854. URL https://link.springer.com/book/10.1007/b98854.

Chris Burgess and Hyunjik Kim. 3D Shapes dataset, 2018. URL https://github.com/google-deepmind/3d-shapes.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pp. 2649–2658, 2018. URL https://proceedings.mlr.press/v80/kim18b.html.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset, 2017. URL https://github.com/google-deepmind/dsprites-dataset.

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

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, pp. 4114–4124, 2019. URL https://proceedings.mlr.press/v97/locatello19a.html.

Jiří Matoušek. *Using the Borsuk–Ulam Theorem*. Springer, 2003. doi: 10.1007/978-3-540-76649-0. URL https://link.springer.com/book/10.1007/978-3-540-76649-0.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do CIFAR-10 classifiers generalize to CIFAR-10? *arXiv preprint arXiv:1806.00451*, 2018. doi: 10.48550/arXiv.1806.00451. URL https://arxiv.org/abs/1806.00451.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. doi: 10.48550/arXiv.1907.02893. URL https://arxiv.org/abs/1907.02893.

A NOTATION

Table 3 lists every symbol used in the paper, in the order in which the symbols are introduced. Table 2 sets out the scope of each formal result.

Table 2: Scope of the formal results. **None of them bounds a classification or out-of-distribution error**: the step from semantic information to downstream failure is a prediction we test, not a theorem. The three parts of Theorem 1 are separated because they are met to different degrees. Propositions 1 and 3 make no distributional assumptions and are checked in Appendices D.15 and D.17.

Result	Assumes	Tested by
Thm 1(ii)	local linear-Gaussian, $\Sigma_n = I$, $\sigma > 0$, $\operatorname{rank} J_S = r$, J_N fixed	§4.1: met exactly, law confirmed
Thm 1(i)	the same, plus $J'_N = J_N Q$ with $\|Q\|_2 \le 1$	§4.1: **not satisfiable** by a representation-space contraction
Thm 1(iii)	J with orthonormal columns	§4.1: **false here** ($a_S \approx 0.99$)
Prop. 2	exact 2D Gaussian construction	Figure 1b
Thm 2	C^3 code, nondegenerate extremum	§D.9, Figure 2

B PROOFS

Proof of Theorem 1. (i) Since $\|Q\|_2 \le 1$ we have $QQ^\top \preceq I$, hence $J_N Q Q^\top J_N^\top \preceq J_N J_N^\top$ and $\Sigma'_z = J_N Q Q^\top J_N^\top + \sigma^2 I \preceq \Sigma_z$. Both matrices are positive definite and matrix inversion is operator antitone, so $\Sigma_z'^{-1} \succeq \Sigma_z^{-1}$. Congruence with J_S preserves the Loewner order, which gives $I'_S = J_S^\top \Sigma_z'^{-1} J_S \succeq J_S^\top \Sigma_z^{-1} J_S = I_S$; the log-determinant is monotone in the Loewner order, so $\mathrm{SID}'_\sigma \le \mathrm{SID}_\sigma$. (ii) $I'_S = B^\top J_S^\top \Sigma_z^{-1} J_S B = B^\top I_S B$, so $\log\det(\sigma^2 I'_S) = \log\det(\sigma^2 I_S) + 2\log|\det B|$ and $\mathrm{SID}'_\sigma = \mathrm{SID}_\sigma - \frac{1}{r}\log|\det B|$. Contracting one semantic direction u by β corresponds to $B = I - (1-\beta)uu^\top$ with $\det B = \beta$. (iii) Let $P_i = I - (1-\beta)e_i e_i^\top$ scale the i-th column of J. Because $J^\top J = I$, $(JP_i)^\top (JP_i) = P_i^2$, whose eigenvalues are $\{\beta^2, 1, \dots, 1\}$ for every i; the singular values of the modified Jacobian therefore do not depend on whether the scaled column is semantic or nuisance. If a semantic column is scaled, (ii) applies. If a nuisance column is scaled, $\mathrm{range}(J'_N) = \mathrm{range}(J_N)$ remains orthogonal to $\mathrm{range}(J_S)$; on $\mathrm{range}(J_N)^\perp$ the matrix Σ_z acts as $\sigma^2 I$, so $I_S = J_S^\top J_S/\sigma^2 = I/\sigma^2$ before and after the change. □

Behaviour of I_S at a semantic overlap degeneracy. With $\Sigma_n = I$ we have $\sigma^2\Sigma_z^{-1} = \sigma^2(J_N J_N^\top + \sigma^2 I)^{-1} \to P_\perp$ as $\sigma \to 0$, where $P_\perp$ is the orthogonal projection onto $\mathrm{range}(J_N)^\perp$, because the two matrices are simultaneously diagonalisable and $\sigma^2/(\lambda+\sigma^2) \to \mathbf{1}[\lambda = 0]$ for each eigenvalue $\lambda \ge 0$ of $J_N J_N^\top$. Hence $\sigma^2 I_S \to J_S^\top P_\perp J_S$. For a unit u, $u^\top J_S^\top P_\perp J_S u = \|P_\perp J_S u\|^2$, which vanishes exactly when $J_S u \in \mathrm{range}(J_N)$; in that case $u^\top I_S u = u^\top J_S^\top \Sigma_z^{-1} J_S u$ stays bounded as $\sigma \to 0$ (it is at most $\|v\|^2$ for any v with $J_S u = J_N v$), while for $P_\perp J_S u \ne 0$ it grows as $\|P_\perp J_S u\|^2/\sigma^2$. Therefore $\det(\sigma^2 I_S) \to \det(J_S^\top P_\perp J_S)$, which is zero exactly at a semantic overlap degeneracy, i.e. $\mathrm{SID}_0 = +\infty$ there and is finite otherwise. □

When the predicate is a dimension count. An overlap degeneracy exists iff $\ker J_S \ne \{0\}$ or $\mathrm{range}(J_S) \cap \mathrm{range}(J_N) \ne \{0\}$. For generic J_S, J_N the two ranges have dimensions $\min(r,d)$ and $\min(q,d)$, and two generic subspaces of $\mathbb{R}^d$ meet nontrivially exactly when their dimensions sum to more than d; if $r > d$ then $\ker J_S \ne \{0\}$ and $\min(r,d) = d$, so the same inequality holds. The predicate is therefore generically equivalent to $\min(r,d) + \min(q,d) > d$. With our $r = 3$ and $q = 5$, dimension alone *forces* an overlap degeneracy for every $d \le 7$. For $d = 16$ and $d = 64$ it does not force one, but genericity is exactly what a trained network need not satisfy, and ours do not: about 99% of the squared semantic secant of an unregularised $d = 16$ network lies in the nuisance span (Appendix D.7). Dimension counting therefore settles the question in one direction only, which is why the graded index at a fixed σ, and not the predicate, is what we measure.

Proof of Proposition 1. At a fold point $F'(s_c) = 0$, so $J_S u = 0$ for the folded direction u and in particular $J_S u \in \mathrm{range}(J_N)$: a fold is an overlap degeneracy there. For the first converse, take

Table 3: Notation used throughout the paper.

Symbol	Meaning
$x \in \mathbb{R}^D$, $z = f(x) \in \mathbb{R}^d$	input and its representation under the encoder f
$x = \psi(s, n)$	local parametrisation of the data manifold
$s \in \mathbb{R}^r$, $n \in \mathbb{R}^q$	semantic coordinates (the label depends on them) and nuisance coordinates (it does not)
$F = f \circ \psi$	representation as a function of (s, n)
$J_S = \partial_s F$, $J_N = \partial_n F$	semantic and nuisance Jacobians; $J = [J_S \; J_N]$
Σ_n, σ^2	local nuisance covariance and representation noise level
$\Sigma_z = J_N \Sigma_n J_N^\top + \sigma^2 I$	representation covariance induced by nuisance and noise
$I_S = J_S^\top \Sigma_z^{-1} J_S$	semantic Fisher information (Eq. 1)
$\mathrm{SID}_\sigma = -\frac{1}{2r} \log\det(\sigma^2 I_S)$	semantic information deficit; $\mathrm{SID}_0 = +\infty$ at an overlap degeneracy
overlap degeneracy	$J_S u \in \mathrm{range}(J_N)$ for some $u \neq 0$: what the index measures
fold caustic	loss of injectivity of F along a semantic coordinate: what Theorem 2 describes, and what the index does *not* measure
C_S, C_N	plain semantic and nuisance contraction indices (log-volume of J_S, J_N)
$\hat{C}_S$, $\hat{C}_N$, $\hat{C}_F$	secant estimators of C_S, C_N and of SID_σ (the subscript F is for Fisher; Section 3.1)
a_S, a_{rand}	fraction of a squared semantic secant, and of a generic representation difference, that lies inside the span of the nuisance secants
u	a semantic direction in the tangent space of the data manifold
$\mathbf{1}[\cdot]$	indicator function, 1 when the condition holds and 0 otherwise
V, W	$\mathrm{range}(J_S)$ and $\mathrm{range}(J_N)$, the semantic and nuisance response subspaces (Section 4.1)
$A = I - (1-\beta)P$	the exact operator intervention of Section 4.1: P an orthogonal projector of rank k, $\beta \in (0, 1]$ the contraction factor
p	local order of an extremum of the code ($p = 2$ for a fold, $p = 1$ for a kink)
α	angle between a fold's kernel direction and the semantic axis (Proposition 2)
κ, s_c, z_c	curvature, location and value of a fold (Theorem 2)
ρ	strength of the planted spurious correlation
η	fraction of training labels replaced by random labels (label-noise regime)
d	bottleneck (representation) width
δz, δx	change of the representation and of the input when one factor changes (a secant)
S, N	matrices of normalised semantic and nuisance secants (Section 3.1)
$\hat{\Sigma}$	covariance of z over a training batch (used by the regularisers)
$\ell_k^2 = \delta z^\top \hat{\Sigma}^{-1} \delta z / d$	Fisher-normalised secant length along factor k
contract[k], protect[k]	regulariser that shrinks ℓ_k, or keeps it above a floor, along factor k
λ, τ	regulariser weight and protection floor (log scale)
θ, K	orientation of the rotating pattern and number of orientation classes ($K = 12$) in Section D.9

$J_S = J_N = e_1$ in $\mathbb{R}^d$: the semantic response lies in the nuisance span, so the predicate holds, while F is injective in s for fixed n. For the second, take $F(s) = s^2$ on $s \in (-a, a) \setminus \{0\}$: $F' \neq 0$ off the origin, so F is a local diffeomorphism and every first-order quantity at those points is that of an injective map, yet F identifies s with $-s$ globally. □

Proof of Proposition 2. Write $q = s\cos\alpha + n\sin\alpha$ and $r = -s\sin\alpha + n\cos\alpha$. Rotations preserve $\mathcal{N}(0, I_2)$, so $(q, r) \sim \mathcal{N}(0, I_2)$ and $z_\alpha = (q^2, r)$ has the same law for every α; any functional of P_z is therefore constant. For the Bayes error, $s = q\cos\alpha - r\sin\alpha$. Given z_α, the value r is known and q is known up to its sign; because q is symmetric and independent of r, both signs have posterior probability $\frac{1}{2}$. The two candidate values are $s_\pm = \pm|q|\cos\alpha - r\sin\alpha$. If they have the same sign, y is determined; otherwise $\Pr(y = 1 \mid z_\alpha) = \frac{1}{2}$. The signs differ iff $|q|\cos\alpha > |r|\sin\alpha$. Writing $(q, r) = \rho(\cos\phi, \sin\phi)$ with ϕ uniform on $[0, 2\pi)$, this is the event $|\tan\phi| < \tan(\frac{\pi}{2} - \alpha)$, whose probability is $\frac{4(\pi/2-\alpha)}{2\pi} = 1 - \frac{2\alpha}{\pi}$. The Bayes error is $\frac{1}{2}(1 - \frac{2\alpha}{\pi}) = \frac{1}{2} - \frac{\alpha}{\pi}$. □

Theorem 2, in full. Let $F : (-a, a) \to \mathbb{R}$ be C^3 along a semantic coordinate with a nondegenerate critical point at s_c ($F'(s_c) = 0$, $F''(s_c) = \kappa \neq 0$), and let s have a density p that is Lipschitz and positive near s_c. With $z_c = F(s_c)$: **(i)** there is a neighbourhood $U \ni s_c$ with $\Pr\big(|F(s) - z_c| \leq u,\ s \in U\big) = 2p(s_c)\sqrt{2u/|\kappa|}\,\big(1+O(\sqrt{u})\big)$, i.e. the representation density contributed by U diverges as $|z - z_c|^{-1/2}$; **(ii)** each z near z_c on the attainable side of the fold has exactly two preimages *in* U, $s_\pm = s_c \pm \sqrt{2|z - z_c|/|\kappa|} + O(|z - z_c|)$, so any decoder errs by at least $\sqrt{2|z - z_c|/|\kappa|}\,(1 + o(1))$ on one branch, and representation noise of size δ is amplified to a semantic error of order $\delta^{1/2}$; **(iii)** if the semantic variable is periodic ($s \in S^1$) and $d = 1$, every continuous F identifies an antipodal pair (Borsuk–Ulam, Matoušek, 2003) and every non-constant C^1 F has at least two critical points; for $d \geq 2$ a fold-free embedding exists. If instead the extremum is a non-smooth *kink*, as when it lies on a ReLU switching boundary, the CDF exponent is 1. Both statements are local: (i) and (ii) describe the branch through U, and other preimages elsewhere in the domain contribute separately.

Proof of Theorem 2. (i) Since $F \in C^3$ and $F''(s_c) = \kappa \neq 0$, we can write $F(s) - z_c = \frac{\kappa}{2}(s - s_c)^2\, h(s)$ with $h \in C^1$ and $h(s_c) = 1$. The one-dimensional Morse lemma gives the C^1 change of variables $t = (s - s_c)\sqrt{h(s)}$, with $t'(s_c) = 1$, in which $F = z_c + \frac{\kappa}{2}t^2$. Because $h \in C^1$ with $h(s_c) = 1$, $t'(s) = 1 + O(s - s_c)$ and hence $s'(t) = 1 + O(t)$; since p is Lipschitz near s_c, $p(s(t))\, s'(t) = p(s_c)\,(1 + O(t))$. Then $|F(s) - z_c| \leq u$ iff $|t| \leq \tau := \sqrt{2u/|\kappa|}$, and $\Pr(|F(s) - z_c| \leq u) = \int_{-\tau}^{\tau} p(s(t))\, s'(t)\, dt = 2\tau\, p(s_c)\,(1 + O(\tau))$. Differentiating in u gives the density $p(s_c)\sqrt{2/|\kappa|}\, u^{-1/2}(1 + o(1))$ on the side of the image. (ii) Inverting the normal form, the two preimages of z are $t_\pm = \pm\sqrt{2(z - z_c)/\kappa}$, i.e. $s_\pm = s_c \pm \sqrt{2|z - z_c|/|\kappa|} + O(|z - z_c|)$. Any decoder $\phi(z)$ satisfies $\max_\pm |\phi(z) - s_\pm| \geq \frac{1}{2}|s_+ - s_-| = \sqrt{2|z - z_c|/|\kappa|}\,(1 + o(1))$. If z_c is observed with noise of size δ, its preimages lie at distance $\sqrt{2\delta/|\kappa|}$ from s_c, whereas at a regular point with $F'(s) = g \neq 0$ they lie at distance $\delta/|g| + O(\delta^2)$. (iii) Let $F : S^1 \to \mathbb{R}$ be continuous and $g(\theta) = F(\theta) - F(\theta + \pi)$. Then $g(\theta + \pi) = -g(\theta)$, so by the intermediate value theorem g vanishes somewhere, i.e. F identifies an antipodal pair (the one-dimensional Borsuk–Ulam theorem). S^1 is compact, so a non-constant F attains a maximum and a minimum at distinct points, at which $F' = 0$ if $F \in C^1$. If these critical points are nondegenerate, (i) and (ii) apply at each of them. For $d \geq 2$, $\theta \mapsto (\cos\theta, \sin\theta)$ is an embedding without critical points. Finally, at a kink extremum $F(s) = z_c - c|s - s_c| + o(|s - s_c|)$ with $c > 0$, $\Pr(|F - z_c| \leq u) = 2p(s_c)u/c\,(1 + o(1))$, i.e. CDF exponent 1. □

Assumptions and how we check them. $\Sigma_n = I$ (Theorem 1) is a choice of units for nuisance variation; in the estimator it is fixed by the one-step convention of Appendix C, and Appendix D.6 reports what depends on it. Theorem 1(i) and (ii) hold the other Jacobian block exactly fixed, which a soft penalty does not do, so no trained arm of Section 4.2 satisfies either hypothesis and the comparison there is matched on the representation's spectrum (Table 7) rather than on the Jacobian. Section 4.1 closes part of that gap directly: a frozen encoder with a fixed post-hoc operator meets (ii) to 10^{-16} and reproduces its law, and shows that (i) cannot be met this way at all. The orthonormal columns of Theorem 1(iii) are a construction, not a claim about trained networks: in our networks a fraction $a_S \approx 0.99$ of the squared semantic response lies inside the nuisance span, so the matched pair reported in Section 4.2 is an empirical analogue, matched on the statistics rather than on the Jacobian. The smoothness and nondegeneracy of Theorem 2 are checked by the measurements themselves: a nondegenerate C^3 critical point has local order 2 and CDF exponent $\frac{1}{2}$, and we measure 1.71 to 1.99 and 0.50 to 0.52. The local linearisation behind Eq. (1) is replaced in the estimators by finite secants, which is why Proposition 2 is stated for a global fold rather than for an infinitesimal one.

C Experimental details

Data. *3D Shapes* (Kim & Mnih, 2018) contains all 480,000 combinations of six factors: floor hue (10 values), wall hue (10), object hue (10), scale (8), shape (4) and orientation (15). We downsample images from 64×64 to 32×32 by 2×2 averaging and verify that the factor index of every image matches the stored factor values. The task is 4-way shape classification. A spurious correlation of strength ρ is planted by assigning, with probability ρ, an object hue from the pair $\{2k, 2k +$

$1\}$ reserved for shape k, and a uniformly random hue otherwise. Every training set has $20{,}000$ images; the in-distribution (ID) test set ($4{,}000$ images) uses the training ρ, the out-of-distribution (OOD) test set ($4{,}000$ images) uses $\rho = 0$. *dSprites* (Matthey et al., 2017) contains all $737{,}280$ combinations of shape (3 values), scale (6), orientation (40), horizontal position (32) and vertical position (32); we downsample the 64×64 binary images to 32×32. The task is 3-way shape classification. The spurious nuisance is the horizontal position, drawn with probability ρ from the ten positions $\{10k, \ldots, 10k + 9\}$ reserved for shape k; the neutral nuisance is the vertical position. Set sizes and protocol match 3D Shapes. *Colored-CIFAR* is CIFAR-10 (Krizhevsky, 2009) with a planted nuisance: a 3-pixel border painted in one of ten colours, drawn from the class-aligned colour with probability ρ and uniformly otherwise. The border is applied after standard augmentation (random crop with 4-pixel reflect padding and horizontal flip). OOD evaluation uses uniformly random border colours on the CIFAR-10 test set, on CIFAR-10.1 (v6) (Recht et al., 2018), and on the 15 corruptions of CIFAR-10-C at severities 3 and 5 (Hendrycks & Dietterich, 2019). The CIFAR-10-C archive was verified by hand against the MD5 checksum published with the dataset before it was extracted. *Forced folds* (Section D.9) use 32×32 grey images of an asymmetric pattern of three Gaussian blobs (radii $7, 7, 6$ pixels at angles $0, 2.2, 4.0$ rad, widths $3.2, 2.2, 1.6$, amplitudes $1, 0.7, 0.5$) rotated by a continuous angle $\theta \sim U[0, 2\pi)$, with nuisance jitter: translation in $[-2, 2]^2$ pixels, scale in $[0.85, 1.15]$, brightness gain in $[0.6, 1]$ and background level in $[0, 0.3]$. The task is 12-way classification of θ.

Architectures and training. 3D Shapes and dSprites: a 4-layer convolutional encoder (channels $32, 64, 128, 128$, kernel 4, stride 2, ReLU), a 256-unit hidden layer and a linear bottleneck of width d (the representation z), followed by a linear classifier; AdamW, learning rate 10^{-3}, batch 256, 8 epochs on 3D Shapes and 16 on dSprites, whose shape task is still unlearned at epoch 4 (8.2% median training error) and fits only by epoch 16 (2.0%). The factorials use $d = 16$ and no weight decay. Colored-CIFAR: ResNet-20 (He et al., 2016) with widths $(16, 32, 64)$ and the 64-dimensional pooled feature as z; SGD with Nesterov momentum 0.9 and weight decay $5 \cdot 10^{-4}$, one-cycle schedule with peak learning rate 0.1, batch 128, 15 epochs, $\rho = 0.9$. Forced folds: a 3-layer convolutional encoder with GELU or ReLU activations, a 256-unit hidden layer and a linear bottleneck of width $d \in \{1, 2\}$; Adam with cosine decay, 4,000 steps of 256 freshly rendered images.

Orientation interventions. For a base batch x we render partners x' that differ in exactly one factor k: shape (semantic), object hue (spurious nuisance) or floor hue (neutral nuisance) on 3D Shapes; a different-class image with the same border colour (semantic) or the same image with a different border colour (planted nuisance) on Colored-CIFAR. The placebo partner is $x + \xi$ with ξ a Gaussian pixel perturbation whose norm matches the treatment it controls for: the nuisance secant norm for contraction and the semantic secant norm for protection. With $\delta z = z(x') - z(x)$ and the second-moment matrix $\hat{\Sigma}$ of the centred representations of the batch, the Fisher-normalised secant length is $\ell^2 = \delta z^\top \hat{\Sigma}^{-1} \delta z / d$ (a ridge of $10^{-3} \operatorname{tr} \hat{\Sigma}/d$ is added, and gradients flow through $\hat{\Sigma}$). *Contract*$[k]$ adds $\lambda \, \mathbb{E}\, \ell^2$ to the cross-entropy; *protect*$[k]$ adds $\lambda \, \mathbb{E} \operatorname{softplus}(\tau - \log(\ell / \|\delta x\|_{\mathrm{rel}}))$, with $\lambda = 1$ and $\tau = 0.5$ for every arm. The arms differ *only* in k. A variant that normalises $\|\delta z\|^2$ by the total variance $\operatorname{tr} \hat{\Sigma}$ is reported as an ablation in Appendix D.1. A Mahalanobis penalty can in principle be satisfied by inflating the representation's variance along the penalised direction instead of shrinking the response; the outcomes rule this out where it would matter, because the arm that contracts the semantic direction loses 37 points of probe accuracy, which an inflated but preserved response could not produce.

Regime sweeps. For Section 4.3 we train without any regulariser. In the spurious regime we use every combination of $\rho \in \{0, 0.9, 0.99\}$, bottleneck width $d \in \{2, 4, 16, 64\}$ and weight decay in $\{0, 10^{-3}\}$ (24 configurations) for 8 epochs (16 on dSprites); failure is the probe OOD error. In the label-noise regime we use $\rho = 0$, a fraction $\eta \in \{0.2, 0.4\}$ of training labels replaced by uniformly random labels, $d \in \{4, 16, 64\}$ and weight decay in $\{0, 5 \cdot 10^{-4}\}$ (12 configurations) for 30 epochs; failure is the test error on clean labels. All statistics are computed on training-distribution inputs at the final checkpoint.

Metrics. Secant indices use 400 (3D Shapes, dSprites) or 300 (Colored-CIFAR) base inputs drawn from the *training* distribution. On 3D Shapes the semantic secants change the shape to each of the three other shapes, and the nuisance secants step object hue, floor hue, wall hue, scale and orien-

tation by one value. On dSprites the semantic secants change the shape to each of the two other shapes, and the nuisance secants step the horizontal and vertical position by two pixels and the scale and orientation by one value. On Colored-CIFAR the semantic secants are three different-class images with the same border colour, and the nuisance secants are three border recolourings plus a flip, a 2-pixel shift and a $1.15\times$ brightness change. Representations are divided by their global scale $\sqrt{\mathbb{E}\|z - \bar{z}\|^2}$ and inputs by theirs. Baselines are computed on 3,000 training-distribution samples: effective rank (Roy & Vetterli, 2007), the exponential entropy of the normalised eigenvalues of the *centred* covariance of z; RankMe (Garrido et al., 2023), the same functional of the *uncentred* singular values of the embedding matrix as its authors define it, which is why the two take different values on one representation (4.07 against 8.32 for protect[semantic]) and can order a pair differently; participation ratio, NC1 (Papyan et al., 2020), class-distance normalised variance (Galanti et al., 2022), our implementation of its directional variant (Luthra et al., 2026) (variance along the axis joining two class means), an LDA effective rank in the spirit of LiDAR (Thilak et al., 2024) with class labels as surrogate classes, and the total sensitivity along random pixel directions (Novak et al., 2018). The representation-level OOD error is the error of a multinomial logistic regression (L-BFGS, $\ell_2 = 10^{-4}$) trained on frozen features of decorrelated data and tested on held-out decorrelated data, in the spirit of last-layer retraining (Kirichenko et al., 2023).

Analysis protocol for the forced folds. Each trained encoder is evaluated on a grid of $8{,}192$ orientations at 40 fixed nuisance draws. *Critical points* are local extrema of $z(\theta)$ found with a standard peak finder on the circular signal; we report both a strict set (topographic prominence above 2% of the range of z, called prominent folds) and a permissive set (above 0.1%). The *local order* at a critical point is the slope of a least-squares line through $\log|z(\theta_c \pm k\Delta\theta) - z(\theta_c)|$ against $\log(k\Delta\theta)$ for 14 logarithmically spaced k between 2 and 150 grid steps (0.0015 to 0.115 rad). The *CDF exponent* is the slope of a least-squares line through $\log \Pr(|z - z_c| \le u)$ against $\log u$ over the 0.1% to 3% quantiles of $|z - z_c|$, which excludes the flat part below the grid resolution. A misclassification at θ_i is a *collision* if the predicted bin contains another exact preimage of $z(\theta_i)$, where the preimages are the level crossings of $z - z(\theta_i)$ computed by linear interpolation on the grid, excluding crossings within 1.5 grid steps of θ_i. We report the collision rate against two nulls. The *uniform* null is the hit rate of a uniformly random wrong bin, $|B_i|/(K-1)$ with B_i the set of bins of the other preimages minus the true bin. That null is too generous, because a classifier's errors are not uniform over wrong bins: they concentrate near decision boundaries, and the second preimage of a fold is close in θ, so a locally confused prediction lands on a preimage for free. The *distance-matched* null keeps how far the prediction missed and randomises only which side it missed on: with d_i the circular distance between the predicted and the true bin, the candidates are the (at most two) bins at distance d_i from the true bin, and the null rate is the fraction of them that lie in B_i. It raises the null from 23% to 34% to 50% to 65% on the circle and from 14% to 22% to 41% to 60% on the interval, and it is the one we quote; the uniform rate is in `t4_summary.json` as `chance_collision`.

Compute. All experiments ran on one laptop (Intel Core i7-1195G7, 32 GB RAM, integrated Intel Iris Xe GPU through PyTorch 2.9.1 with the XPU backend). The GPU does not support float64 or dtype conversion, so all metrics are computed on the CPU in float64. Each logged run records its own wall-clock time; a 3D Shapes training run with an orientation penalty takes 30 to 420 s, a fold network 380 to 710 s, and a Colored-CIFAR arm 30 to 90 min, with several runs sharing the GPU.

D Additional results

D.1 Naive normalisation of the orientation regulariser

Our first implementation of the factorial normalised the secant length by the *total* variance of the representation, $\|\delta z\|^2 / \operatorname{tr}\hat{\Sigma}$, instead of the Fisher (Mahalanobis) length $\delta z^\top \hat{\Sigma}^{-1} \delta z / d$ used in Section 4.2. Table 4 reports this version (3D Shapes, $\rho = 0.9$, seed 0, all nine arms). It was run before the Fisher-normalised factorial and at a weaker correlation, where ERM already almost solves the task, so its numbers are not comparable with Table 1; only the qualitative failure mode matters here. Every contraction arm drives the effective rank towards one, whichever direction it targets: a network can shrink a total-variance-normalised secant by inflating one irrelevant direction, so the penalty induces global dimensional collapse instead of targeted invariance. The naive version there-

fore does not isolate orientation, and we report it only as an ablation. Even so, semantic protection is the only arm that improves on ERM, while protecting a random direction of matched norm destroys training. The Fisher normalisation removes the degenerate solution, because the Mahalanobis length of a secant is invariant to rescaling any direction of z.

Table 4: Naive (total-variance) normalisation of the factorial, 3D Shapes, $\rho = 0.9$, seed 0. Errors in %; $\hat{C}_S$ is the plain semantic contraction index. ERM is the unregularised reference.

Arm	ID error	OOD error	probe OOD error	$\hat{C}_S$	effective rank
ERM	0.4	2.4	1.8	-0.93	1.91
contract semantic	1.6	14.5	8.3	$+1.49$	1.02
contract spurious nuisance	1.4	10.0	9.0	$+0.80$	1.16
contract neutral nuisance	1.7	13.3	7.4	-0.25	1.70
contract random	0.9	6.9	4.5	-0.11	1.33
protect semantic	**0.2**	**1.2**	**0.5**	-1.44	2.16
protect spurious nuisance	1.1	10.7	4.9	-1.10	2.71
protect neutral nuisance	1.4	11.2	7.1	$+0.72$	2.26
protect random	74.5	74.5	46.9	-0.80	3.37

D.2 The exact operator test of Theorem 1

Section 4.1 asks when the hypotheses of Theorem 1 can be met exactly by a fixed linear operator on a frozen encoder's representation, $A = I-(1-\beta)P$. Part (ii) requires P inside $V \cap W^{\perp}$ and part (i) requires P inside $W \cap V^{\perp}$, for $V = \mathrm{range}(J_S)$ and $W = \mathrm{range}(J_N)$; their generic dimensions are $\max(0, r-q)$ and $\max(0, q-r)$, so which part is testable is decided by the declared intervention set rather than by the encoder. We evaluate at every one of the 400 base points of Section 3.1, for three seeds, with the ERM protocol of Section 4.2. Two readings of Table 5 matter. First, where the hypothesis is met the theorem is exact, not approximate: the deviation between the measured change in the deficit and the predicted $(k/r)\log(1/\beta)$ never exceeds $3 \cdot 10^{-10}$ over $4{,}800$ point-β pairs, while the block held fixed moves by at most 10^{-16}. Second, a representation-space contraction of nuisance directions is *not* the operation part (i) covers. It produces $J'_N = J_N Q$ with $\|Q\|_2$ up to 174, and the Loewner comparison $J'_N J'^{\top}_N \preceq J_N J_N^{\top}$ on which the proof depends fails by up to 0.24; the deficit then rises at all $4{,}800$ point-β pairs. This is why a penalty that contracts a nuisance factor can cost semantic information, as contract[neutral] on 3D Shapes and contract[border] on Colored-CIFAR both do (Table 1), and it replaces the generic "cost of a penalty" explanation we previously offered for those arms.

D.3 Pre-registered predictions and outcomes

Before running the Fisher-normalised factorial on 3D Shapes we recorded, in our experiment log, the predicted sign of each arm's effect on OOD error relative to ERM and two kill criteria: the orientation claim would be abandoned if contracting the spurious nuisance were not better than contracting the semantic direction, or if protecting the semantic direction were not better than protecting a random one. Table 6 lists the predictions with the observed outcome. The plan said "about equal" without a tolerance. Rather than score those four predictions with a threshold chosen after seeing the results, we record them as *not confirmed*: every penalty on an irrelevant factor moved the error, so "no change" was wrong whatever tolerance one picks. We apply that rule to every row, which costs us one more: protect[spurious] was predicted "up or equal" and its OOD error fell by 3.6 points, so it is scored not confirmed. An earlier version of this table used a five-point tolerance on every predicate, which scored that row as confirmed while this appendix claimed no tolerance was used; the rule stated here is now the one the table applies. The two kill criteria were fully pre-specified and are scored as written. Six of the eight comparisons with ERM failed and **kill criterion 1 fires**. It asked for contract[spurious] to beat contract[semantic] on both the model's OOD error and the probe error; at five seeds it wins on the probe (45.0% against 65.7%) and loses on the model's OOD error (69.2% against 68.3%), so the conjunction fails. The four contract[spurious] networks that never fit their training data (in-distribution errors 74.9%, 76.4%, 75.2% and 71.6% against a

Table 5: **Theorem 1 tested where its hypotheses hold exactly**, on a frozen encoder (400 base points, 3 seeds, mean over points). The operator is $A = I - (1-\beta)P$ with P of rank k. The fourth column is the largest movement of the block the hypothesis holds fixed. Part (ii) reproduces its predicted law to $3 \cdot 10^{-10}$. Part (i)'s operator has $\|Q\|_2$ up to 174, so its hypothesis $\|Q\|_2 \leq 1$ fails and the bound does not apply: the measured rise is not a counterexample to the theorem but a demonstration that contracting nuisance responses in representation space is a different operation.

Part	contracted subspace	β	fixed block	predicted Δ	measured Δ
(ii) $q=2$	$V \cap W^{\perp}$ $(k=1)$	0.75	$6 \cdot 10^{-17}$	$+0.096$	$+0.096$
		0.50	$8 \cdot 10^{-17}$	$+0.231$	$+0.231$
		0.25	$8 \cdot 10^{-17}$	$+0.462$	$+0.462$
		0.10	$6 \cdot 10^{-17}$	$+0.768$	$+0.768$
(i) $q=5$	$W \cap V^{\perp}$ $(k=2)$	0.75	$4 \cdot 10^{-16}$	≤ 0	$+0.007$
		0.50	$3 \cdot 10^{-16}$	≤ 0	$+0.013$
		0.25	$6 \cdot 10^{-16}$	≤ 0	$+0.017$
		0.10	$4 \cdot 10^{-16}$	≤ 0	$+0.019$
(iii) matched	contract semantic, eff. rank 1.79	0.10	—	$+1.535$	$+0.823 \pm 0.052$
	contract nuisance, eff. rank 2.35	0.10	—	$+0.000$	$+0.093 \pm 0.016$
	contract random, eff. rank 2.34	0.10	—	–	$+0.035 \pm 0.015$

chance level of 75%) are what drags that mean: on the one network that fit, the criterion passes on both measures (49.4% and 42.4%), and under the post hoc warm-up of Appendix D.5, where every arm fits, contract[spurious] beats contract[semantic] by 19.6% against 71.4% on OOD error. We report the criterion as fired because that is what it says, and we do not let the warm-up overturn it. Kill criterion 2 passes on both measures. The plan attached a consequence to criterion 1: the orientation claim "would be abandoned". We do not abandon it, and we record that as a *deviation from the pre-registered analysis plan* rather than quietly narrowing the criterion. Our ground is that the criterion is mis-specified: it scores an arm whose networks never optimised, so it compares a trained representation with four untrained ones and cannot test what it was written to test. A reader who rejects that judgement should read the pre-registered factorial as having falsified the claim. The consequence we do accept is that the factorial does not establish that contracting a spurious nuisance is beneficial. What it does establish is the same-operator contrast between factors, with the caveat that one of the five contrasts we report, contract semantic against spurious, is itself computed on that arm and inherits its optimisation failure; the other four do not.

Table 6: Pre-registered predictions for the 3D Shapes factorial ($\rho = 0.99$, $\lambda = 1$, $\tau = 0.5$, 5 seeds) and outcomes. The plan named the model's own OOD error, so that is what the Outcome column scores; the probe error, which the rest of the paper reports, is shown beside it and is not scored. Values are percentages, mean over seeds; ERM is at 35.2% OOD and 28.6% probe, and the kill rows compare two arms instead. $^{\dagger k}$: k of the five networks did not fit their training data and are included in the mean.

Arm	Prediction	OOD (%)	Outcome	Probe (%)
contract semantic	up	68.3	confirmed	65.7
contract spurious	down or equal	$69.2^{\dagger 4}$	**not confirmed**	45.0
contract neutral	about equal	48.8	**not confirmed**	42.4
contract random	about equal	44.6	**not confirmed**	38.7
protect semantic	down	22.8	confirmed	6.4
protect spurious	up or equal	31.6	**not confirmed**	15.6
protect neutral	about equal	45.0	**not confirmed**	38.5
protect random	about equal	47.2	**not confirmed**	33.0
Kill 1	contract spurious not better than semantic	69.2 vs 68.3	**triggered**	45.0 vs 65.7
Kill 2	protect semantic not better than random	22.8 vs 47.2	not triggered	6.4 vs 33.0

D.4 PER-ARM STATISTICS OF THE 3D SHAPES FACTORIAL

Table 7 lists, for every arm, the in-distribution error that decides the exclusion rule of Table 1, both OOD errors, and every statistic compared in the paper. All networks are included, so the contract-spurious row averages four networks that did not fit their training data; the correlation analyses in Section 4.2 exclude those four, and it is their inclusion that fires kill criterion 1 (Appendix D.3).

Table 7: Every arm of the Fisher-normalised 3D Shapes factorial ($\rho = 0.99$, mean $\pm$ sd). Errors in %. $^{\dagger k}$: k of the seeds did not fit their training data. NC1 and CDNV are ratios of within-class to between-class scatter and are not bounded above, so they span orders of magnitude across these arms (0.02 to 271); that spread is real and is why Section 4.2 restricts the "label-free statistics differ little" comparison to effective rank and RankMe, which are on a bounded scale.

Arm	ID	OOD	probe	$\hat{C}_F$	$\hat{C}_S$	$\hat{C}_N$	e.rank	NC1	dCDNV	sens.
ERM	0.3±0.0	35.2±1.7	28.6±1.5	-0.44	+0.09	+2.40	2.36	0.04	0.010	-1.85
contract semantic	0.6±0.1	68.3±4.9	65.7±5.1	+2.20	+2.69	+4.13	1.38	2.32	0.013	-2.43
contract spurious nuis.$^{\dagger 4}$	60.3±31.9	69.2±11.1	45.0±10.2	+1.24	+1.79	+2.22	2.01	270.74	393.655	-1.12
contract neutral nuis.	0.4±0.0	48.8±3.9	42.4±3.7	+0.19	+0.71	+3.18	1.97	0.06	0.012	-1.88
contract random	0.4±0.1	44.6±4.8	38.7±3.7	+0.13	+0.59	+2.97	2.37	0.06	0.016	-2.30
protect semantic	0.2±0.1	22.8±3.2	6.4±1.8	-2.50	-1.64	+1.24	4.07	0.02	0.004	-0.84
protect spurious nuis.	0.3±0.1	31.6±2.2	15.6±1.4	-2.47	-1.61	+1.31	5.08	0.03	0.005	-1.06
protect neutral nuis.	0.4±0.1	45.0±1.7	38.5±2.6	-0.31	+0.25	+1.31	3.93	34.41	0.013	-2.00
protect random	0.5±0.1	47.2±4.7	33.0±3.5	-0.44	+0.12	+1.65	2.32	0.06	0.008	+0.70

D.5 WARMING UP THE PENALTY

In the pre-registered factorial the penalty is active from the first step, and four of five contract-spurious networks then never fit their training data: suppressing the easy colour cue before any shape feature exists removes the only signal the network has found so far, an optimisation failure that invariant-risk methods avoid by annealing their penalty (Arjovsky et al., 2019). Before running it we declared, in our experiment log, a post hoc variant: the identical factorial with the penalty switched on after two epochs for *all* arms, with the same seeds and hyper-parameters. Table 8 reports both. With the warm-up every network fits, and the ordering is the one Theorem 1 predicts. This variant is post hoc and we do not count it towards the pre-registered predictions; we report it because it identifies the cause of their failure and because it supplies the case that the pre-registered run lacks, a network whose collapse is harmless.

Table 8: Pre-registered factorial (penalty from the first step, 5 seeds) and the declared post hoc variant (penalty after two epochs, 3 seeds), 3D Shapes, $\rho = 0.99$, errors in %. $^{\dagger k}$: k seeds did not fit their training data.

	pre-registered (penalty from step 1)		post hoc (penalty after 2 epochs)			
Arm	OOD	probe	OOD	probe	$\hat{C}_F$	eff. rank
ERM	35.2±1.7	28.6±1.5	36.1±2.3	30.4±3.1	-0.42	2.34
contract semantic	68.3±4.9	65.7±5.1	71.4±1.2	72.3±1.0	+1.71	2.94
contract spurious nuisance	69.2±11.1$^{\dagger 4}$	45.0±10.2$^{\dagger 4}$	19.6±2.1	14.9±5.0	-0.38	1.35
contract neutral nuisance	48.8±3.9	42.4±3.7	37.9±3.2	34.2±2.4	-0.22	2.16
contract random (placebo)	44.6±4.8	38.7±3.7	41.6±1.3	35.4±1.0	-0.04	2.30
protect semantic	22.8±3.2	6.4±1.8	28.5±4.4	14.0±3.6	-2.32	3.71
protect spurious nuisance	31.6±2.2	15.6±1.4	34.4±5.7	17.1±2.8	-2.27	4.40
protect neutral nuisance	45.0±1.7	38.5±2.6	46.2±3.1	41.1±3.6	-0.40	4.99
protect random (placebo)	47.2±4.7	33.0±3.5	44.5±1.1	31.3±2.4	-0.75	2.89

D.6 SENSITIVITY OF THE SEMANTIC INFORMATION DEFICIT

The estimator of Eq. (2) has one free constant, the noise floor σ^2, which decides how much nuisance variation counts as noise. Table 9 repeats the analysis of Section 4.3 with $\sigma^2 = 10^{-1}$ instead of 10^{-2}: the two versions rank networks almost identically, so no conclusion depends on this constant. A second convention matters more: the columns of N are one-step changes of each nuisance factor, which fixes the unit in which nuisance variation is measured, and the columns of S are single class changes. Doubling a nuisance step or dropping a nuisance factor changes $\hat{C}_F$; the index is therefore defined relative to a declared set of factor interventions, which Appendix C lists.

Table 9: Sensitivity of the index to its noise floor: rank agreement between $\hat{C}_F$ at $\sigma^2 = 10^{-2}$ and at 10^{-1}, and the Spearman correlation of each with failure, at the final checkpoint of each sweep.

Regime	n	$\rho(\hat{C}_F^{0.01}, \hat{C}_F^{0.1})$	$\rho(\hat{C}_F^{0.01}, \text{fail})$	$\rho(\hat{C}_F^{0.1}, \text{fail})$	$\rho(\hat{C}_S, \text{fail})$
3D Shapes, spurious	72	+0.99	+0.45	+0.42	+0.40
3D Shapes, label noise	24	+0.97	+0.85	+0.85	+0.65
dSprites, spurious	24	+0.99	+0.88	+0.88	+0.87
dSprites, label noise	12	+0.91	+0.81	+0.87	+0.78

D.7 CONTROLS FOR THE ORIENTATION EFFECT

Table 10 reports the four controls of Section 4.2 on 3D Shapes at $\rho = 0.99$, with the penalty annealed over two epochs so that every arm fits, 3 seeds. The *alignment* block holds the Fisher length of the semantic response above the same floor in both regularised arms and adds, in one of them, a term that keeps the response inside the span of the nuisance responses. We measure the effect with a_S, the fraction of the squared semantic secant that lies in that span, and compare it with a_{rand}, the same fraction for random differences of representations, which says how much of a generic direction the span already absorbs when the representation is low rank. The *exposure* arm uses the same counterfactual partners as the protection arm but adds them to the cross-entropy with their own labels, which is counterfactual data augmentation with the same partner budget and no orientation term. The *transfer* block intervenes on, and evaluates, the object's scale while the training label remains the shape. Two of the four controls are summarised rather than quantified in Section 4.2, so their numbers are here. *Not label supervision*: with the intervened and evaluated factor the object's scale while the label stays the shape, so that no arm sees a scale label, protecting scale lowers a scale probe's error from 81.8% to 52.6%, against 76.7% for a matched placebo; the contract side of that design does not separate from its placebo. *Not a substitute for data*: adding the same counterfactual partners to the cross-entropy with their own labels brings the OOD error to 0.1%, so where labelled counterfactuals exist they are the better remedy and the penalties are a way to read and move the geometry rather than a competitor.

D.8 CHOOSING A MODEL WITH EACH STATISTIC

A diagnostic is used to choose. In each sweep of Section 4.3 we pretend that only training-distribution data are available, pick the network that a statistic calls best (least collapse for effective rank and RankMe, most collapse for NC1 and CDNV, least input sensitivity, smallest index for ours), and report where the picked network falls in the failure distribution of that sweep, as a percentile (0 is the best network of the sweep, 50 the median, 100 the worst). Table 11 collects the result.

D.9 FORCED FOLDS: THE FAILURE THE INDEX CANNOT SEE

A one-dimensional bottleneck cannot represent a periodic semantic variable injectively (Theorem 2(iii)), which turns a design choice into a forced fold. Classifying the orientation of a rotating pattern into 12 bins through $d = 1$ and $d = 2$ bottlenecks (Appendix C, Figure 2), the six $d = 1$ codes carry every predicted signature: local order 1.71 to 1.99, CDF exponent 0.50 to 0.52 against 1 for a kink, and 80% to 96% of errors are collisions between preimages of one code value against 50% to 65% for a distance-matched null. Three controls qualify it and one of our predictions failed, and the $d = 2$ control also doubles capacity, so topology and capacity are not separated

Table 10: Controls on 3D Shapes ($\rho = 0.99$, penalty after a 2-epoch warm-up, 3 seeds, errors in %). a_S: fraction of the squared semantic secant inside the nuisance span; a_{rand}: the same for random representation differences. In the transfer block the column "OOD" is the error of a probe for the scale factor and "probe" the usual shape probe.

Arm	a_S	a_{rand}	$\hat{C}_F$	eff. rank	OOD	probe
alignment at a matched Fisher length, and exposure to counterfactual pairs						
ERM	0.99±0.00	0.94±0.01	-0.42±0.10	2.36±0.08	36.5±2.0	29.1±2.6
protect semantic	0.55±0.05	0.55±0.02	-2.33±0.16	3.72±0.21	27.5±6.3	14.1±5.8
align semantic	0.98±0.00	0.74±0.07	-1.04±0.32	2.73±1.11	40.8±9.4	20.6±3.3
counterfactual augmentation	0.99±0.00	0.99±0.01	-0.90±0.45	1.43±0.43	0.1±0.1	0.1±0.1
transfer: the intervened and evaluated factor (scale) is not the training label						
ERM	0.99±0.00	–	-0.41±0.12	2.35±0.08	81.8±1.3	29.3±1.8
contract scale	1.00±0.00	–	-0.13±0.08	2.10±0.25	87.1±0.2	33.9±0.5
protect scale	0.83±0.04	–	-1.10±0.18	4.28±0.23	52.6±8.0	28.6±3.1
contract random	1.00±0.00	–	-0.08±0.10	2.27±0.07	86.4±0.2	34.4±1.5
protect random	0.89±0.03	–	-0.70±0.15	2.76±0.36	76.7±0.6	31.2±1.0

Table 11: Model selection with each statistic: percentile of the failure measure of the selected network within its sweep (lower is better; 50 is what a coin would give).

Sweep	n	$\hat{C}_F$	$\hat{C}_S$	eff. rank	RankMe	NC1	dCDNV	sensitivity
3D Shapes, spurious	72	35	46	67	78	72	76	92
3D Shapes, label noise	24	0	12	50	46	0	0	4
dSprites, spurious	24	0	0	58	62	92	79	50
dSprites, label noise	12	0	0	50	17	0	0	0

(Appendix D.18). We report this as a scope boundary rather than a third result: it is the one failure our index cannot see (Proposition 1), and a practitioner who needs it must test for it separately.

D.10 Fold diagnostics

Figure 2 shows the three measurements of Section D.9 on the trained one-dimensional codes.

D.11 Which checkpoint the shootout reads

Section 4.3 reads every statistic at the final checkpoint of each run, on the grounds that this is the object a practitioner would measure: a trained model. The choice matters, and it does not favour us in the way one might guess. Pooling every logged checkpoint of every run instead, three or four depending on the sweep, (Table 12) leaves both of our indices positive in all four sweeps, but it also rescues two baselines. CDNV becomes sign-stable only because its dSprites spurious correlation, the one negative entry at -0.22, is no longer distinguishable from zero, and total input sensitivity becomes genuinely and consistently *negative*: pooled over training, a network with a larger total Jacobian norm has lower failure in all four sweeps, which is the opposite of the sign it takes among converged networks (-0.85 and -0.65 in the spurious sweeps, $+0.83$ and $+0.91$ under label noise). Early checkpoints are both insensitive and inaccurate, so pooling them inserts a training-time confound that no statistic can be blamed for. We report the converged analysis in the main text and this one here; the honest statement is that our index keeps its sign under both, and that the cross-regime instability of the baselines is a statement about trained networks.

D.12 Information matching: the pool and the summary are different contributions

Our index queries the network on counterfactual pairs, and the baselines of Section 4.3 see training-distribution samples only. A fair reading of that comparison has to ask what a baseline would do

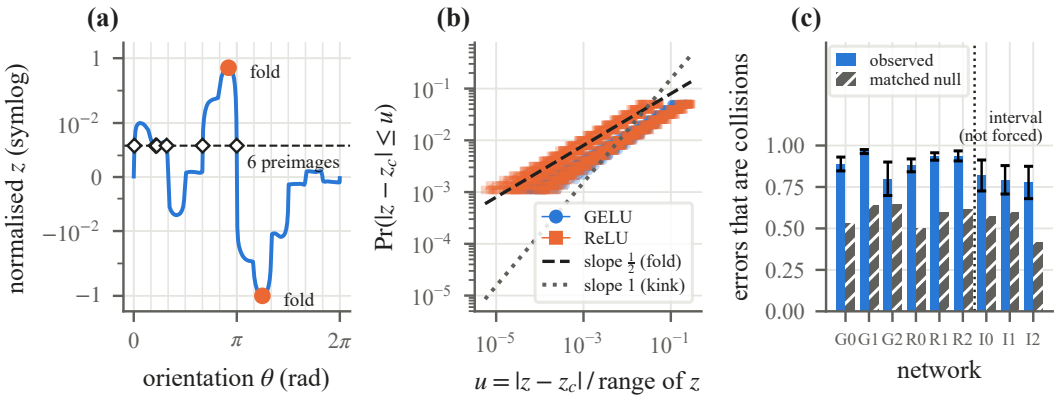


Figure 2: **A forced semantic fold is a caustic.** (a) The learned code $z(\theta)$ of a trained GELU network at one nuisance draw, on a symmetric-log axis because it spans four decades; grey lines are the 12 class boundaries, and one code value (dashed) has six preimages (diamonds) falling in four classes. (b) Cumulative distribution of $|z - z_c|$ near the global extrema for all six $d = 1$ networks (every fifth nuisance draw): both activations follow the fold law with slope $\frac{1}{2}$, not the slope 1 of a kink. (c) Share of errors that are collisions with another exact preimage, against a distance-matched null that misses the true bin by the same distance the network did (Appendix C); the uniform random-wrong-bin null, which is far more generous, is not plotted (G: GELU, R: ReLU, I: the interval control where no fold is forced; digit: seed; error bars: sd over 40 draws).

Table 12: The shootout of Figure 3 recomputed over all logged checkpoints rather than the final one. Entries are Spearman correlations with the regime's failure measure; the last column asks whether the statistic keeps one sign across all four sweeps, counting a correlation whose bootstrap interval covers zero as no evidence of a sign either way.

Statistic	3DS spurious	3DS label noise	dSp spurious	dSp label noise	one sign?
$\hat{C}_F$ (ours)	$+0.62$	$+0.88$	$+0.29$	$+0.87$	yes
$\hat{C}_S$ (ours)	$+0.66$	$+0.84$	$+0.30$	$+0.87$	yes
effective rank	$+0.43$	-0.54	$+0.13$	-0.75	no
RankMe	$+0.17$	-0.34	-0.18	-0.19	no
participation ratio	$+0.44$	-0.56	$+0.13$	-0.76	no
NC1	-0.18	$+0.94$	-0.48	$+0.92$	no
CDNV	$+0.06$	$+0.91$	-0.22	$+0.92$	yes
directional CDNV	0.00	$+0.93$	-0.26	$+0.91$	no
LDA rank	$+0.20$	-0.64	$+0.42$	-0.77	no
total sensitivity	-0.94	-0.36	-0.17	-0.67	yes

with the same pairs. The relevant baseline is LiDAR (Thilak et al., 2024), because it is built on the same pairing: it whitens a between-group scatter by the within-group scatter that augmentations generate, and with the nuisance interventions of Section 3.1 playing the role of augmentations, the matrix it whitens is the one in Eq. (1). What remains different is the summary. LiDAR reports the effective rank $\exp(-\sum_i p_i \log p_i)$ of that matrix, with p_i the eigenvalues normalised to sum to one, and we report $-\frac{1}{2r}\log\det$.

That difference is not cosmetic. Replacing the semantic scatter $S^\top S$ by $c\,S^\top S$ for $c > 0$ multiplies every eigenvalue by c, leaves the normalised spectrum p unchanged, and therefore leaves the effective rank exactly unchanged, while the log-determinant moves by $-\frac{1}{2}\log c$. A uniform contraction of the semantic response is precisely what *contract*[semantic] applies, so the effective-rank summary cannot register the intervention Section 4.2 is built on, at any strength, even when it is given our pool. Table 13 shows this on a constructed example: shrinking the semantic response by a factor of 100 leaves LiDAR on the pool at 2.705, unchanged to three decimals, while $\hat{C}_F$ rises by 4.6 nats. The effective rank does move when the loss is anisotropic, for instance when one of three

semantic directions is removed, which is the case it was designed for. The pool and the summary therefore do different jobs: the pool is what lets a statistic separate semantic from nuisance variation at all, and the log-determinant is what lets it see a loss spread evenly over the semantic directions. Appendix D.13 measures both on trained networks and on the observational sweeps; the division of labour is not the one we assumed, and the pool carries most of the difference. The same table shows the cost of our own summary. A log-determinant is a geometric mean, so it cannot separate a uniform contraction from a concentrated one of the same volume: $\hat{C}_F$ reads -0.96 both for a tenfold shrink of all three semantic directions and for a thousandfold shrink of one of them, where LiDAR separates them (2.705 against 1.733). The two summaries are blind to different things, and neither dominates; what we claim is only that the blindness of an effective rank falls on the case our intervention creates.

Table 13: Two summaries of the same matrix, given the same counterfactual pool. Constructed secants with $d = 16$, $r = 3$ semantic and $q = 5$ nuisance directions, ridge 10^{-2}. An effective rank is a function of the normalised eigenvalue spectrum, so it is invariant to a uniform contraction of the semantic response; a log-determinant is not.

Representation	LiDAR on the pool	$\hat{C}_F$ (ours)
reference representation	2.705	-3.27
semantic response $\times 10^{-1}$	2.705	-0.96
semantic response $\times 10^{-2}$	2.705	$+1.34$
semantic response rotated into $\mathrm{range}(J_N)$	2.510	$+0.48$
one of three semantic directions $\times 10^{-3}$	1.733	-0.96

D.13 The information-matched baseline on trained networks

Table 13 separates the pool from the summary on a constructed example. Here we measure both on trained networks, by re-running the pre-registered factorial and the two 3D Shapes sweeps with `lidar_secant` logged beside $\hat{C}_F$, so that the published baseline reads exactly the counterfactual secants our index reads. An effective rank is a "more is better" statistic, so for LiDAR a *negative* correlation with failure is the correct sign; our index is "more is worse", so ours should be positive.

The pool carries most of the difference, and we did not expect that. Across the 14 evaluable strata of the two sweeps, LiDAR on our pool keeps one sign in 13, against 11 of 14 for effective rank computed the usual way on training samples and 8 of 14 for total input sensitivity; our index keeps one sign in all 14. Giving a published statistic the counterfactual pool therefore takes it most of the way to a direction-resolved diagnostic. This is a concession to the objection that motivated the experiment: the information, not only the geometry we build on it, is doing real work.

What the pool does not buy is the intervention. On the factorial the two part company. Over the 25 networks of the replicate that fit (0.89 here against 0.92 in the original run of the same design; the two differ only by the accelerator's non-determinism), $|\rho|$ with the probe out-of-distribution error is 0.89 for $\hat{C}_F$ and 0.66 for LiDAR on the same pool, and the arm means show why (Table 14). LiDAR on the pool spans 1.127 to 1.558 across the nine arms while $\hat{C}_F$ spans -2.45 to $+2.28$; it reads 1.222 for ERM and 1.127 for the semantic contraction, a gap of 0.10 across a 39-point change in probe error, which is the invariance of Table 13 showing up in trained networks; and it awards its best score to contract[spurious nuisance] at 36.8% probe error when the best arm is protect[semantic] at 5.9%. That last ranking does not survive our own filter, and we withdraw the reading we first drew from it: the arm is carried by the two networks that never fit, and keeping only the networks with in-distribution error below 50%, as we do for the correlations above, moves LiDAR's top arm to protect[semantic], the right choice. On arm means the two summaries are close, $|\rho| = 0.82$ for LiDAR against 0.88 for ours.

The honest summary is that the counterfactual pool is the larger half of what a direction-resolved statistic needs, and the log-determinant is what makes the result usable for choosing between interventions rather than only for ranking unregularised networks.

Table 14: LiDAR computed on our counterfactual pool against our index, on the arms of the pre-registered 3D Shapes factorial (3 seeds, mean). Both statistics read the same secants of the same networks; higher LiDAR and lower $\hat{C}_F$ both mean "healthier". Starred: placebo.

Arm	LiDAR on our pool	$\hat{C}_F$ (ours)	probe OOD error (%)
erm	1.222	-0.38	30.4
contract semantic	1.127	$+2.28$	69.1
contract spurious nuisance	1.558	$+0.60$	36.8
contract neutral nuisance	1.128	$+0.17$	42.0
contract random*	1.151	$+0.21$	42.1
protect semantic	1.310	-2.45	5.9
protect spurious nuisance	1.231	-2.42	15.2
protect neutral nuisance	1.167	-0.28	39.7
protect random*	1.185	-0.45	32.1

An unplanned replication. That re-run used the same seeds as the original factorial, but the XPU kernels are not bit-deterministic, so it is an independent replicate of all 27 networks rather than a copy. It reproduces the original closely: Spearman $+0.94$ between the two runs over the 27 networks on probe out-of-distribution error ($+0.95$ on the model's own OOD error, $+0.94$ on effective rank), mean absolute difference 1.9 points, and Spearman $+0.95$ between the nine arm means. Arm means agree to within 2.8 points, the largest gap being contract[random] at 39.3% against 42.1%. The replicate used the three seeds available at the time and the factorial now uses five, so the two are not compared arm for arm in the main text. The reordering is confined to contract[random], which moves two places, past protect[neutral] and contract[neutral]; these are arms the paper treats as indistinguishable.

D.14 The same comparison inside every stratum of the grid

Each sweep varies three knobs, and a rank correlation pooled over that grid can take a sign that no part of the grid has. Table 15 therefore repeats the comparison of Section 4.3 inside every level of every knob of every sweep: four sweeps, three knobs each, 32 candidate strata, of which 30 have an outcome that varies enough to test anything. We require a range of at least two percentage points, which removes two: the $\rho = 0$ level of the 3D Shapes spurious sweep, where the probe error spans 0.0% to 0.2%, and the $\eta = 0.2$ level of its label-noise sweep, where the clean-label error spans 0.2% to 2.0%. Both belong to 3D Shapes sweeps, so of the 132 networks 96 appear in three strata and 36 in two, for 360 stratum memberships; that dependence is why we report resolvable counts rather than treating the 30 signs as independent. A stratum with fewer than four networks, or one on which a statistic is constant so that its rank correlation is undefined, would also be dropped, and neither case arises here. The pattern that the pooled numbers only suggest is unambiguous here. Effective rank correlates *positively* with failure inside each bottleneck width of the 3D Shapes spurious sweep ($+0.65$, $+0.89$, $+0.90$ at $d = 4, 16, 64$) and *negatively* inside each correlation strength (-0.22 at $\rho = 0.9$, -0.55 at $\rho = 0.99$); its pooled $+0.55$ is an artefact of how the grid distributes width and correlation strength, not a property of the statistic. The participation ratio and RankMe behave the same way, and the label-aware statistics reverse between the two regimes instead.

D.15 What the index does not see: the fold of Proposition 2

The construction of Proposition 2 squares the direction at angle α from the semantic axis, so at $\alpha = 0$ the representation is $z = (s^2, n)$. Its semantic Jacobian is $J_S = (2s, 0)$, which is *large* for a typical s, and its nuisance Jacobian is $(0, 1)$, so $\sigma^2 I_S = 4s^2$ and $\mathrm{SID}_\sigma = -\frac{1}{2}\log(4s^2)$, which is $+\infty$ on the fold line $s = 0$ and negative wherever $|s| > \frac{1}{2}$. At $\alpha = \pi/2$ the semantic response is intact, $\sigma^2 I_S = 1$ and $\mathrm{SID}_\sigma = 0$. Averaged over $s \sim \mathcal{N}(0, 1)$ the first is the smaller of the two, so the index calls the chance-level representation the healthier one. Measuring this over the seven angles (4,000 points each, $\sigma^2 = 10^{-2}$) gives a Spearman correlation with the Bayes error of -0.54 for the infinitesimal index and -0.75 for a secant long enough to cross the fold.

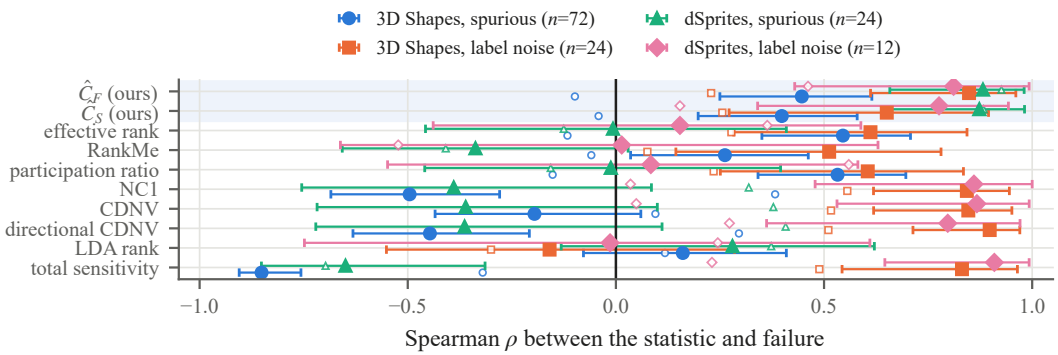


Figure 3: **Which statistic tracks failure in both regimes?** Spearman correlation between each statistic and the regime's failure measure, over unregularised networks at their final checkpoint (large filled markers: pooled over the configuration grid, with bootstrap 95% intervals; small open markers: the same correlation after removing what the training knobs explain). A statistic is useful as a diagnostic only if its association keeps one sign across regimes. The shaded band is our index and its nuisance-free variant.

Table 15: Spearman correlation with failure inside every stratum of every training knob, over the four sweeps (30 strata holding 4 to 36 networks each; every network appears in two or three of them, so the counts are not independent). "Resolvably" counts strata whose 95% bootstrap interval over networks excludes zero, which is how much of the sign count the sample sizes actually support. The last column counts strata whose sign disagrees with the statistic's own majority sign: how often it would have pointed the wrong way.

Statistic	strata	lowest ρ	highest ρ	resolvably + / −	against majority
$\hat{C}_F$ (ours)	30	+0.03	+1.00	19 / 0	**0**
$\hat{C}_S$ (ours)	30	−0.31	+1.00	14 / 0	1
effective rank	30	−0.95	+0.94	9 / 3	10
RankMe	30	−1.00	+0.86	5 / 4	11
participation ratio	30	−0.95	+0.94	10 / 3	9
NC1	30	−0.91	+1.00	7 / 5	13
CDNV	30	−0.83	+1.00	9 / 1	13
directional CDNV	30	−0.83	+1.00	10 / 5	14
LDA rank	30	−0.87	+0.89	2 / 2	13
total sensitivity	30	−0.95	+0.94	7 / 9	13

We report this because it fixes the scope of the index rather than qualifying it. A fold destroys information by identifying two semantic states that are far apart, and a quantity built from first derivatives, or from secants between nearby states, cannot register an identification between distant ones. Contraction and misalignment are what it does register, and those are what Sections 4.2 and 4.3 measure; folds have their own signatures, and Section D.9 uses those. A diagnostic that covered both would need to compare states that are far apart in the semantic coordinate, which is exactly the counterfactual pool a categorical factor supplies and a continuous one does not.

D.16 Two practical readings of the results

Which remedy. The orientation picture orders the interventions we tried. Labelled single-factor counterfactuals used as training data beat every penalty (0.1% on both error measures against 27.5% OOD and 14.1% probe for the best penalty), so where such pairs and their labels exist they are the remedy and the penalties are a way to read and move the geometry rather than a competitor. Without labels for the counterfactuals, protecting the semantic direction is the useful intervention; without either, the orientation analysis is still a measurement of which collapse to worry about. Effective rank orders our factorial reasonably well ($\rho = -0.69$), and it does put the two matched protection

arms the right way round, but by 3.4% of rank against a sixfold difference in error: on the case the whole paper is about its margin carries almost no information.

Which directions count as nuisance. A factor that is a nuisance for the label need not be a nuisance in the representation's input space. Object hue in 3D Shapes does not determine the shape, yet the pixels it changes and those a shape change changes overlap with an intersection over union of 0.49, against 0.11 for floor hue (absolute secant cosine 0.18 against 0.04), so penalising the hue response partly penalises the shape response, and the spurious-nuisance arms behave accordingly. The lesson is that the semantic and nuisance directions of Section 3 are directions in input space, and identifying them with named generative factors is an approximation whose quality can and should be measured.

D.17 What the index is invariant to

Proposition 3 says that SID_σ is unchanged by an orthogonal change of representation coordinates, unchanged by a global rescaling when the readout floor is rescaled with it, and not invariant under a general invertible map with the floor held fixed. The first two are immediate from Eq. (1). For $z' = Uz$ with $U^\top U = I$ we have $S' = US$ and $N' = UN$, so $N'N'^\top + \sigma^2 I = U(NN^\top + \sigma^2 I)U^\top$ and the whole quadratic form $S'^\top (N'N'^\top + \sigma^2 I)^{-1} S' = S^\top (NN^\top + \sigma^2 I)^{-1} S$ is unchanged. For $z' = az$ with $\sigma' = |a|\sigma$, numerator and denominator both scale by a^2 and cancel. For a general invertible A with the floor held at $\sigma^2 I$, the floor no longer transforms with the data and the cancellation fails.

Table 16 confirms this numerically on five independent sets of secants with the shapes used in our experiments ($d = 16$, $r = 3$, $q = 5$), with no retraining: the two exact invariances hold to machine precision and a moderate diagonal rescaling moves the index by up to 0.43 nats. The size of that last number is a choice of how anisotropic a map to apply, and it can be made arbitrarily large, which is the point: the index answers how much semantic information survives at a *stated* readout resolution. We therefore hold σ^2 and the representation's global scale fixed in every comparison, and we read the index only across networks that share an architecture and an output scale.

Table 16: Invariance class of the index, measured on constructed secants with the shapes used in the experiments (five independent sets). The first two rows are exact; the third is the intended dependence on the readout metric, not a defect.

Transformation of z	readout floor	max $\lvert\Delta\ \mathrm{SID}_\sigma\rvert$
orthogonal $z' = Uz$	held at $\sigma^2 I$	$8e-15$
global $z' = az$	rescaled to $\lvert a\rvert\sigma$	$2e-14$
diagonal anisotropic $z' = Az$	held at $\sigma^2 I$	0.43 nats

D.18 Fold controls in detail

Section D.9 states three qualifications compactly; here are the measurements behind them. *Capacity.* With $d = 2$ the circle can be embedded and the same training reaches 1.2% to 1.4% error instead of 14.5% to 38.6% at $d = 1$. That comparison doubles the bottleneck and therefore the capacity at the same time as it removes the topological obstruction, so it does not by itself attribute the difference to topology; we report it as consistent with Theorem 2(iii) rather than as an isolation of it. *Non-periodic control.* On an interval, where no fold is forced, the networks fold anyway: 78% to 82% of errors are collisions against a distance-matched null of 41% to 60%, at a similar error rate. Two of the three interval networks also match the fold law (local order 1.91 and 2.00, CDF exponent 0.50 in both) and the third does not (order 1.61, exponent 0.71); it is the one that formed a single fold rather than three or four, so its extremum is not in the generic quadratic regime the theorem describes. We report all three because the failure is informative: the square-root law is a signature of folds that take the generic form, not an automatic property of every learned one-dimensional code.

D.19 SAME-OPERATOR CONTRASTS WITH PAIRED INTERVALS

Section 4.2 reports the orientation effect as a contrast between arms that share the intervention operator and differ only in the factor it acts on. Because the arms share seeds, the difference is paired, and Table 17 resamples seeds rather than networks to give a bootstrap interval for each contrast ($10{,}000$ resamples). The last column is the collapse-amount mismatch that remains inside each contrast: it is what decides whether the comparison is amount-matched, and we report it rather than asserting that the arms are matched. The protection contrast is the cleanest, with the two arms within a few per cent of each other in effective rank and their probe errors differing by a factor of about six.

Table 17: Same-operator contrasts on the pre-registered 3D Shapes factorial. Δ is the mean paired difference in probe out-of-distribution error, in percentage points, with a 95% bootstrap interval over seeds; a negative value means the semantic arm is better. The last column gives the mean paired difference in effective rank, the collapse amount the contrast does not control.

Same-operator contrast	Δ probe OOD (pts)	95% CI	Δ effective rank
contract semantic vs spurious	$+20.7$	$[+9.5, +31.9]$	-0.63
contract semantic vs neutral	$+23.3$	$[+17.0, +28.4]$	-0.59
contract semantic vs placebo	$+27.0$	$[+22.4, +32.1]$	-0.99
protect semantic vs neutral	-32.0	$[-35.3, -29.4]$	$+0.14$
protect semantic vs placebo	-26.5	$[-30.6, -23.3]$	$+1.75$