ONE-HOT IS ALREADY A GEOMETRY: WHAT A LABEL SPACE BUYS, IN BITS

Anonymous authors
Paper under double-blind review

ABSTRACT

One-hot targets are a choice of geometry: n equidistant points, the regular simplex of neural collapse. Nothing forces that equidistance, and since labels a classifier already confuses could sit close together, placing classes at informative distances is common. But no unit says what that buys, or whether the gain comes from where the points lie or from which label sits where. We define one, in bits. With a frozen encoder and a ridge map onto the label points, scored by distance, the held-out information gain of a label geometry splits exactly into a reference level, an arrangement gain and an assignment gain, each against its own null. Three things follow. Averaging any geometry over all its labellings returns the regular simplex, so one-hot is the non-informative prior over assignment: at fixed temperature it arranges as well as any geometry of the same scale, and assigns nothing at all. An informative geometry trades arrangement for the power to assign. Near the simplex the arrangement gain depends on just two numbers read off the label Gram matrix, its distance from the simplex, and how much of that distance lies on the radii. A second-order expansion weighs them, and although it misses the level, only the ratio of its two weights fixes the order. Both weights come in closed form from the ridge, so screening a candidate geometry is a matrix product, not a training run. Applied to affective and topic classification, those two numbers order candidate geometries where the log-determinant, blind to the radii, fails.

1 INTRODUCTION

One-hot is already a geometry. Take a text classification problem: n classes, one encoder, one head. The routine decision is to encode each class as a one-hot vector, and Figure 1 draws what that decision is. On the left, the target of the regression is the matrix $Z = I$: row i is the point class i is regressed onto, and every pair of rows sits at the same distance. Centred, those n points are the regular simplex (Massion et al., 2026, Prop. III.2), the equiangular tight frame of neural collapse (Papyan et al., 2020); centring changes the origin and not the geometry (Appendix B). On the right, the same matrix Z is built from the class names: each label sits where its name sits among the others, and the distances vary. Both panels are one object, a label geometry, and the left one is the case that rarely registers as a choice.

The right panel differs from the left in two ways. The right panel of Figure 1 places joy beside love, grief beside sadness, and surprise far from both. What it adds over the left panel comes in two kinds. One is *where the points lie* relative to one another, the arrangement: some pairs close, some far, the cloud elongated or round. The other is *which label occupies which point*, the assignment: that joy and not grief sits at the point beside love. In the left panel the second kind does not exist: all points are equidistant, so swapping who sits where changes nothing. The question of this paper is what the right panel buys over the left, in a unit, and how much comes from each kind.

Contributions. Four families of prior work already put information into the labels: label embedding for zero-shot classification (Akata et al., 2015), fixed prototype heads, including heads frozen at the equiangular frame (Hoffer et al., 2018; Mettes et al., 2019; Yang et al., 2022), declared taxonomies, and codes that erase or swap the assignment (Appendix D); neural collapse (Papyan et al., 2020; Han et al., 2022) says where a trained head ends, and $\mathcal{V}$-information (Xu et al., 2020) gives our

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.

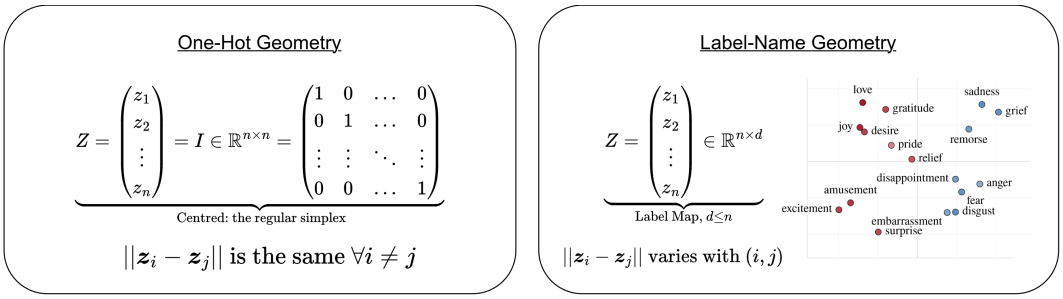


Figure 1: **The target is a matrix, and writing the identity is a choice.** Both panels show the same object, the label geometry $Z \in \mathbb{R}^{n \times d}$ whose row $\boldsymbol{z}_i$ is the point class i is regressed onto. **Left:** $Z = I$, which centred is the regular simplex: every pair of labels at the same distance. **Right:** the label-name geometry of Section 2, in which the distances vary; the inset draws the 17 classes of Section 4 in two dimensions for legibility.

unit. We supply three things. (1) **A unit (Section 2).** The held-out information gain of a label geometry, in bits above the uniform forecast, splits exactly into a reference level, an arrangement gain and an assignment gain, the first the level a random geometry reaches and the other two each read against a declared null. (2) **A zero, and a trade (Section 3).** Averaging any arrangement over every way of assigning its labels returns the regular simplex, so the one-hot design is the non-informative prior over assignment. At a fixed softmax temperature, and attained only when $d \geq n-1$, it collects the largest arrangement gain any geometry of the same scale allows and assigns nothing; leaving it costs a rent that Section 3 names. (3) **A screen of two scalars for the arrangement gain (Section 3).** Near the simplex the arrangement gain depends on two numbers read off the label Gram matrix, its distance from the simplex and how much of that distance lies on the radii, the distances of the label points from their centroid, with two weights in closed form from the ridge, so screening a candidate geometry for its arrangement gain is a matrix product. The screen orders what a candidate gives up; the assignment gain, which decides the trade, is what the fibre null measures, and no scalar of K predicts it. Only the ratio of the weights fixes the order, the expansion misses the level, and Appendix B proves that to second order there is no third number; where the radii carry the ordering the log-determinant is blind to one of the two. Section 4 is the demonstration under the ridge: on the adjudicating task the label-name geometry gives up 0.51 bits of arrangement gain to the standard design and collects 0.62 bits of assignment gain in exchange, mean over four encoders; Appendix C.3 varies the decoder and the task family.

Scope. We evaluate on affective text classification, three tasks and four frozen encoders. The decoder is a ridge with intercept followed by a softmax over squared distances with one temperature, and the encoder is frozen (Appendix C.1 says what an encoder trained on the target changes). What a trained network changes, and whether the construction survives a task outside affect, are the second experiments of Appendix C.3.

2 The measure: reference level, arrangement gain, assignment gain

The geometry is the target, and the admissible set fixes one representative of it. The top-left box of Figure 2 is the choice under study: the target of the regression is a row of $Z \in \mathbb{R}^{n \times d}$, d free; the one-hot code is $Z = I$, which centred has rank $n-1$ and is read at $d = n-1$. The decoder consumes the labelled point configuration (Kendall, 1984, §2) and nothing else, so we fix, without loss of generality, one representative of each: Z centred over its rows and scaled to a fixed trace, and we call the set of such Z the *admissible set*,

$$K = ZZ^\top, \qquad K\mathbf{1} = 0, \qquad \operatorname{tr} K = nd. \tag{1}$$

The diagonal of K holds the squared *radii*, the distances of the label points from their common centroid. Rotation leaves the admissible representative unchanged, translation leaves it unchanged because the ridge below has an intercept and its predictions sum to one, and the scale the trace constraint fixes is absorbed by the temperature below, refit for every geometry.

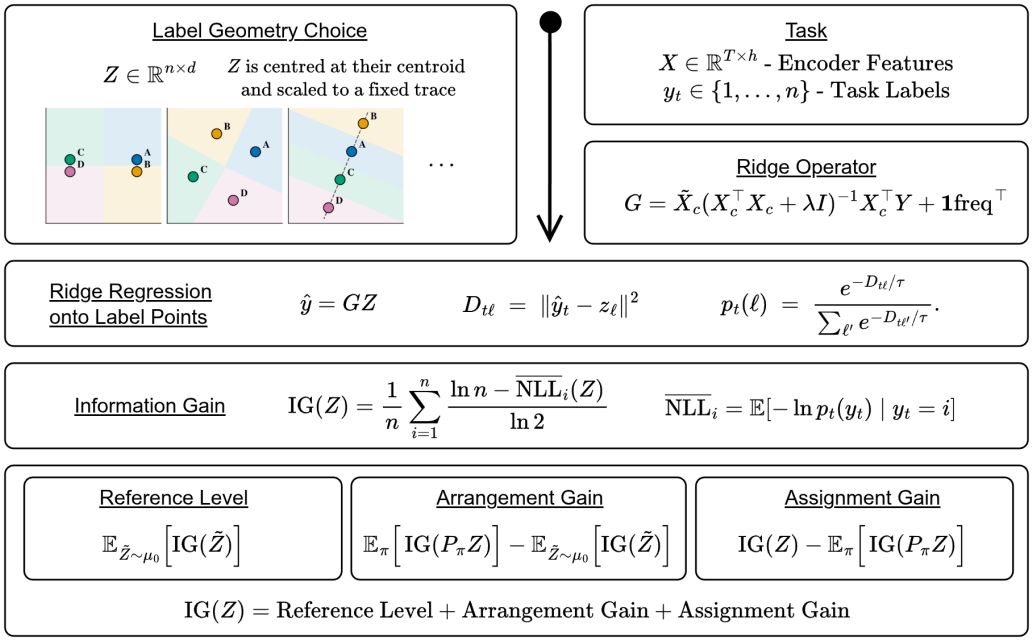


Figure 2: **The map of the measure, read top to bottom.** The label geometry Z (top left) and the task (top right) are independent inputs; the task and the encoder fix the ridge operator G, and Z enters after it, in one product. The prediction is scored by squared distance to each label point and turned into a distribution by one temperature; the information gain averages the surprisal over labels, in bits above the uniform forecast. The three boxes at the foot are the three parts that Figure 3 defines.

The task fixes one operator, and the geometry enters after it. A task is a labelled text corpus with features X from a frozen encoder, and its whole content for the measure is the ridge operator $G \in \mathbb{R}^{T' \times n}$ of Figure 2: the ridge fit once against the one-hot target and evaluated on the T' held-out texts, one row g_t of scores per text. The prediction for any geometry is the one product $\hat{y} = GZ$, and because ridge is linear in its target, fitting it directly onto Z returns the same GZ: the one-hot target privileges nothing, and every assignment and random constellation below is a matrix product on the same G. The squared distance from the prediction to label point ℓ is $D_{t\ell} = \|Z^\top(g_t - e_\ell)\|^2$, and with $v_{t\ell} := g_t - e_\ell$,

$$D_{t\ell} \;=\; v_{t\ell}^\top K\, v_{t\ell}. \tag{2}$$

so the task enters through vectors no geometry moves and the choice through K alone, and every distance is affine in K: this is the hypothesis of every closed form below. The softmax $p_t(\ell) \propto e^{-D_{t\ell}/\tau}$ has one temperature, and two conventions for it travel with every number: *canonical*, refit by maximum likelihood on the training texts for every geometry read, the convention of every measured number; and *fixed*, one τ across the geometries compared, the convention of every theorem. Every theorem below is stated at fixed τ, and each prediction it makes for measured numbers is then checked at canonical τ, with the check named where it is used.

The unit is bits above the uniform forecast, averaged over labels. The unit is the ignorance score (Good, 1952; Roulston & Smith, 2002) and the object is a held-out plug-in estimate of predictive $\mathcal{V}$-information (Xu et al., 2020; Ethayarajh et al., 2022) applied to a label geometry,

$$\mathrm{IG}(Z) \;=\; \frac{1}{n}\sum_{i=1}^{n} \frac{\ln n - \overline{\mathrm{NLL}}_i(Z)}{\ln 2}, \qquad \overline{\mathrm{NLL}}_i = \mathbb{E}\big[-\ln p_t(y_t) \,\big|\, y_t = i\big], \tag{3}$$

where the step through the label makes the aggregation macro. Two properties govern every number below. IG admits negative values, because the decoder is fit by squared error and scored by surprisal. And the score reads $p_t(y_t \mid Z)$ and nothing else, so clumping confusable labels costs nothing where

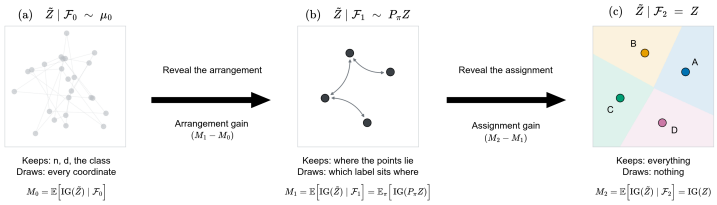


Figure 3: **Three levels of knowledge, one mean each.** The filtration $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2$ reveals first where the points lie, then which label sits where. At each level $\tilde{Z}$ is what the prior still leaves random: (a) nothing is known beyond n, d and the admissible set, and $\tilde{Z} \sim \mu_0$ is a random constellation, with mean M_0; (b) the arrangement is revealed and only the names move, $\tilde{Z} = P_\pi Z$ with π uniform over the $n!$ relabellings, with mean M_1; (c) everything is revealed and the value is $\mathrm{IG}(Z)$.

the decoder did not separate them, and that is where an assignment can earn bits. One number for one geometry does not say what it is made of, and Figure 3 is the construction that splits it.

Three levels of knowledge give three means, and the parts are the steps between them. The prior of the paper is a measure μ_0 on constellations of the admissible set, a Gaussian cloud of n points in d dimensions centred and scaled into it, together with an assignment drawn uniformly from the symmetric group S_n of the $n!$ relabellings, $\mu_0 \otimes \mathrm{Unif}(S_n)$: one draw gives a cloud of n points *and* a way of hanging the names on it. Figure 3 reveals it in three steps, with mean M_0, M_1 and $M_2 = \mathrm{IG}(Z)$. The set of relabellings of one arrangement, $\{P_\pi Z : \pi \in S_n\}$ (P_π a permutation matrix), is its *fibre*, and M_1 is the uniform mean over it; for the one-hot design every relabelling is the same geometry, so its M_1 is its own information gain and its assignment gain is zero by construction. The three means are a Doob martingale (Williams, 1991), and the law given the arrangement is uniform on the fibre because μ_0 is exchangeable.

Each level is a null, and each null concedes something declared. Frame (a) is the first null, the reference μ_0, which concedes only the size of the problem, and every real-data number below uses it. Frame (b) is the label-permutation null, which concedes the arrangement and destroys the assignment, the quadratic-assignment lineage of Mantel (1967), Hubert & Schultz (1976) and Ojala & Garriga (2010).

The identity is exact, and its content is that each part has a null of its own. The three parts are the increments of the martingale,

$$\mathrm{IG}(Z) \;=\; \underbrace{M_0}_{\text{reference level}} \;+\; \underbrace{(M_1 - M_0)}_{\text{arrangement gain}} \;+\; \underbrace{(\mathrm{IG}(Z) - M_1)}_{\text{assignment gain}}, \tag{4}$$

the first under frame (a) and the other two under the two arrows of the figure, and the sum telescopes for every geometry, task and encoder. The residual of equation 4 is checked at the machine floor by a *guard*, an assertion that aborts before any table is written (Appendix B). The identity carries no content by itself; the content is that each part is a difference against a declared null, so each has a sign and a test, and nothing forces a part to be positive: a negative arrangement gain says the arrangement does worse than a random constellation of its dimension, a negative assignment gain that the published pairing does worse than the average pairing of its own points. Revealing the arrangement first is a declared precedence, $\mathsf{arr} \prec \mathsf{asg}$ (Appendix B).

A declared criterion says what counts as zero. Each part is a Monte Carlo estimate. Take a contrast in arrangement gain between two geometries at the same d, and combine the Monte Carlo error of the two fibre means with a bootstrap over the held-out texts; on the contrast between the label-name and the appraisal geometry at $d = 4$ it runs from 0.012 to 0.017 bits across the four encoders named below. The *resolution criterion* of the paper is three of it, one *resolution unit*, reused at every dimension as the declared zero. A difference below three combined standard errors reads as zero and not as "small".

The standard design is the simplex, and what it collects above the reference level is the budget. Raw bits do not answer "is this space good?", because M_0 moves between tasks and between dimensions (Appendix C.1), so bits compare only within one d. Write K_s for the admissible representative of the one-hot design, the centred identity of Section 1 scaled to trace nd, the regular simplex. The *budget* of a task and encoder is $\mathrm{IG}(K_s) - M_0$, what the standard design collects above the reference level, and since its assignment gain is zero the budget is its arrangement gain. Section 3 proves that the budget is the largest arrangement gain the admissible set allows at fixed temperature.

Three affective tasks, four encoders and a roster of four geometries instantiate the measure. The adjudicating task is GoEmotions (Demszky et al., 2020) on 17 single-label classes, replicated on EmpatheticDialogues (Rashkin et al., 2019) and TweetEval emotion (Barbieri et al., 2020) (Appendix C.1); four encoders replicate every measurement, mpnet as the primary one, GTE-large, E5-large-v2 and a TF-IDF control, and a *cell* is one task, one encoder and whatever else a table holds fixed. The affective roster is the one-hot design at $d = n - 1 = 16$; the *label-name geometry*, a classical multidimensional scaling to $d = 16$ of the distances among the label names as embedded by a committee of 12 text encoders that shares no member with the four encoders of the texts (Appendix C.1); the *appraisal geometry* (Fontaine et al., 2007; 2013) at its published $d = 4$; and the *VAD lexicon* (Mohammad, 2018) at $d = 3$. Both come from psychometrics, where the dimensionality of emotional meaning is an open question (Appendix A); sizes, splits, constants and the status of each number are in Appendix C.1.

3 Theory: the simplex as the zero, and a two-scalar predictor

Two scalars order the arrangement gain where the radii carry it, and the log-determinant does not. The log-determinant of the Gram matrix, taken over its $n - 1$ non-zero eigenvalues, is the standard descriptor of "how much a geometry separates". Three scalars of K appear throughout, computed on the admissible representative and divided by the squared trace so that they are dimensionless: the *radius spread*, the variance of the squared radii of Section 2; the *spectral concentration* $\mathrm{tr}\, K^2$; and the log-determinant. The two-scalar coefficient is

$$\bar{V}(K) \;=\; \mathrm{Var}(\mathrm{diag}\, K) + c_G\, \mathrm{tr}\, K^2, \tag{5}$$

with c_G a constant of the task and encoder that the derivation below puts in closed form, and a larger $\bar{V}$ predicts a *smaller* arrangement gain. Section 4 measures which of the three orders the arrangement gain on three candidate families, and on which of them the differences resolve. The radius term carries the order wherever it is measured, and on a family built at one spectrum the log-determinant has no order to give, while the same coefficient misses the *level* by an order of magnitude. Finding the scale on which the same two numbers give a level needs a point to expand around: a zero for the assignment itself, a geometry that assigns nothing.

The mean of the fibre is always the same Gram matrix, and it is the simplex. This uses the admissible set equation 1 and nothing about the task. Relabelling sends K to $P_\pi K P_\pi^\top$ and preserves the admissible set, so the mean over the fibre is a point of the *relaxed set* $\{K \succeq 0,\ K\mathbf{1} = 0,\ \mathrm{tr}\, K = nd\}$, the admissible set with the rank constraint $\mathrm{rank}\, K \le d$ dropped, that every permutation fixes. There is exactly one such point, because a symmetric matrix fixed by all permutations is $aI + b\mathbf{1}\mathbf{1}^\top$ and the two constraints fix a and b. Hence

$$\mathbb{E}_\pi[P_\pi K P_\pi^\top] \;=\; \frac{\mathrm{tr}\, K}{n-1}\, C \;=\; \frac{nd}{n-1}\, C \;=\; K_s, \qquad C = I - \tfrac{1}{n}\mathbf{1}\mathbf{1}^\top, \tag{6}$$

and C is the Gram matrix of the centred regular simplex, the K_s of Section 2. It is the mean under a finite group read as the projection on the fixed-point subspace (Shah & Chandrasekaran, 2012, Lemma 2.2), with the subspace cut to one point by the two constraints; the entry-by-entry check and the guard are in Appendix B. Marginalising who sits where therefore returns the standard design whatever the arrangement one starts from: the simplex is the *barycentre* of the admissible set, the non-informative prior over assignment. K_s has rank $n - 1$, so no configuration of n points in $\mathbb{R}^d$ realises it unless $d \ge n - 1$. A published label space, read against this, is an informative prior over which confusions are cheap. The same lemma says why the zero of this paper and the terminal geometry of unconstrained training coincide. Neural collapse (Papyan et al., 2020; Zhu et al., 2021) finds the class means of a trained network at the equiangular frame, and that result is theirs, including

the uniqueness that the fourth hypothesis below assumes (Han et al., 2022). Under four hypotheses, balanced classes, class means centred and scaled to a fixed trace, $d \geq n-1$, and a unique optimal class-mean Gram, a loss that sees the labels only as a partition is invariant under relabelling, so its optimum is a fixed point of every permutation, and the admissible set has exactly one: the frame is the simplex K_s. What the lemma adds is that uniqueness and symmetry alone, with no reference to the loss, force the frame.

At fixed temperature the simplex is the arrangement ceiling of the admissible set, and assigns nothing. Hypotheses: fixed τ; fixed ridge operator, with its held-out score matrix G of full column rank n; the admissible set with $\operatorname{tr} K = nd$; μ_0 exchangeable and supported on it; M_1 the uniform mean on the fibre.

Theorem 1 (concavity and the ceiling)**.** *At fixed τ, $K \mapsto \mathrm{IG}(K)$ is concave on $\mathrm{Sym}(n)$, and strictly concave along every direction of the admissible set under the rank hypothesis. Hence, for every K of the relaxed set $\{K \succeq 0,\ K\mathbf{1} = 0,\ \operatorname{tr} K = nd\}$,*

$$\textit{arrangement gain}(K) \ \leq\ \mathrm{IG}(K_s) - M_0, \tag{7}$$

with equality if and only if $K = K_s$, a point the admissible set contains only when $d \geq n-1$; the right-hand side is non-negative, by the same Jensen step applied to μ_0; and $\textit{assignment gain}(K_s) \equiv 0$ at every τ, because the fibre of the simplex is a point.

The proof is Jensen on the fibre mean equation 6, with the strictness argument and the guards in Appendix B; the rank hypothesis is itself a guard, and it holds in every encoder. The bound is the budget of Section 2 as a theorem. The simplex discriminates as much as the admissible set allows and knows nothing about what it discriminates. That is the design advice, under the ridge: a head frozen at the equiangular frame collects the largest arrangement gain and leaves the assignment gain on the table, and whether that is a loss is a question about the task, whether the pairing of labels with points must carry information.

The rent is the Jensen gap, and gain plus rent is the budget. Under the hypotheses of Theorem 1, with $\mathrm{IG}(K_s)$ and M_1 read under one and the same τ, the loss from trading the standard design for K is the slack of equation 7 read as a price, the *rent*:

$$\begin{aligned} \textsf{arrangement rent}(K) &= \mathrm{IG}(K_s) - M_1(K) \ \geq\ 0, \\ \textsf{arrangement gain}(K) + \textsf{arrangement rent}(K) &= \mathrm{IG}(K_s) - M_0. \end{aligned} \tag{8}$$

The right-hand side is the budget, a constant of the cell at each d; positivity of the rent is a theorem at fixed τ and a verified prediction at canonical τ (Appendix B). So an informative geometry trades arrangement for the power to assign, and comes out ahead if and only if the *exchange rate* $\eta = \textsf{assignment gain}/\textsf{arrangement rent}$ exceeds one. Equation equation 8 holds at every d; at $d = n-1$ the standard design is itself admissible, so the rent is the price of a trade between two geometries of the same dimension, and below that a gap to a ceiling no geometry of that dimension attains. The ceiling is of the *arrangement*; whether a geometry that also assigns can exceed $\mathrm{IG}(K_s)$ in total is a question equation 8 leaves open.

Near the simplex, two numbers govern the arrangement gain, and to second order there is no third. By equation 8 all the dependence on the geometry sits in the rent, so we expand the rent around the simplex in $\Theta := K - K_s$. The first order vanishes, because its mean over the fibre is $v^\top(\mathbb{E}_\pi K_\pi - K_s)v = 0$ by equation 6: the identity that made the simplex the zero also kills the first order. What remains is a quadratic form in Θ that relabelling leaves invariant, and the quadratic forms on such Θ that the permutation group leaves invariant form a space of dimension two, spanned by the squared distance to the simplex, $\|K - K_s\|_F^2$, and the part of that distance which sits on the diagonal, $\|\operatorname{diag}\Theta\|^2$ (Theorem 7, stated and proved in Appendix B). The two are the scalars of equation 5 under other names: $K\mathbf{1} = 0$ and $\operatorname{tr} K = nd$ fix $\langle K, K_s\rangle = \operatorname{tr} K_s^2$, so $\|K - K_s\|_F^2 = \operatorname{tr} K^2 - \operatorname{tr} K_s^2$, and $\operatorname{diag} K_s$ is constant, so $\|\operatorname{diag}\Theta\|^2 = n\operatorname{Var}(\operatorname{diag} K)$. Hence

$$\textsf{arrangement gain}(K) \ =\ \text{const} \ -\ \tfrac{1}{2}\big[\alpha\, n \operatorname{Var}(\operatorname{diag} K) + \beta \operatorname{tr} K^2\big] \ +\ O(\|\Theta\|^3), \tag{9}$$

where the constant is the budget, α and β are two weights of the cell in closed form from the ridge scores (Appendix B), and $c_G = \beta/(\alpha n)$ in the limit $\tau \to \infty$ is the constant of equation 5. Theorem 1

gives $\beta > 0$; α has the sign of the difference of the two block prices of Appendix B, positive as $\tau \to \infty$ and, by a guard, in every cell measured, where the price ratio at canonical temperature runs from 12 to 34. Screening a candidate geometry for its arrangement gain is therefore one pass over the texts and a matrix product, with no training and no permutation.

What the formula says. A geometry can leave the simplex in two ways, and the formula charges each at its own weight. It can spread its *radii*, some labels farther from the common centre than others; or it can concentrate its *spectrum*, flattening the cloud along some axes at fixed radii. Three things follow. The log-determinant, like every function of the spectrum alone, sees the second kind and not the first, so a family of geometries that differ in their radii at a fixed spectrum leaves it unchanged, which Section 4 tests. The *level* in bits is not recovered: the weights depend on the temperature, c_G is their ratio at infinite temperature and not at the operating one, and the expansion holds only near the simplex, at a distance no published geometry reaches (Section 4). And yet the *order* can survive, because ordering by equation 9 depends on the two weights only through their ratio, so a scale error common to both leaves it intact; Section 4 measures the order with the infinite-temperature ratio c_G.

4 Results: what published geometries deliver, and what they pay for it

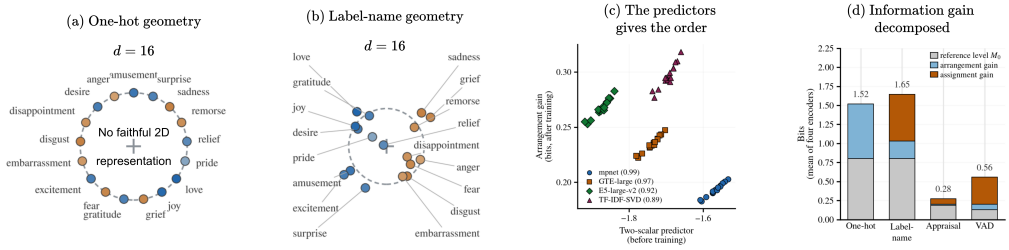


Figure 4: **Two geometries, two readings.** (a) The standard design at $d = n - 1 = 16$, drawn as a ring because all pairwise distances are equal. (b) The label-name geometry at $d = 16$ on its first two axes, over the same 17 classes; colour is valence in the VAD lexicon. (c) The order: measured arrangement gain against the two-scalar coefficient $\bar{V}$ of Section 3 over the 13 variants of the label-name geometry at $d = 4$, the committee and its 12 leave-one-encoder-out subsets; one line per task encoder with its Spearman correlation. (d) What each of the four geometries collects: reference level, arrangement gain and assignment gain, in bits and summing exactly to the information gain written above each bar, mean over the four encoders.

The affective roster is four geometries, each in its own dimension, and one table reads them. Figure 4a and b draw the two geometries the paper opposes, the standard design and the label-name geometry, both at $d = 16$; Table 1 reads the four under the ridge, averaged over the four encoders, and Figure 4d stacks its three parts. Appendix C.2 prints every encoder separately, with the exchange rate, the two scalars and the log-determinant beside each geometry. The trade is read at $d = n - 1 = 16$, where the rent is an identity, and the two scalars at $d = 4$, where the resolution criterion was calibrated.

geometry	d	IG	M_0	arr.	asg.	rent
		(bits, mean over four encoders)				
one-hot (standard)	16	1.520	0.800	0.720	0	0
label-name geometry	16	1.630	0.800	0.214	0.616	0.506
appraisal geometry, 4 axes	4	0.276	0.189	0.011	0.076	—
VAD lexicon	3	0.556	0.131	0.071	0.354	—

Table 1: **GoEmotions under the ridge decoder: what each label geometry collects and what it paid.** IG, M_0, the arrangement gain (arr.) and the assignment gain (asg.) are in bits and sum exactly, each the mean over the four encoders. The rent is printed only at $d = n - 1$, where the standard design is admissible; Appendix C.2 gives it below that dimension, with the exchange rate and every encoder. The standard design assigns nothing and pays no rent by construction.

The unit separates: the label-name geometry collects both parts, the two psychometric geometries collect mostly assignment. The first of three readings of Table 1 is the unit at work. The label-name geometry carries an arrangement gain and an assignment gain that are positive in each of the four encoders (Appendix C.2). Read against the standard design in the same dimension, it gives up seven tenths of the arrangement gain and collects 0.62 bits of assignment in exchange: only the sum of the two says whether the choice was worth it. The two published geometries sit at the other end. Their arrangement gain is close to nothing, and what they collect is assignment. By the resolution criterion of Section 2, the arrangement gain of the appraisal geometry does not resolve from zero in any encoder, while the VAD lexicon clears it with a small gain. Both collect a positive assignment gain in the three neural encoders.

Under the ridge the trade pays in the three neural encoders, and the paired contrast says the same with an interval. The second reading is the trade, read only where the simplex lies inside the admissible set, at $d = n - 1$, where the rent is the identity equation 8. In Table 1 the label-name geometry pays a rent of 0.51 bits and collects 0.62 of assignment, an exchange rate of 1.20 on average; encoder by encoder (Appendix C.2) the rate exceeds one in the three neural encoders and falls below it in the non-neural control, where the assignment the map buys is worth less than the arrangement it sold. The identity carries no interval, so the same trade is read as a paired contrast over the held-out texts: the label-name geometry gains $+0.15$ to $+0.24$ bits of information over the one-hot code in the three neural encoders, each 95% interval excluding zero (Appendix C.2). The map comes out ahead of the standard design where the encoder gives it an assignment worth paying for, so a geometry that assigns can exceed $\mathrm{IG}(K_s)$ in total, which equation 8 left open; the ratio decides the sign. Under a two-layer network on the same features the map's exchange rate falls below one in every encoder and in each of six training seeds (Appendix C.3): the sign of each part travels, the balance does not.

The two scalars order the arrangement gain, including where no spectral function can, and the level is not recovered. The third reading is the coefficient $\bar{V}$ of equation 5, with c_G taken from the ridge in closed form and nothing fitted, read on three candidate sets at $d = 4$. The first is the 13 variants of the label-name geometry, cut to $d = 4$: the committee of 12 encoders and the 12 geometries a practitioner gets by dropping one member. The second is 13 constructions that flatten the smallest axis in steps, collapse nearest pairs, and add the appraisal geometry and a random constellation (Appendix C.1). Table 2 gives the Spearman correlation of each scalar with the measured arrangement gain, signed as the theory fixes; $\bar{V}$ orders above the log-determinant on both sets, and the two sets answer different questions. The 13 constructions move the spectrum by design, and the last row of Table 2 puts their range at 4.7 to 7.3 resolution units, so their order rests on resolved differences, and there every spectral function orders too. The 13 variants differ in their radii and barely in their spectrum, and their whole range sits within 0.6 to 1.3 resolution units, so at most the extreme pair resolves, and only in the non-neural control. What Figure 4c draws is a trend over thirteen dependent points, with a margin over the log-determinant that no single difference among the variants can adjudicate. On all three sets the radius spread alone orders as well as the two scalars; Theorem 7 and the 13 constructions, where the spectral concentration alone reaches 0.79, justify the second term, and no measured margin does. The third set is synthetic and isolates the mechanism. Fixing the spectrum of the label-name geometry and drawing the eigenvectors of K

uniformly gives 48 geometries that share one spectrum and differ in how the label points sit against one another. Every function of the spectrum is constant across the set, so the log-determinant and the spectral concentration cannot order it and Table 2 reads *constant*. The arrangement gain still ranges over 4.8 to 7.7 resolution units, and half to seven tenths of the $\binom{48}{2}$ pairs separate by more than one of them, against at most one pair among the 13 variants. The radius spread orders them at 0.96, beyond the permutation null in all four encoders, and splitting the relabellings in half and ordering the set twice returns that order again, at 0.95 to 0.98. Nobody publishes a random orientation, so the set settles the mechanism and leaves merit to Table 1. The level is not recovered: on the published geometries it is off by more than an order of magnitude, because they sit 1.9 to 2.7 from the simplex where the expansion requires below one (Appendix B).

predictor	committee + 12 subsets	13 constructions	48 orientations
	(Spearman, mean ± sd over four encoders)		
two scalars	0.88 ± 0.09	0.92 ± 0.04	0.96 ± 0.02
radius spread alone	0.90 ± 0.03	0.94 ± 0.04	0.96 ± 0.02
spectral concentration alone	0.12 ± 0.10	0.79 ± 0.02	*constant*
log-determinant	0.06 ± 0.11	0.75 ± 0.04	*constant*
span of the gain, in resolution units	0.6–1.3	4.7–7.3	4.8–7.7

Table 2: **Two scalars of the Gram matrix order the arrangement gain on all three candidate sets, and on the third no spectral function can order at all.** Spearman correlation between each scalar, on the admissible representative and divided by the squared trace, and the arrangement gain at canonical temperature, signed so that positive is the predicted order; mean and sd over the four encoders. The last row gives the range the arrangement gain covers inside each set, in units of the resolution criterion of Section 2. The first set is one committee and its 12 leave-one-out subsets, so its members are strongly dependent. The third fixes one spectrum and draws the orientation, so the spectral scalars take one value and $\bar{V}$ and the radius spread give the same order.

5 Conclusion

A label geometry is a choice, and this paper gives it a unit and a zero. The unit splits the held-out information gain of a geometry, in bits, into a reference level, an arrangement gain and an assignment gain, the first the level of the task and decoder and the other two each read against a declared null. The zero is the regular simplex: averaging any arrangement over all assignments of its labels returns it, so the one-hot code, centred, is the non-informative prior over assignment; at fixed temperature and $d \geq n - 1$ it is also the arrangement ceiling. A geometry that leaves the simplex pays an arrangement rent to gain the power to assign, and near the simplex the arrangement gain depends, to second order, on two numbers read off the Gram matrix, with no third number. On the adjudicating task the label-name geometry pays 0.51 bits of rent for 0.62 bits of assignment gain and comes out ahead of the standard design in the three neural encoders; in the non-neural control, and under a trained two-layer network in every encoder, the assignment it buys is worth less than the arrangement it sells. Where the differences resolve the two scalars order candidate geometries, including a family on which every spectral criterion is constant.

A lens on collapse. The simplex that this paper takes as its zero is also where unconstrained training ends (Papyan et al., 2020), and Section 3 reads the meeting through the same lemma; read through the unit, the frame sits at the arrangement ceiling and assigns nothing. When the scores form a partition of the classes, as they do once a decoder interpolates its training set, every relabelling of a geometry scores the same, because the score averages over labels: the assignment gain vanishes and, at fixed temperature, the simplex becomes the optimum of the whole gain, not only of the arrangement. Collapse to the frame is then the optimum of the sample's partition, and where the task's posterior is not a partition, reaching it would be a form of overfitting: the head keeps the partition the sample shows and discards the confusion structure the population holds. Measured on a network's own features, the held-out assignment gain would tell the two apart; our frozen encoders leave that open.

USE OF LARGE LANGUAGE MODELS

We used a large language model assistant to help develop the conceptual framework, formulate mathematical statements and draft some of their proofs, design the pre-registered experiments, implement the measurement code and the figure scripts, screen the literature for precedence, and draft, translate and edit the manuscript. It generated no data: deterministic code produces the nulls. We verified the assisted work by the mechanism this paper reports on, not by reading alone: every closed form carries an executable guard that aborts before writing a table, every number in the text is read from a released table produced by that code, a released check confirms that each cited value exists in its source table and that no number lacks one, and every citation key passed through a bibliographic ledger. The corpus filters, the derivational bridge of Appendix C.1 and the grids of the ridge penalty and of the temperature live in code written with its help; the choices themselves are ours. We take responsibility for the final content of this work, including text, claims and artefacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The formal statements and their hypotheses are in Sections 2 and 3, with proofs and guards in Appendix B; the data, the splits, the election of the ridge penalty, the temperature and the training constants of the decoder ladder are in Appendix C.1. Every wave aborts by assertion before writing a table: 24 guards on the topic task, with the accounting identity at exactly zero and the trace of the Gram matrix at nd exactly in the eight cells; 15 on the dimension control, which reproduces the 144 cells of the previous wave at the 10^{-16} level; 33 assertions on the partition layer against the exhaustive fibre of the toy. The embeddings are single precision, and the floor of `float32` puts any promise of bit-exact reproduction below about 10^{-7} bits out of reach, since changing the thread count alone moves the information gain at the 10^{-8} level (`tp7_piso_numerico.csv`). The anonymous artefact contains the measurement code, the guards, every table the paper reads and the scripts that fetch the third-party data; it does not redistribute the lexicon or the appraisal coordinates.

ETHICS STATEMENT

This work involves no human subjects. We use four public corpora: GoEmotions (CC BY 4.0, verified), EmpatheticDialogues (CC BY-NC 4.0, academic use), the emotion subset of TweetEval, and 20 Newsgroups. The TweetEval subset repackages the SemEval-2018 affect-in-tweets data under the Twitter terms of service, and we do not redistribute its tweet text; 20 Newsgroups is a public corpus distributed without a formal licence, used only as a labelled corpus with a declared taxonomy, with no text from it reproduced. We use the valence–arousal–dominance lexicon and the appraisal coordinates under their research terms and do not redistribute them. The labels of the affective tasks are English emotion words, and nothing in the paper generalises across languages.

REFERENCES

Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 819–826, 2013. doi: 10.1109/CVPR.2013.111.

Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. doi: 10.1109/CVPR.2015.7298911.

Aaron F. Alexander-Bloch, Haochang Shou, Siyuan Liu, Theodore D. Satterthwaite, David C. Glahn, Russell T. Shinohara, Simon N. Vandekar, and Armin Raznahan. On testing for spatial correspondence between maps of human brain structure and function. *NeuroImage*, 2018. doi: 10.1016/j.neuroimage.2018.05.070.

Sam Allen, Christopher A. T. Ferro, and Frank Kwasniok. A conditional decomposition of proper scores: quantifying the sources of information in a forecast. *Quarterly Journal of the Royal Meteorological Society*, 149:1704–1725, 2023. doi: 10.1002/qj.4478.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1644–1650, 2020. doi: 10.18653/v1/2020.findings-emnlp.148.

Lisa Feldman Barrett. The theory of constructed emotion: an active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, 12(1):1–23, 2017. doi: 10.1093/scan/nsw154.

Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A. Lord. Making better mistakes: Leveraging class hierarchies with deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. doi: 10.1109/CVPR42600.2020.01252.

Jochen Bröcker. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135(643):1512–1519, 2009. doi: 10.1002/qj.456.

Clemens-Alexander Brust and Joachim Denzler. Integrating domain knowledge: Using hierarchies to improve deep classifiers. *Lecture Notes in Computer Science (Pattern Recognition)*, 2020. doi: 10.1007/978-3-030-41404-7_1.

Arthur Charpentier and Agathe Fernandes Machado. Decomposing probabilistic scores: Reliability, information loss and uncertainty, 2026. URL https://arxiv.org/abs/2603.15232.

Shuxiao Chen, Edgar Dobriban, and Jane H. Lee. A group-theoretic framework for data augmentation. *Journal of Machine Learning Research*, 21(245):1–71, 2020. URL https://www.jmlr.org/papers/v21/20-163.html.

Alan S. Cowen and Dacher Keltner. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114(38):E7900–E7909, 2017.

Alan S. Cowen and Dacher Keltner. What the face displays: mapping 28 emotions conveyed by naturalistic expression. *American Psychologist*, 2020. doi: 10.1037/amp0000488.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4040–4054, 2020.

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

Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, 6(3–4):169–200, 1992.

Gamaleldin F. Elsayed and John P. Cunningham. Structure in neural population recordings: an expected byproduct of simpler phenomena? *Nature Neuroscience*, 20(9):1310–1318, 2017. doi: 10.1038/nn.4617.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with V-usable information. In *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5988–6008. PMLR, 2022. URL https://proceedings.mlr.press/v162/ethayarajh22a.html.

Itay Evron, Ophir Onn, Tamar Weiss Orzech, Hai Azeroual, and Daniel Soudry. The role of codeword-to-class assignments in error-correcting codes: An empirical study. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 206 of *Proceedings of Machine Learning Research*, pp. 8053–8077, 2023. URL https://arxiv.org/abs/2302.05334.

Ulrich Faigle and Walter Kern. The Shapley value for cooperative games under precedence constraints. *International Journal of Game Theory*, 21(3):249–266, 1992. doi: 10.1007/BF01258278.

Alan Page Fiske. The lexical fallacy in emotion research: mistaking vernacular words for psychological entities. *Psychological Review*, 127(1):95–113, 2020. doi: 10.1037/rev0000174.

Johnny R. J. Fontaine, Klaus R. Scherer, Etienne B. Roesch, and Phoebe C. Ellsworth. The world of emotions is not two-dimensional. *Psychological Science*, 18(12):1050–1057, 2007.

Johnny R. J. Fontaine, Klaus R. Scherer, and Cristina Soriano (eds.). *Components of Emotional Meaning: A Sourcebook*. Oxford University Press, Oxford, 2013.

Johnny R. J. Fontaine, Christelle Gillioz, Cristina Soriano, and Klaus R. Scherer. Linear and non-linear relationships among the dimensions representing the cognitive structure of emotion. *Cognition and Emotion*, 36(3):411–432, 2022. doi: 10.1080/02699931.2021.2013163.

Dan Friedman and Adji Bousso Dieng. The Vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=g97OHbQyk1.

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*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10929–10974. PMLR, 2023. URL https://proceedings.mlr.press/v202/garrido23a.html.

I. J. Good. Rational decisions. *Journal of the Royal Statistical Society: Series B*, 14(1):107–114, 1952. doi: 10.1111/j.2517-6161.1952.tb00104.x.

X. Y. Han, Vardan Papyan, and David L. Donoho. Neural collapse under MSE loss: Proximity to and dynamics on the central path. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=w1UbdvWH_R3.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275.

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher D. Manning. Conditional probing: measuring usable information beyond a baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1626–1639, 2021. doi: 10.18653/v1/2021.emnlp-main.122.

Elad Hoffer, Itay Hubara, and Daniel Soudry. Fix your classifier: the marginal value of training the last weight layer. In *International Conference on Learning Representations*, 2018. doi: 10.48550/arXiv.1801.04540.

Geoff Hollis and Chris Westbury. The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin & Review*, 23(6):1744–1756, 2016.

Lawrence J. Hubert and James V. Schultz. Quadratic assignment as a general data analysis strategy. *British Journal of Mathematical and Statistical Psychology*, 29:190–241, 1976. doi: 10.1111/j.2044-8317.1976.tb00714.x.

David G. Kendall. Shape manifolds, procrustean metrics, and complex projective spaces. *Bulletin of the London Mathematical Society*, 16(2):81–121, 1984. doi: 10.1112/blms/16.2.81.

Aris Kosmopoulos, Ioannis Partalas, Eric Gaussier, Georgios Paliouras, and Ion Androutsopoulos. Evaluation measures for hierarchical classification: a unified view and novel approaches. *Data Mining and Knowledge Discovery*, 29(3):820–865, 2015. doi: 10.1007/s10618-014-0382-x.

Nikitas C. Koussis, James C. Pang, Richa Phogat, Jayson Jeganathan, Bryan Paton, Alex Fornito, P. A. Robinson, Bratislav Misic, and Michael Breakspear. Generation of surrogate brain maps preserving spatial autocorrelation through random rotation of geometric eigenmodes. *Imaging Neuroscience*, 3, 2025. doi: 10.1162/IMAG.a.71. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC12330862/.

Kuppens, Tuerlinckx, Russell, and Barrett. The relation between valence and arousal in subjective experience. *Psychological Bulletin*, 139(4):917–940, 2013. doi: 10.1037/a0030811.

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. Do supervised distributional methods really learn lexical inference relations? In *NAACL-HLT 2015*, pp. 970–976, 2015. doi: 10.3115/v1/N15-1098.

Zexi Li, Xinyi Shang, Rui He, Tao Lin, and Chao Wu. No fear of classifier biases: Neural collapse inspired federated learning with synthetic and fixed classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. doi: 10.1109/ICCV51070.2023.00490.

Nathan Mantel. The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2):209–220, 1967.

Haggai Maron, Heli Ben-Hamu, Nadav Shamir, and Yaron Lipman. Invariant and equivariant graph networks. In *International Conference on Learning Representations*, 2019. doi: 10.48550/arXiv.1812.09902.

Bastien Massion, Roy Makhlouf, and Estelle Massart. The role of class encoding in neural collapse, 2026. URL https://arxiv.org/abs/2606.00344.

Albert Mehrabian. Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292, 1996. doi: 10.1007/bf02686918.

Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka. Distance-based image classification: Generalizing to new classes at near-zero cost. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013. doi: 10.1109/TPAMI.2013.83.

Pascal Mettes, Elise van der Pol, and Cees G. M. Snoek. Hyperspherical prototype networks. In *Advances in Neural Information Processing Systems 32*, pp. 1487–1497, 2019.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *EMNLP 2022*, pp. 11048–11064, 2022. doi: 10.18653/v1/2022.emnlp-main.759.

Saif M. Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 174–184, 2018.

Markus Ojala and Gemma C. Garriga. Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, 11:1833–1863, 2010. doi: 10.5555/1756006.1859913.

Charles E. Osgood, George J. Suci, and Percy H. Tannenbaum. *The Measurement of Meaning*. University of Illinois Press, 1957.

Art B. Owen. Sobol' indices and Shapley value. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):245–251, 2014. doi: 10.1137/130936233.

Mark Palatucci, Dean Pomerleau, Geoffrey E. Hinton, and Tom M. Mitchell. Zero-shot learning with semantic output codes. In *Advances in Neural Information Processing Systems 22*, pp. 1410–1418, 2009. URL https://proceedings.neurips.cc/paper/2009/hash/1543843a4723ed2ab08e18053ae6dc5b-Abstract.html.

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.

Amey P. Pasarkar and Adji Bousso Dieng. Cousins of the Vendi score: A family of similarity-based diversity metrics for science and machine learning. In *International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3808–3816. PMLR, 2024. URL https://proceedings.mlr.press/v238/pasarkar24a.html.

Robert Plutchik. A general psychoevolutionary theory of emotion. In Robert Plutchik and Henry Kellerman (eds.), *Emotion: Theory, Research, and Experience, Vol. 1*, pp. 3–33. Academic Press, New York, 1980.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: a new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5370–5381, 2019. doi: 10.18653/v1/P19-1534.

Mark S. Roulston and Leonard A. Smith. Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, 130:1653–1660, 2002. doi: 10.1175/1520-0493(2002)130⟨1653:EPFUIT⟩2.0.CO;2.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *European Signal Processing Conference (EUSIPCO) 2007*, pp. 606–610, 2007. doi: 10.5281/zenodo.40328.

Russell and Mehrabian. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273–294, 1977. doi: 10.1016/0092-6566(77)90037-x.

James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39 (6):1161–1178, 1980.

Mohammad Saeed Ebrahimi Saadabadi, Ali Dabouei, Sahar Rahimi Malakshan, and Nasser M. Nasrabadi. Hyperspherical classification with dynamic label-to-prototype assignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. doi: 10.1109/CVPR52733.2024.01641.

Klaus R. Scherer. What are emotions? And how can they be measured? *Social Science Information*, 44(4):695–729, 2005.

Parikshit Shah and Venkat Chandrasekaran. Group symmetry and covariance regularization. *Electronic Journal of Statistics*, 6:1600–1640, 2012. doi: 10.1214/12-ejs723.

Lloyd S. Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games, Volume II*, pp. 307–317. Princeton University Press, 1953. doi: 10.1515/9781400881970-018.

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems 30*, 2017. doi: 10.48550/arXiv.1703.05175.

Eunhye Song, Barry L. Nelson, and Jeremy Staum. Shapley effects for global sensitivity analysis: Theory and computation. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):1060–1083, 2016. doi: 10.1137/15M1048070.

van Heijst, Kret, and Ploeger. Basic emotions or constructed emotions: Insights from taking an evolutionary perspective. *Perspectives on Psychological Science*, 2023. doi: 10.1177/17456916231205186.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4):1191–1207, 2013.

David Williams. *Probability with Martingales*. Cambridge University Press, 1991. doi: 10.1017/CBO9780511813658.

Anderson M. Winkler, Matthew A. Webster, Diego Vidaurre, Thomas E. Nichols, and Stephen M. Smith. Multi-level block permutation. *NeuroImage*, 123:253–268, 2015. doi: 10.1016/j.neuroimage.2015.05.092.

Wu and Mondelli. Neural collapse beyond the unconstrained features model: Landscape, dynamics, and generalization in the mean-field regime. In *International Conference on Machine Learning*, 2025. URL `https://proceedings.mlr.press/v267/wu25u.html`.

Cinna Wu, Mark Tygert, and Yann LeCun. A hierarchical loss and its problems when classifying non-hierarchically. *PLOS ONE*, 2019. doi: 10.1371/journal.pone.0226222.

Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning - a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2251–2265, 2019. doi: 10.1109/TPAMI.2018.2857768.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/2002.10689.

Yibo Yang, Shixiang Chen, Xiangtai Li, Liang Xie, Zhouchen Lin, and Dacheng Tao. Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network? In *Advances in Neural Information Processing Systems 35*, 2022. doi: 10.52202/068431-2753.

Michelle Yik, James A. Russell, and James H. Steiger. A 12-point circumplex structure of core affect. *Emotion*, 11(4):705–731, 2011. doi: 10.1037/a0023980.

Jinxin Zhou, Xiao Li, Tianyu Ding, Chong You, Qing Qu, and Zhihui Zhu. On the optimization landscape of neural collapse under MSE loss: Global optimality with unconstrained features. In *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27179–27202. PMLR, 2022. URL https://proceedings.mlr.press/v162/zhou22c.html.

Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. In *Advances in Neural Information Processing Systems 34*, 2021. doi: 10.48550/arXiv.2105.02375.

A EMOTION SPACES: A SHORT HISTORY, AND THE DIMENSIONALITY DEBATE

A label geometry is a set of coordinates, one point per class, onto which the decoder of Section 2 regresses; d is the dimension of the space those points live in. The roster of Section 4 holds four of them over the same 17 classes: the one-hot code, the label-name geometry, the appraisal geometry at $d = 4$ and the VAD lexicon at $d = 3$. The classes are emotion words of the GoEmotions corpus, and the 77-term vocabulary the label points are cut from was inherited from psychology rather than written by us. Two of the four geometries were not built for machine learning at all. This appendix is the short history that places them: where each came from, what each fixed before collecting a rating, and what the measure does and does not decide about the dispute.

The question is old, and the field has not closed it. How many dimensions does emotional meaning have, and which ones? Answers on record run from two axes to 27 categories, and they disagree about the kind of object as much as about the count. Opening a study of the dimensions themselves, Fontaine et al. (2022) write that "no consensus has been reached about their number and nature". That sentence comes from inside the tradition that argues for four dimensions, and not from its critics. Three families outline the disagreement: a dimensional one that compresses, a categorical one that counts kinds, and a componential one that derives its axes from something else.

The dimensional lineage compresses affective meaning onto two or three axes. The root is the semantic differential of Osgood et al. (1957), who had people rate words on bipolar adjective scales and recovered three recurring factors, evaluation, potency and activity. Russell & Mehrabian (1977) carried the same three into affect as pleasure, arousal and dominance, and Mehrabian (1996) formalised them as the PAD framework, named after the initials of those three words. Russell (1980) proposed the reading that made the tradition portable: emotion terms lie on a circle in the plane of valence and arousal, so what distinguishes two terms is an angle.

Lexical norms turned that geometry into a table of coordinates, one row per word and one column per axis. Warriner et al. (2013) collected nine-point ratings of valence, arousal and dominance, the three axes usually abbreviated VAD, for close to fourteen thousand English lemmas. Mohammad (2018) collected the same three for nearly twenty thousand terms by best–worst scaling, in which

an annotator picks the extremes of a small set instead of placing one item on a scale. The second of those two is the VAD lexicon of the roster, read at its published $d = 3$. Valence and dominance can be recovered from co-occurrence statistics alone, as components of a principal component analysis of word vectors, while no single component there matches arousal (Hollis & Westbury, 2016). That is evidence that label names carry part of this structure, and not an account of why they carry it.

Two later results say what the dimensional side settles and what it leaves open. Yik et al. (2011) fit a 12-point circumplex across instruments and samples, and the 12 points are angular resolution rather than extra dimensions, so the result reinforces the two-dimensional reading instead of qualifying it. Its object is momentary core affect, the raw pleasure and activation of a felt state, and not the meaning of an emotion word, which is what the lexical norms and the instruments below are about. Kuppens et al. (2013) review the relation between valence and arousal in subjective experience and find no single shape: across six proposals in the literature it appears as a straight line, as a V, and, when valence is read as a function of arousal rather than the reverse, as an inverted V. On average the relation is a weak asymmetric V, and the variation across individuals is large enough that no static shape survives. The axes are therefore not as fixed as a table of per-word coordinates makes them look.

The categorical lineage counts kinds instead of axes. A second family holds that emotions are discrete kinds and that a continuous space is the wrong object. Ekman (1992) argued for a small set of basic emotions marked by distinct expressions and physiology, and Plutchik (1980) proposed a psychoevolutionary scheme of primaries that combine into blends. The modern form of the position is empirical and much larger. Cowen & Keltner (2017) had people report their reactions to thousands of short videos and recovered 27 categories, with the further finding that the categories are bridged by continuous gradients rather than separated by boundaries. That study is the categorical row of Table 3. Cowen & Keltner (2020) carry the same programme to facial expression and report 28 kinds of expression.

Against both families, Barrett (2017) argues that emotion categories are constructed by the perceiver out of core affect and concepts, so no natural set is there to be counted. van Heijst et al. (2023) review the basic-versus-constructed dispute from an evolutionary angle and leave it open, which is the state the present paper inherits.

The componential programme derives its axes instead of asking for them. The third family treats an emotion as an episode with components: an appraisal of the event, a bodily reaction, an expression, an action tendency and a subjective feeling (Scherer, 2005). That paper also supplies the two instruments the tradition uses for elicitation, a lexicon of affect terms and a wheel on which respondents place their state. GRID is the name of the componential study that Scherer, Fontaine and colleagues ran to carry the idea over to the meaning of emotion words, with judges in several languages. Each term was rated on 142 componential features, and the dimensions were extracted as factors from those ratings rather than asked for directly: valence, potency, arousal and novelty. Fontaine et al. (2007) report the result under the title that the world of emotions is not two-dimensional. Fontaine et al. (2022) publish the coordinates for 80 terms across languages, and those are the appraisal geometry of the roster at its published $d = 4$.

The fourth dimension is what a comparison with the VAD lexicon would turn on. GRID potency corresponds closely to VAD dominance, so three of the four axes have counterparts in the older tradition and novelty has none. A test between three and four dimensions is therefore a test about novelty, and this paper does not run it. The evidence inside the tradition is itself mixed. Fontaine et al. (2022) report that two dimensions fit their data well in general, and that four are needed for the whole gamut of emotions. Terms of anger and surprise drive that requirement, while terms of joy and sadness are about as well represented by two.

What an instrument fixes before it sees data is a gradient, not a switch. It is tempting to say that every instrument here is confirmatory, in the sense of committing to its dimensions before collecting a rating, and that is false for the GRID. The instruments sit at three levels of prior commitment, and the rows of Table 3 are ordered by them. At the first level a VAD questionnaire fixes the axes themselves, asking for valence, arousal and dominance and able to return nothing else; the VAD lexicon and the Warriner norms are both of this kind. At the second the GRID fixes an inventory of features and derives the axes from their covariance. Its four-dimensional solution was therefore a

finding, given the inventory, and not an assumption, and the categorical account commits the same way, to an inventory of categories. At the third the label-name geometry of Section 2 fixes neither, and commits instead to a vocabulary and to the corpora the encoders were trained on, which is less of one kind of commitment rather than none of any kind.

One criticism reaches all of them at once. Fiske (2020) argues that vernacular emotion words lack meanings stable and mutually transparent enough to serve as units of psychological measurement, so any structure derived from them inherits that instability. The argument applies to the feature ratings of the appraisal instrument, to the per-word coordinates of the VAD lexicon, and to the emotion vocabulary this paper cuts its label points from. Nothing in the measure of Section 2 answers it. What the measure does is hold that vocabulary fixed across the geometries compared, which makes the defect common to all of them. Whether it also moves their order is not tested here, because no experiment in this paper varies the vocabulary.

instrument	**elicitation**	**fixed before data**	**nominal dimensions**	**reference**
VAD lexicon	best–worst choices on each axis	the axes	3, *asked*	Mohammad, 2018
Warriner norms	nine-point scales on each axis	the axes	3, *asked*	Warriner et al., 2013
appraisal geometry	ratings on 142 componential features	the feature inventory	4, *derived*	Fontaine et al., 2013
categorical account	34 categories and 14 scales on video	the category inventory	27, *derived*	Cowen & Keltner, 2017
label-name geometry	none (text corpora)	neither	none fixed	this paper

Table 3: **The gradient of prior commitment.** What each instrument fixes before it collects a rating, and how many dimensions it reports. The first two rows sit at the first level of commitment, fixing the axes and asking for them directly. The next two sit at the second, fixing an inventory of items and deriving their structure from it, so their dimension counts are outputs. The last sits at the third, fixing neither and committing to a vocabulary and a set of corpora instead, with its dimension a free parameter. The categorical row is the 2017 self-report study, not the later one on facial expression. The first, third and last rows are roster geometries of Section 4; the Warriner norms are listed as the other widely used lexicon of the same three axes.

This paper takes two points from the debate, and settles none of it. Three inheritances come from the history above. The vocabulary from which the label points are cut is an intersection: the terms carried at once by the appraisal instrument, by the VAD lexicon and by every encoder of the committee. It is therefore a vocabulary of the psychometric traditions, and not a list of our own. The two published geometries enter the roster at their published dimensions, as instances of a human-authored assignment of labels to points. And the classes of the adjudicating task are the labels of a corpus, restricted to those that are members of that vocabulary, which Appendix C.1 states term by term.

What the measure returns about them is narrower than the debate. It scores a label geometry in bits, the held-out information gain of a decoder that reads the geometry, counted above the reference level that a random geometry of the same dimension reaches. It then splits that score into what the arrangement buys, where the points lie relative to one another, and what the assignment buys, which label sits on which point. Both are read in one cell: one task, one frozen encoder, one decoder and one d. A geometry that collects fewer bits in a cell is worse in that cell, under that decoder, at that dimension, and it is not thereby a worse account of emotional meaning.

No contrast in this paper adjudicates the dimension count of the psychometric question. The two published geometries are read at $d = 4$ and $d = 3$, below the dimension $n - 1$ at which the one-hot code over n classes is itself admissible. What they give up in arrangement is therefore a gap to a ceiling that no geometry of their dimension reaches, and not the price of a trade. Both collect a positive assignment gain in the three neural encoders; in the non-neural control the appraisal geometry does not. Section 4 reads both against the resolution criterion of Section 2, and Appendix C gives the numbers cell by cell.

B PROOFS, STEP BY STEP

This appendix proves what the body states, in the order the body states it. Two results are written out step by step, because everything else rests on them: the fibre mean of Section 3, and the ceiling of Theorem 1. Each result opens with where the body uses it and what the body reads from it, and then comes as a named statement, its hypotheses and a proof. Every check, and every producer of a table or a figure in this paper, is in the code released with it.

B.1 NOTATION, AND THE TWO CONSTRAINTS THAT CARRY EVERY PROOF

Everything below lives in one slice of the symmetric matrices. A label geometry is $Z \in \mathbb{R}^{n\times d}$, one row per label, and the decoder reads it only through $K = ZZ^\top$. The admissible set equation 1 imposes two affine constraints on K, centring and scale, $K\mathbf{1} = 0$ and $\operatorname{tr} K = nd$, plus the rank constraint $\operatorname{rank} K \le d$. Dropping the rank constraint gives the *relaxed set* $\overline{\mathcal{K}}_d = \{K \succeq 0,\ K\mathbf{1} = 0,\ \operatorname{tr} K = nd\}$, which is convex. Both constraints are affine, so means and convex combinations of admissible matrices stay in the relaxed set, and that is the only closure any proof below needs.

Relabelling is an isometry of the slice, and its orbit is the fibre. An assignment $\pi \in S_n$ sends K to $K_\pi := P_\pi K P_\pi^\top$, with $(P_\pi K P_\pi^\top)_{ij} = K_{\pi^{-1}(i)\pi^{-1}(j)}$. The entries are permuted and none changes value, so the map preserves the Frobenius distance and both constraints. The orbit $\{K_\pi : \pi \in S_n\}$ is the *fibre* of the arrangement, and M_1 is the uniform mean of IG on it.

The matrix C is the Gram of the centred regular simplex. Write $C = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$. Its rows are the centred vertices $e_i - \frac{1}{n}\mathbf{1}$ of the standard basis, and $\|e_i - e_j\|^2 = 2$ for every $i \ne j$: all pairs are at one distance, which is what *regular simplex* means. Then $K_s = \frac{nd}{n-1}C$ is the admissible representative of that configuration, the one-hot code centred and scaled, with constant diagonal, constant off-diagonal and rank $n-1$.

The accounting identity is exact, and its residual is arithmetic. The residual of equation 4 measures only the arithmetic of the estimator: it sits at floating-point precision over the comparison grid, and exactly at zero on the topic task.

The precedence between the two parts is a choice, and it has a name. The value of a game under the precedence constraint $\mathsf{arr} \prec \mathsf{asg}$ (Faigle & Kern, 1992) is exactly the sequence of increments of equation 4, whereas the symmetric Shapley value (Shapley, 1953) would halve the interaction between the two parts. Chain decompositions of proper scores with martingale increments are due to Charpentier & Fernandes Machado (2026) and, in conditional form, Allen et al. (2023); their conditioning variable is never the geometry of the labels.

B.2 THE FIBRE MEAN IS THE SIMPLEX, AND TWO ROUTES PROVE IT

The body reads this proposition as the zero of the assignment. Section 3 states it, and Section 1 and Section 5 carry the reading: "the simplex is the barycentre of the admissible set, the non-informative prior over assignment"; it is what the ceiling rests on.

Proposition 2 (the fibre mean)**.** *Let K be symmetric with $\mathbf{1}^\top K\mathbf{1} = 0$, and let π be uniform on S_n. Then*

$$\mathbb{E}_\pi\left[P_\pi K P_\pi^\top\right] \;=\; \frac{\operatorname{tr} K}{n-1}\,C.$$

On the admissible set $\operatorname{tr} K = nd$, so the mean is K_s for every geometry, whatever its arrangement, rank or spectrum.

The hypothesis is weaker than it looks. The proof uses only that π is uniform on a group whose action is transitive on indices and on ordered distinct pairs, which S_n is, and that the entries of K sum to zero, which $K\mathbf{1} = 0$ gives. Nothing about the task, the encoder or the decoder enters.

Route 1: entry by entry. Fix i. Under a uniform π the index $\pi^{-1}(i)$ is uniform on $\{1,\dots,n\}$, so every diagonal entry of the mean is the average of the n diagonal entries of K,

$$\mathbb{E}_\pi\big[(K_\pi)_{ii}\big] \;=\; \frac{1}{n}\sum_a K_{aa} \;=\; \frac{\operatorname{tr} K}{n}.$$

Fix now $i \neq j$. The ordered pair $(\pi^{-1}(i), \pi^{-1}(j))$ is uniform on the $n(n-1)$ ordered distinct pairs, so every off-diagonal entry of the mean is the average of the $n(n-1)$ off-diagonal entries of K,

$$\mathbb{E}_\pi\big[(K_\pi)_{ij}\big] \;=\; \frac{1}{n(n-1)}\sum_{a\neq b} K_{ab} \;=\; \frac{\mathbf{1}^\top K\mathbf{1} - \operatorname{tr} K}{n(n-1)} \;=\; -\frac{\operatorname{tr} K}{n(n-1)},$$

the last step by the hypothesis. The mean therefore has constant diagonal and constant off-diagonal, and regrouping the two values row by row,

$$\mathbb{E}_\pi[K_\pi] \;=\; \frac{\operatorname{tr} K}{n} I - \frac{\operatorname{tr} K}{n(n-1)}\big(\mathbf{1}\mathbf{1}^\top - I\big) \;=\; \operatorname{tr} K\Big[\frac{1}{n-1} I - \frac{1}{n(n-1)}\mathbf{1}\mathbf{1}^\top\Big] \;=\; \frac{\operatorname{tr} K}{n-1}\, C.$$

Substituting $\operatorname{tr} K = nd$ gives K_s. □

The second route proves more than the value: it proves that no other value was available, and it needs one lemma.

Uniqueness is what the body asks of the lemma. Section 3 uses it for the fibre mean and for the meeting with the collapse literature, and Section 5 repeats the second: "uniqueness and symmetry alone, with no reference to the loss, force the frame"; it is what makes the ceiling a statement about one point.

Lemma 3 (the fixed point is unique)**.** *Let $\mathcal{K}_t = \{K \in \operatorname{Sym}(n) : K\mathbf{1} = 0,\ \operatorname{tr} K = t\}$. The fixed-point subspace of the action $K \mapsto P_\pi K P_\pi^\top$ on $\operatorname{Sym}(n)$ is the commutant $\operatorname{span}\{I, \mathbf{1}\mathbf{1}^\top\}$, of dimension two. The two affine constraints cut it to the single point $K = \frac{t}{n-1}C$. Hence any S_n-equivariant construction that returns an invariant element of $\mathcal{K}_t$ returns that point.*

Proof. A symmetric matrix commuting with every P_π is an intertwiner of the permutation representation. Over $\mathbb{R}$ that representation decomposes as $\mathbb{R}^n \cong \mathrm{triv} \oplus \mathrm{std}$, with each irreducible appearing once, so by Schur the commutant has dimension two and is $\{aI + b\mathbf{1}\mathbf{1}^\top\}$. Impose the constraints. Centring gives $K\mathbf{1} = (a+bn)\mathbf{1} = 0$, hence $b = -a/n$ and $K = aC$, which spends one degree of freedom. The trace gives $\operatorname{tr} K = a(n-1) = t$, which spends the other and fixes $a = t/(n-1)$. □

Route 2: the mean is invariant, and there is only one invariant. Write $M(K) = \frac{1}{n!}\sum_\pi P_\pi K P_\pi^\top$. First, $M(K)$ is invariant: conjugating by any σ sends the summand indexed by π to the one indexed by $\sigma\pi$, and $\sigma\pi$ runs over the whole group as π does, so the sum is only reindexed. Second, $M(K)$ lies in $\mathcal{K}_{\operatorname{tr} K}$: both constraints are affine and each summand satisfies them, so their average does. Invariant and in $\mathcal{K}_{\operatorname{tr} K}$, it is the single point of Lemma 3. On the admissible set that point is K_s. □

Route 2 is the mean under a finite group read as the orthogonal projection on the fixed-point subspace, the Reynolds operator (Shah & Chandrasekaran, 2012, Lemma 2.2). The commutant has dimension two at $n = 5$, 6, 7, 8 and 10, and the mean agrees with the exhaustive fibre to floating-point precision. The one-hot target placed in the admissible set *is* K_s, with equal radii and equiangular off-diagonal.

The terminal geometry of unconstrained training is the same fixed point. Section 3 and Section 5 read the meeting with neural collapse through it. Four hypotheses are named: balanced classes; class means centred and scaled to a fixed trace; $d \geq n-1$; and a unique optimal class-mean Gram matrix. Under them a loss that sees the labels only as a partition is invariant under relabelling, so its optimum is fixed by every permutation, and Lemma 3 leaves the admissible set one candidate. The frame and that uniqueness are theirs (Papyan et al., 2020; Zhu et al., 2021; Han et al., 2022); what the lemma adds is that symmetry and uniqueness alone force it. Read through this appendix, the same geometry appears as the fixed point of the label-permutation null.

Three collinear points average to the equilateral triangle. The smallest case that is not already symmetric makes the mechanism visible. Take $n = 3$ and $d = 1$, with the three labels on a line at $z = (1, 0, -1)^\top$, so that

$$K = zz^\top = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix}, \qquad K\mathbf{1} = 0, \qquad \operatorname{tr} K = 2, \qquad \operatorname{rank} K = 1.$$

The fibre has the six permutations of S_n. On the diagonal each of the three entries $(1, 0, 1)$ visits each position twice, so the mean diagonal entry is $\frac{2}{3}$, which is $\operatorname{tr} K/n$. Off the diagonal the three distinct pairs carry the values 0, -1 and 0, and each of them lands on a given position twice among the six permutations, so the mean is $-\frac{1}{3}$, which is $-\operatorname{tr} K/[n(n-1)]$. The mean is therefore

$$\mathbb{E}_\pi[K_\pi] = \tfrac{2}{3} I - \tfrac{1}{3}\big(\mathbf{1}\mathbf{1}^\top - I\big) = C = \frac{\operatorname{tr} K}{n-1}\, C,$$

since $\operatorname{tr} K/(n-1) = 1$ here. Read the ranks. The arrangement has rank one and the mean has rank two: averaging over assignments erases the arrangement, and erases it onto the equilateral triangle, which no member of the fibre resembles. That is why the ceiling of the next subsection holds for every geometry at once.

Without the centring constraint the mean has one more term. The hypothesis is not decoration. For a general symmetric K the same two averages give

$$\mathbb{E}_\pi[K_\pi] \;=\; \frac{\operatorname{tr} K}{n} I \;+\; s\,\big(\mathbf{1}\mathbf{1}^\top - I\big), \qquad s = \frac{\mathbf{1}^\top K \mathbf{1} - \operatorname{tr} K}{n(n-1)},$$

which collapses to a multiple of C only when $\mathbf{1}^\top K \mathbf{1} = 0$. On an uncentred configuration the formula of Proposition 2 is simply false, and the general form is what closes. We confirm the general form against an exhaustive mean taken with $K\mathbf{1} \neq 0$. Restricting a vocabulary and then measuring without re-applying the centring is the practical way to lose the hypothesis, which is why the admissible representative is recomputed on whatever set is measured.

The same computation holds for the reference measure. Let $\tilde{Z} \sim \mu_0$ be exchangeable, that is, let $P\tilde{Z}$ have the law of $\tilde{Z}$ for every permutation matrix P. Then $\mathbb{E}[\tilde{K}] = \mathbb{E}[P\tilde{K}P^\top]$ for every P, so averaging that identity over the group and applying Proposition 2 inside gives

$$\mathbb{E}_{\mu_0}[\tilde{K}] \;=\; \mathbb{E}_{\mu_0}\mathbb{E}_\pi\big[P_\pi \tilde{K} P_\pi^\top\big] \;=\; \mathbb{E}_{\mu_0}\Big[\tfrac{\operatorname{tr}\tilde{K}}{n-1}\Big]\, C \;=\; K_s,$$

because the trace is nd almost surely on the admissible set. The two averaging operations of the filtration, over the reference and over the fibre, therefore share one barycentre, which is why there is one ceiling below and not two.

Four constructions reach C, and two of them are one fact. Lemma 3 turns a list of coincidences into corollaries of one statement. The fibre mean reaches C; the reference mean reaches it; the centred and rescaled one-hot target reaches it, because the identity matrix is equivariant and the admissible set places it in $\mathcal{K}_{nd}$; and the Hessian of the log-sum-exp at the uniform weight is C/n, the only weight the group does not see. The last two are one identity and not two, since the Hessian of the log-sum-exp at p *is* the covariance of the categorical with parameter p. Three paths are separate, and they arrive at the same place for the same reason.

The proposition is about the Gram matrix, and two readings of it are false. First, the mean of the Gram is not the mean of the information gain. Proposition 2 computes $\mathbb{E}_\pi[K_\pi]$, whereas M_1 is $\mathbb{E}_\pi[\mathrm{IG}(K_\pi)]$, and the two would agree only if IG were affine in K. It is concave, and the gap between them is the subject of the next subsection. The level of the standard design is $\mathrm{IG}(K_s)$; $M_1(K)$ is the level of *that* arrangement without its assignment, which is never larger and is equal only at the simplex (Theorem 1). Second, the statement is true in Gram and not in coordinates. K_s has rank $n-1$, so no configuration of n points in $\mathbb{R}^d$ with $d < n-1$ has it as its Gram matrix: below that dimension the barycentre of the fibre is a matrix the admissible set does not contain.

The precedence is on the group mean, and what is ours is the cut. Reading the mean over an orbit as the orthogonal projection on the fixed-point subspace is Shah & Chandrasekaran (2012, Lemma 2.2), stated there for an arbitrary finite subgroup. The basis of the linear maps that commute with every permutation is due to Maron et al. (2019). What this paper adds is the last step of Lemma 3. The two affine constraints of the admissible set cut that subspace to a single point, so the mean has no freedom left, and the point is the centred one-hot code.

B.3 At fixed temperature the simplex is the arrangement ceiling, and assigns nothing

Concavity is what the fixed-temperature convention is for. Section 2 declares that convention, "one τ across the geometries compared, the convention of every theorem", and Section 3 spends it in Theorem 1. It gives the ceiling its Jensen step and the screen the sign of β.

Proposition 4 (concavity)**.** *At fixed τ and fixed ridge operator G, the map $K \mapsto \mathrm{IG}(K)$ is concave on all of* $\mathrm{Sym}(n)$.

Proof. By equation 2 each squared distance $D_{t\ell} = v_{t\ell}^\top K v_{t\ell}$ is linear in K, so $x_\ell = -D_{t\ell}/\tau$ is affine in K. The log-sum-exp $\mathrm{lse}(x) = \ln \sum_\ell e^{x_\ell}$ is convex: its Hessian is $\mathrm{diag}(p) - pp^\top$ with p the softmax of x, and for any u,

$$u^\top \big(\mathrm{diag}(p) - pp^\top\big) u \;=\; \sum_\ell p_\ell u_\ell^2 - \Big(\sum_\ell p_\ell u_\ell\Big)^2 \;=\; \mathrm{Var}_p(u) \;\geq\; 0.$$

A convex function composed with an affine map is convex, its negative is concave, and a sum with positive weights preserves concavity. By equation 3 that sum is IG. Neither $K \succeq 0$ nor the rank was used, which is what lets the argument run on the relaxed set. □

Concavity is in K and at fixed τ, and both qualifications are measured. It is not concavity in Z: the map $Z \mapsto ZZ^\top$ is quadratic, and the composition fails. It is also not concavity under a temperature refit per geometry. Concavity in K holds along every chord between admissible geometries we test; with τ refit along the chord it can fail. That is why every theorem here is stated at fixed τ, and why a prediction it makes about numbers measured at canonical τ is checked rather than asserted.

Strictness is what makes the equality case an if and only if. Section 3 carries the rank hypothesis into Theorem 1, where it "holds in every encoder" we measure. It gives the ceiling its equality case and the screen $\beta > 0$.

Lemma 5 (strictness)**.** *Let $W = \{\Theta \in \mathrm{Sym}(n) : \Theta\mathbf{1} = 0,\ \mathrm{tr}\,\Theta = 0\}$ be the space of admissible directions at K_s. If the held-out score matrix G has full column rank n, then $K \mapsto \mathrm{IG}(K)$ is strictly concave along every $\Theta \in W \setminus \{0\}$, at every K.*

Proof. Strictness is a statement about directions and not about K. For a text t, $-\ln p_t(y_t) = D_{t,y_t}/\tau + \mathrm{lse}(-D_t/\tau)$; the first term is linear in K, and the log-sum-exp is strictly convex in every direction except along $\mathbf{1}$. Since $D_{t\ell}$ is affine in K with direction $\xi_{t\ell}(\Theta) = v_{t\ell}^\top \Theta v_{t\ell}$, the map $K \mapsto \mathrm{IG}_t(K)$ fails to be strictly concave along $\Theta \in W$ if and only if $\xi_{t\ell}(\Theta)$ is constant in ℓ, that is $L_t(\Theta) = -2\Theta g_t + \mathrm{diag}\,\Theta \in \mathrm{span}\{\mathbf{1}\}$; and $\mathbf{1}^\top L_t(\Theta) = 0$ on W, so $L_t(\Theta) = 0$. The softmax weights are strictly positive at every K, so this condition does not depend on K: IG is strictly concave along Θ on the whole relaxation if and only if $L_t(\Theta) \neq 0$ for some held-out text. Now $L_t(\Theta) = 0$ for every t reads $\Theta g_t = \frac{1}{2}\,\mathrm{diag}\,\Theta$ for every t, the same vector for all texts, hence $\Theta(g_t - g_s) = 0$ for all pairs. If the differences of the held-out scores span $\mathbf{1}^\perp$, then $\ker\Theta \supseteq \mathbf{1}^\perp$, and $\Theta\mathbf{1} = 0$ gives $\Theta = 0$. The ridge carries an intercept and is fit against a one-hot target, whose rows sum to one, so the rows of G sum to one and the differences span $\mathbf{1}^\perp$ exactly when G has full column rank n, which is the hypothesis. □

The hypothesis is measured rather than assumed. The rank of G is full in every encoder we measure, and so is the stronger statement that the linear map $\Theta \mapsto (L_t(\Theta))_t$ has trivial kernel on W, which is the conclusion without the intermediate step.

The ceiling is the statement the body sends here for proof. Section 3 states the theorem and Section 4 reads the trade it opens, drawing from it that "the simplex discriminates as much as the admissible set allows and knows nothing about what it discriminates".

Proof of Theorem 1. Hypotheses, named: (H1) τ fixed, the same for every geometry compared; (H2) G fixed and independent of Z, with full column rank n, the hypothesis Lemma 5 needs and the equality case of (f) uses; (H3) the trace constraint $\operatorname{tr} K = nd$; (H4) μ_0 exchangeable and supported on the admissible set; (H5) M_1 the uniform mean on the fibre.

(a) The arrangement gain is bounded above. For every K in the relaxed set $\overline{\mathcal{K}}_d$, Jensen on the group mean (Shah & Chandrasekaran, 2012, Lemma 4.1), which appears there with the opposite curvature, with Proposition 4 and then Proposition 2, give

$$M_1(K) \;=\; \mathbb{E}_\pi \operatorname{IG}(K_\pi) \;\le\; \operatorname{IG}\big(\mathbb{E}_\pi K_\pi\big) \;=\; \operatorname{IG}(K_s).$$

The step through Jensen is why the bound is over the relaxed set and not the admissible one: the fibre mean of a rank-d geometry has rank up to $n-1$, so it leaves the rank constraint, and Proposition 4 holds on the ambient space. Nothing here uses the rank.

(b) The reference level obeys the same bound. By (H4) and the exchangeable version of Proposition 2, $\mathbb{E}_{\mu_0}\tilde{K} = K_s$, so Jensen again gives $M_0 = \mathbb{E}_{\mu_0}\operatorname{IG}(\tilde{K}) \le \operatorname{IG}(K_s)$, and the right-hand side of equation 7 is non-negative.

(c) Subtract. Removing the same constant M_0 from both sides of (a) gives $\mathsf{arrangement\ gain}(K) \;=\; M_1(K) - M_0 \;\le\; \operatorname{IG}(K_s) - M_0$ for every K of the relaxed set, which is equation 7. There is no circularity: (a) involves M_1 and not M_0, and (b) involves M_0 and not Z.

(d) The fibre of the simplex is a point. Every permutation fixes I and fixes $\mathbf{1}\mathbf{1}^\top$, so $P_\pi C P_\pi^\top = C$ and $P_\pi K_s P_\pi^\top = K_s$. Hence $M_1(K_s) = \operatorname{IG}(K_s)$ and $\mathsf{assignment\ gain}(K_s) \equiv 0$, at every τ, refit or not, because a refit sees the same distances. With (b), $\mathsf{arrangement\ gain}(K_s) = \operatorname{IG}(K_s) - M_0 \ge 0$: the standard design never loses to a random constellation.

(e) Attainment. K_s has rank $n-1$, so it belongs to the admissible set if and only if $d \ge n-1$, and only then is the bound of (c) attained by a member. Below that dimension the bound still holds over the relaxation, and by (f) it is strict inside the admissible set.

(f) Equality forces the fibre to a point. Each $\operatorname{IG}(K_\pi)$ lies below the supporting hyperplane of IG at K_s, and the average of those bounds is $\operatorname{IG}(K_s)$ because $\mathbb{E}_\pi \Theta_\pi = 0$, so equality of the averages forces equality term by term. A concave function that meets its tangent at the far end of a segment is affine on the segment, so the Hessian at K_s vanishes along Θ_π, and Lemma 5 then forces $\Theta_\pi = 0$ for every π. Hence equality holds if and only if the fibre is the point K_s, that is if and only if K commutes with every permutation, and Lemma 3 leaves the admissible set one candidate. □

On real data the fibre amplitude of the standard design is zero to floating-point precision. On a toy small enough for the fibre to be enumerated, the fibre of the simplex closes exactly, and (a) and (b) hold against that exhaustive fibre and against draws of μ_0.

The budget identity is how the body prices a trade. Section 3 states it, and Section 1, Section 4 and Section 5 read the trade from it: "an informative geometry trades arrangement for the power to assign, and comes out ahead if and only if the exchange rate exceeds one". Its slack is read as a price rather than discarded.

Identity 6 (budget)**.** *Under the hypotheses of Theorem 1, with* $\operatorname{IG}(K_s)$ *and* M_1 *read under one and the same* τ*, the loss from trading the standard design for* K *is the Jensen gap of the fibre,*

$$\textit{arrangement rent}(K) \;=\; \operatorname{IG}(K_s) - M_1(K) \;=\; \operatorname{IG}\big(\mathbb{E}_\pi K_\pi\big) - \mathbb{E}_\pi\big[\operatorname{IG}(K_\pi)\big],$$

and $\textit{arrangement gain}(K) + \textit{arrangement rent}(K) = \operatorname{IG}(K_s) - M_0$*, which is equation 8.*

Proof. By Theorem 1(d) the standard design has $\mathsf{arrangement\ gain}(K_s) = \operatorname{IG}(K_s) - M_0$, and by definition $\mathsf{arrangement\ gain}(K) = M_1(K) - M_0$. Subtracting cancels M_0, which is where the matched-dimension hypothesis enters, and leaves $\operatorname{IG}(K_s) - M_1(K)$. Substituting $K_s = \mathbb{E}_\pi K_\pi$

from Proposition 2 and $M_1(K) = \mathbb{E}_\pi \mathrm{IG}(K_\pi)$ from the definition turns that difference into the Jensen gap. Adding arrangement gain(K) back gives the second display. □

The cancellation is arithmetic and holds under any temperature convention; the identification with the Jensen gap needs both terms under one functional, hence fixed τ. So the positivity of the rent is a theorem at fixed τ and a measured prediction at canonical τ. The identity holds to floating-point precision in every dimension-matched cell, and the rent is positive at canonical τ in every cell of the grid.

B.4 The expansion around the simplex kills the first order

The quadratic form of the cell is what survives the fibre mean. Section 3 walks this path and reads from it that screening a candidate geometry is a matrix product, "with no training and no permutation". By Identity 6 the whole dependence of the arrangement gain on the geometry sits in the rent, so what follows expands the rent around the simplex. Write $K = K_s + \Theta$ and $\xi_{t\ell} := v_{t\ell}^\top \Theta v_{t\ell}$. With $\mathrm{IG}_t = \log_2 n - [\mathrm{lse}(x) - x_{y_t}]/\ln 2$, $x_\ell = -D_{t\ell}/\tau$, and $\mathrm{lse}(x+\delta) = \mathrm{lse}(x) + \mathbb{E}_p[\delta] + \frac{1}{2}\mathrm{Var}_p(\delta) + O(\delta^3)$ for p the softmax of x,

$$\mathrm{IG}_t(K_s + \Theta) = \mathrm{IG}_t(K_s) + \frac{1}{\tau \ln 2}\big(\mathbb{E}_{p_t}[\xi_t] - \xi_{t,y_t}\big) - \frac{1}{2\tau^2 \ln 2}\mathrm{Var}_{p_t}(\xi_t) + O(\|\Theta\|^3), \tag{10}$$

with $p_t = \mathrm{softmax}(-D_t(K_s)/\tau)$; in the simplex $D_{t\ell}(K_s) = \frac{nd}{n-1}(\|g_t\|^2 - 2g_{t\ell} + 1)$, so p_t is the softmax of the ridge scores. Averaging over the fibre kills both linear terms, because $\mathbb{E}_\pi \Theta_\pi = 0$ by Proposition 2 while p_t, evaluated at K_s, does not move with π; hence $M_1(K) = \mathrm{IG}(K_s) - \bar{q}(\Theta) + O(\|\Theta\|^3)$, with $\bar{q}$ the fibre mean of the quadratic form q defined next. Write $w_t = 1/(nT_i')$ for a held-out text of class i, with T_i' such texts, the weights that make equation 3 a macro average. The second-order term summed over texts with those weights is $q(\Theta) := \frac{1}{2\tau^2 \ln 2}\sum_t w_t \mathrm{Var}_{p_t}(\xi_t(\Theta))$. Expanding $v_{t\ell} = g_t - e_\ell$ inside ξ gives $\xi_{t\ell} = g_t^\top \Theta g_t - 2(\Theta g_t)_\ell + \Theta_{\ell\ell}$, whose first term is constant in ℓ and drops from every mean and variance under p_t, so the text enters q through the linear map $L_t(\Theta) := -2\Theta g_t + \mathrm{diag}\,\Theta$ alone, and $q(\Theta) = \kappa \sum_t w_t L_t(\Theta)^\top \Omega_t L_t(\Theta)$ with $\kappa = 1/(2\tau^2 \ln 2)$ and $\Omega_t = \mathrm{diag}(p_t) - p_t p_t^\top$ the Hessian of the log-sum-exp at the simplex. In Hessian notation $q(\Theta) = \frac{1}{2}\Theta^\top H \Theta$ with $H = -\nabla^2 \mathrm{IG}|_{K_s} \succeq 0$ by Proposition 4. The identity that made the simplex the zero is therefore also what removes the first order, and the rest of this appendix is about $\bar{q}$.

B.5 The space of directions has two blocks, and Schur leaves two invariants

The admissible set is affine at the simplex, so the directions form a linear space. The constraints $K\mathbf{1} = 0$ and $\mathrm{tr}\,K = nd$ are affine and K_s satisfies them, so $K_s + \varepsilon\Theta$ stays in the relaxed set for small ε if and only if $\Theta \in W = \{\Theta \in \mathrm{Sym}(n) : \Theta\mathbf{1} = 0, \mathrm{tr}\,\Theta = 0\}$, of dimension $n(n+1)/2 - n - 1 = (n-2)(n+1)/2$, which is 135 at $n = 17$. The group acts on W by conjugation and fixes K_s. As an S_n-module, $W \cong \mathrm{std} \oplus S^{(n-2,2)}$: the block std is the image of the maps $v \mapsto C\,\mathrm{diag}(v)C$ for $v \perp \mathbf{1}$, which move the diagonal, of dimension $n-1$, and par is its orthogonal complement in W, of dimension $n(n-3)/2$, on which every element has zero diagonal. The decomposition is the Young rule for the restriction of the permutation representation; the identity $\|\mathrm{diag}\,\Theta\|^2 = \frac{n-2}{n}\|\Theta_{\mathrm{std}}\|^2$ is the projection computed on that image, and the orthogonality $\mathrm{tr}\,\Theta^2 = \|\Theta_{\mathrm{std}}\|^2 + \|\Theta_{\mathrm{par}}\|^2$ is that of non-isomorphic blocks. The two blocks are non-isomorphic because their dimensions differ for every $n \geq 4$; at $n = 3$ the block par is empty, so Theorem 7 does not cover the three-label task of Appendix C.1. The dimension of the space of invariant forms on W is 2 at every cardinality where we compute it.

Two invariants, and the claim the body makes that there is no third. Section 1 and Section 3 state it as "to second order there is no third number", and Section 4 leans on it where "no measured margin" justifies the second term, the ground of the two-scalar screen.

Theorem 7 (two invariants, and what they are)**.** *For $n \geq 4$ the space of permutation-invariant quadratic forms on W has dimension exactly two, spanned by $u_2 = \mathrm{tr}\,\Theta^2 = \|K - K_s\|_F^2$ and $u_1 = \sum_i \Theta_{ii}^2 = \|\mathrm{diag}\,\Theta\|^2$.*

Hypotheses, named: the admissible set is affine around K_s, so the space of directions is W; the group S_n acts on W by conjugation; and $\bar{q}$ is invariant under it because it is a group mean. That invariance is the only property the theorem uses.

Proof. If $V = \bigoplus_\lambda m_\lambda V_\lambda$ as a real module of a finite group whose irreducibles are absolutely irreducible, as those of S_n are, the space of invariant symmetric bilinear forms on V has dimension $\sum_\lambda m_\lambda(m_\lambda+1)/2$. With $m_{\mathrm{std}} = m_{\mathrm{par}} = 1$ the dimension is $1+1=2$ (Figure 5), and Schur says more than the count: on each irreducible block the invariant form is a multiple of the norm, with no cross term between non-isomorphic blocks. This is the structural half of Lemma 9. A counting proof reads the same statement on the 15 partitions of an index quadruple, the basis of the maps that commute with every permutation (Maron et al., 2019): a quadratic form on symmetric matrices sees 7 of them, and the two constraints of W kill five more. □

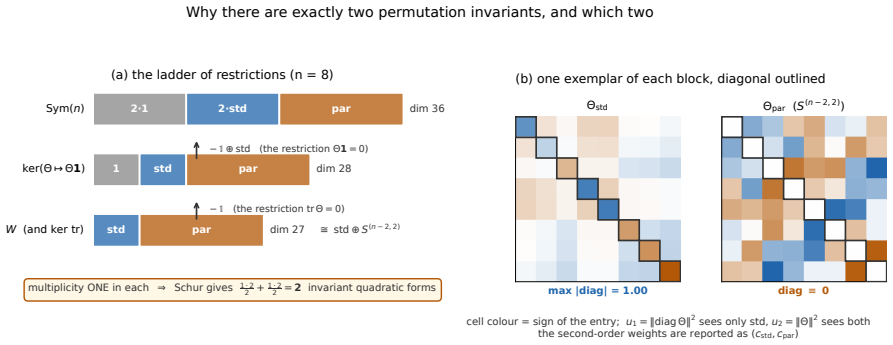


Figure 5: $W = \mathrm{std} \oplus \mathrm{par}$**, and why exactly two.** (a) The ladder of constraints $\mathrm{Sym}(n) \to \ker(\Theta \mapsto \Theta\mathbf{1}) \to W$ with the pieces the group does not mix and their dimensions, at $n = 8$: "why two" without representation theory, with the box "multiplicity one in each ⇒ Schur gives 2". (b) One exemplar of each block with the diagonal marked: in std the diagonal carries most of the norm, in par it is identically zero: "which two". Closed form in n.

The two invariants are a distance and the part of it carried by the radii. Section 3 reads the theorem as "a geometry can leave the simplex in two ways, and the formula charges each at its own weight". It names the two numbers the screen of Section 4 computes.

Theorem 8 (what the invariants are)**.** *$u_2 = \|K - K_s\|_F^2$ is the squared distance from the geometry to the simplex, $u_1 = \|\mathrm{diag}\,\Theta\|^2$ is the part of that distance carried by the radii, and the fraction $\psi = u_1/u_2$ lies in $[0, (n-2)/n]$.*

Proof. The two identifications are immediate from the definitions. For the range of ψ, the block identities of the previous paragraph give $\psi = u_1/u_2 = \frac{n-2}{n}\|\Theta_{\mathrm{std}}\|^2/(\|\Theta_{\mathrm{std}}\|^2 + \|\Theta_{\mathrm{par}}\|^2)$, which runs over $[0,(n-2)/n]$, the maximum at pure std and the zero at pure par. For the first identification read as a spectral excess, $K\mathbf{1} = 0$ gives $KC = K$ and $\mathrm{tr}(KK_s) = (nd)^2/(n-1) = \mathrm{tr}\,K_s^2$, so $\|K - K_s\|_F^2 = \mathrm{tr}\,K^2 - (nd)^2/(n-1)$, which is equation 11 below. □

The domain says where a level may be read, before any level is read. Section 3 and Section 4 use it as the scope of the expansion, which "holds only near the simplex, at a distance no published geometry reaches". It bounds what the screen may claim.

The distance to the simplex has a natural unit, the effective dimension read as a distance:

$$\varepsilon_{\mathrm{equiv}} := \frac{\|K - K_s\|_F}{\|K_s\|_F} = \sqrt{\frac{n-1}{d_{\mathrm{eff}}} - 1}, \qquad d_{\mathrm{eff}} = \frac{(nd)^2}{\mathrm{tr}\,K^2}, \tag{11}$$

and the floor $\varepsilon_{\mathrm{equiv}} \geq \sqrt{(n-1)/d - 1}$ follows from $d_{\mathrm{eff}} \leq \mathrm{rank}\,K \leq d$.

Read against equation 9, the gain falls with the distance to the simplex, and the part of that distance carried by the radii is charged at its own weight. Whether that weight is a surcharge is the sign of

α, which the theorem does not give: it is positive as $\tau \to \infty$ by Theorem 10, and positive at the operating temperature in every cell we measure (Section B.6).

The identity behind u_2 is that the spectral excess is the squared distance to the fixed-point subspace of Lemma 3, and it holds to floating-point precision. At $n = 17$ the ceiling on ψ is 0.882, and over 200 isotropic draws in W the largest fraction is 0.222. That second number reads the two block dimensions and no measured geometry: an isotropic direction splits its norm between the blocks in proportion to their dimensions, so the fraction it realises sits near $\frac{n-2}{n} \cdot \frac{n-1}{\dim W}$, and only an almost pure std direction approaches the ceiling. The radii can still decide an ordering while carrying little of the norm, because weight in the norm and power to decide are different things.

Two ends of the same construction place the fibre mean in a family rather than in isolation. With one block the partitioned mean returns Proposition 2; with n singleton blocks the group is trivial and nothing moves.

B.6 The two weights are two traces of one Hessian

The block traces are why the weights need no training run. Section 1 promises that "screening a candidate geometry for its arrangement gain is a matrix product". Section 3 adds that "α has the sign of the difference of the two block prices", and Section 4 takes c_G from the ridge with nothing fitted.

Hypotheses, named: S_n acts on W by orthogonal maps, which conjugation by a permutation matrix is; and std and par are submodules of W, which the previous subsection established.

Lemma 9 (block trace)**.** *Let q be a quadratic form on W and $\bar{q}$ its mean under S_n. Then $\bar{q}(\Theta) = c_{\mathrm{std}}\|\Theta_{\mathrm{std}}\|^2 + c_{\mathrm{par}}\|\Theta_{\mathrm{par}}\|^2$ with $c_B = \mathrm{tr}(q|_B)/\dim B$.*

Proof. $\bar{q}$ is invariant, so by Theorem 7 it is a multiple of the norm on each block, without cross term. The group acts on each block by orthogonal matrices, so the trace of a quadratic form restricted to a block is invariant under conjugation, $\mathrm{tr}(\bar{q}|_B) = \mathrm{tr}(q|_B)$, and the trace of $c_B\|\cdot\|^2$ on B is $c_B \dim B$. □

The block traces the lemma predicts agree with the exhaustive fibre to floating-point precision.

The traces close in the ridge scores, so the weights need no fibre. With q the second-order form of equation 10, $q(\Theta) = \kappa \sum_t w_t L_t(\Theta)^\top \Omega_t L_t(\Theta)$, $\mathrm{tr}(q|_B) = \kappa \sum_t w_t \,\mathrm{tr}(\Omega_t \Lambda_{t,B})$ with $\Lambda_{t,B} = L_t \Pi_B L_t^*$, where $L_t^* : \mathbb{R}^n \to W$ is the adjoint of L_t and Π_B is the orthogonal projector of W onto the block B; the traces close because Π_{std} is the projector onto the image of $v \mapsto C\,\mathrm{diag}(v)C$ and $\Pi_W = \Pi_{\mathrm{std}} + \Pi_{\mathrm{par}}$. With $\gamma_{t\ell} := p_{t\ell}(1 - p_{t\ell})$, the sensitivity of the softmax at the simplex, and $\mathbb{E}_t$ the macro mean over texts,

$$c_{\mathrm{std}} = \kappa \,\frac{n}{(n-1)(n-2)}\, \mathbb{E}_t\Big[\sum_\ell \gamma_{t\ell}(2g_{t\ell} - 1)^2 - \tfrac{4}{n}\,\mathrm{Var}_{p_t}(g_t)\Big], \tag{12}$$

$$c_{\mathrm{par}} = \frac{\mathrm{tr}(q|_W) - (n-1)c_{\mathrm{std}}}{n(n-3)/2}, \tag{13}$$

$$\mathrm{tr}(q|_W) = \kappa\, \mathbb{E}_t\Big[\sum_\ell \gamma_{t\ell}\big(1 + 2\|g_t\|^2 - 4g_{t\ell}\big) + \tfrac{2(n-3)}{n-1}\,\mathrm{Var}_{p_t}(g_t)\Big]. \tag{14}$$

In the coordinates of equation 9, since $u_1 = \frac{n-2}{n}\|\Theta_{\mathrm{std}}\|^2$ and $u_2 = \|\Theta_{\mathrm{std}}\|^2 + \|\Theta_{\mathrm{par}}\|^2$, the two coefficients are $\beta = 2c_{\mathrm{par}}$ and $\alpha = \frac{2n}{n-2}(c_{\mathrm{std}} - c_{\mathrm{par}})$. Equation equation 13 is a subtraction because equation 14 gives the trace on all of W and equation 12 gives the part of it that sits on std. The task constant c_G is the case $\gamma \equiv (n-1)/n^2$ of equation 12 and equation 14, the uniform weight. On a synthetic reference cell, $n = 6$, $d = 3$, $\tau = 1$, where the fibre is exhaustive, the two prices are 0.0171 and 0.0097 bits per unit of squared deviation from the simplex: a unit spent on the radii costs 1.766 times a unit spent at fixed radii, and it is that ratio, not either price, that the order reads. Lemma 5 fixes one of the two signs and not the other. The Hessian H at K_s is positive definite on W, so q is positive definite, its trace on each block is positive, and $c_{\mathrm{std}}, c_{\mathrm{par}} > 0$ by Lemma 9; hence $\beta = 2c_{\mathrm{par}} > 0$. The sign of α does not follow, because α is proportional to $c_{\mathrm{std}} - c_{\mathrm{par}}$ and

a difference of two positive traces has no sign in advance. It is measured. Table 4 prints the price ratio and the two weights in the four cells, at the canonical temperature and along a grid around it, and the ratio stays well above one everywhere, so $\alpha > 0$ and both prices are positive in every cell we measure.

encoder	τ/τ_{can}	τ	$c_{\text{std}}/c_{\text{par}}$	$\tau^2 \ln 2 \cdot \alpha$	$\tau^2 \ln 2 \cdot \beta$
mpnet	0.25	0.71	15.5	0.0032	0.0002
	0.5	1.42	15.1	0.0072	0.0005
	1	2.85	14.8	0.0176	0.0011
	2	5.70	14.1	0.0344	0.0023
	4	11.39	13.4	0.0482	0.0034
	16	45.58	13.3	0.0567	0.0041
GTE-large	0.25	0.69	13.8	0.0026	0.0002
	0.5	1.38	14.2	0.0060	0.0004
	1	2.76	13.8	0.0146	0.0010
	2	5.52	12.8	0.0311	0.0023
	4	11.03	12.0	0.0467	0.0037
	16	44.13	11.8	0.0565	0.0046
E5-large-v2	0.25	0.69	11.5	0.0018	0.0002
	0.5	1.38	11.5	0.0043	0.0004
	1	2.76	11.9	0.0126	0.0010
	2	5.52	11.6	0.0290	0.0024
	4	11.05	10.9	0.0458	0.0041
	16	44.18	10.7	0.0563	0.0051
TF-IDF-SVD	0.25	0.71	51.8	0.0054	0.0001
	0.5	1.42	45.8	0.0108	0.0002
	1	2.83	34.0	0.0207	0.0006
	2	5.66	22.3	0.0355	0.0015
	4	11.32	16.5	0.0498	0.0028
	16	45.29	13.1	0.0582	0.0042

Table 4: **The sign of α: the price ratio and the two weights in the four cells, at the canonical temperature and along a grid.** $\tau/\tau_{\text{can}} = 1$ is the temperature each cell elects; the other rows rescale it. The two weights are printed multiplied by $\tau^2 \ln 2$, the scale on which the expansion reads them, because the two prices themselves scale as $1/\tau^2$. $\beta = 2c_{\text{par}} > 0$ by Theorem 1; α has the sign of $c_{\text{std}} - c_{\text{par}}$, and the ratio never approaches one. The sign is unchanged up to 10^3 times the canonical temperature. The prices depend on the task and the encoder only, not on the label-name geometry.

The infinite-temperature limit is where c_G comes from. Section 3 and Section 5 read it as why "only the ratio of their weights fixes the order, and the expansion misses the level". Section 4 measures the order with that ratio.

Theorem 10 (the uniform-weight limit)**.** *As $\tau \to \infty$ the softmax goes to the uniform weight, the Hessian becomes C/n, and $\alpha\tau^2 \ln 2 \to 1/n$, $\beta\tau^2 \ln 2 \to c_G$ with $c_G = \frac{4}{n(n-1)}\mathbb{E}_t\|Cg_t\|^2$.*

Proof. The Hessian Ω_t is invariant under the group only at the uniform weight, where $\Omega_t = C/n$ for every t. There $\mathrm{Var}_{1/n}(\xi_t) = \frac{1}{n}\|-2\Theta g_t + \operatorname{diag}\Theta\|^2$, because Θg_t and $\operatorname{diag}\Theta$ already sum to zero, and we take the fibre mean term by term with $\Theta_\pi := P_\pi \Theta P_\pi^\top$: $\|\operatorname{diag}\Theta_\pi\|^2$ is invariant; the cross term dies since $\mathbb{E}_\pi[\Theta_\pi \operatorname{diag}\Theta_\pi] = \frac{1}{n}(\mathbf{1}^\top\Theta \operatorname{diag}\Theta)\mathbf{1} = 0$; and $\mathbb{E}_\pi[\Theta_\pi^2] = \frac{\operatorname{tr}\Theta^2}{n-1}C$ is Proposition 2 applied to Θ^2, giving $g_t^\top C g_t = \|Cg_t\|^2$. Hence $\bar{q}_{1/n}(\Theta) = \kappa[\frac{1}{n}\|\operatorname{diag}\Theta\|^2 + c_G\|\Theta\|_F^2]$ with $c_G = \frac{4}{n(n-1)}\mathbb{E}_t\|Cg_t\|^2$, the second-order identity of equation 5, obtained from the same Proposition 2 that killed the first order. As $\tau \to \infty$ the logits $-D_t(K_s)/\tau$ go to zero uniformly in t, because the simplex bounds the distances, $p_t \to 1/n$, and the variance is continuous in the weight. □

The closed forms reach the limit to a relative $1.8 \cdot 10^{-4}$ at $\tau = 10^4$ on the synthetic reference cell, and the vanishing of the first order holds there as well. Theorem 10 is therefore the evaluation of

$\beta/(\alpha n)$ at infinite temperature, and equation 5 is that evaluation written in the two scalars of the Gram matrix.

B.7 THE DOMAIN IS CHECKED BEFORE THE LEVEL, AND NO SPECTRAL FUNCTION SEES THE RADII

The floor on the distance puts the published geometries outside the domain before any measurement. Since $d_{\text{eff}} \leq d$, $\varepsilon_{\text{equiv}} \geq \sqrt{(n-1)/d-1}$ for any geometry of n labels in d dimensions: at $n = 17$ and $d = 4$ the floor is $\sqrt{3}$, and the five geometries of the roster run from 1.87 to 2.74, one value each and the same in the four encoders because $\varepsilon_{\text{equiv}}$ reads the geometry alone (Figure 6c), structurally outside the domain of any second-order expansion around the simplex. There the evidence for the order is empirical, Table 2, and no level in bits follows from a formula; the rule is to compute d_{eff} before trusting a level. The identity equation 11 and the floor below it are arithmetic, and the relative error of the predicted level grows with the distance to the simplex.

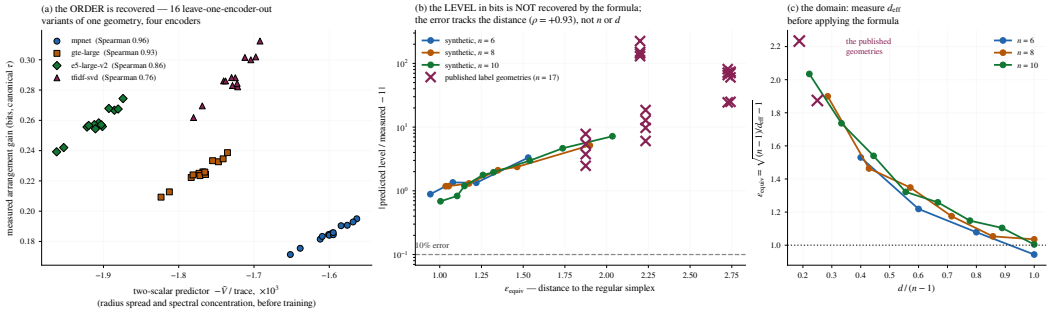


Figure 6: **The order of Figure 4c, with the level and the domain.** (a) as in Figure 4c. (b) The relative error of the predicted level against the distance to the simplex, on a synthetic reference grid with $n = 6, 8, 10$, where the fibre is exhaustive and the error is seen to grow with the distance, and on the published geometries ($n = 17$, crosses), all far above the 10% line. (c) $\varepsilon_{\text{equiv}}$ of equation 11 against $d/(n-1)$ on the synthetic cells, with the published geometries at $d = 3$ and 4: the floor $\sqrt{(n-1)/d-1}$ puts them outside the domain before any measurement.

Every function of the spectrum alone is constant on the block that carries the radii. Section 3 announces that "on a family built at one spectrum the log-determinant has no order to give", and Section 4 measures the arrangement gain on such a family. That the gain is not a function of the spectrum alone is read through the orientation null, which redraws the frame at fixed spectrum (Koussis et al., 2025), and the std block is what sees it. The effective rank (Roy & Vetterli, 2007), RankMe (Garrido et al., 2023) and the Vendi score (Friedman & Dieng, 2023) belong to the same spectral family as the log-determinant (Pasarkar & Dieng, 2024): all are functions of the eigenvalues of K alone, so none sees u_1 of Theorem 8, the part of the distance carried by the radii, and in the regime where the radii carry the ordering that is where they fail. Along the constructions that move the spectrum by design the log-determinant orders through a different channel: at fixed trace and $d = 2$ it orders as $-\operatorname{tr} K^2$ by Schur concavity, which we confirm numerically.

C THE MEASUREMENT IN NUMBERS: DATA, TRAINING, AND EVERY TABLE

C.1 DATA AND TRAINING, IN NUMBERS

This appendix carries the arithmetic behind "four tasks, four encoders": the sizes, the filters, the elected constants and the draw budgets. Each is tabulated from the same measurement path that produces the bits, and none is transcribed by hand. Every number in the paper carries one of three statuses: *identity*, *numerically checked theorem*, or *exploratory*. The last comes with Monte Carlo errors and no interval, and it is the status of every level in bits. The two published geometries enter as instances of a human-authored assignment, never as the thesis. Every check, and every producer of a table or a figure in this paper, is in the code released with it.

Two filters cut the adjudicating task, and the label vocabulary is the second. The adjudicating task is GoEmotions in its `simplified` configuration, whose published training split holds 43,410 texts. Two filters run in that order: the first is the corpus's own label count, the second is the vocabulary from which the label points are cut. We keep the texts carrying *exactly one* label, which leaves 36,308. Among those we keep the ones whose label belongs to the 77-term emotion vocabulary, which leaves 11,621 texts over 17 classes. That vocabulary is the one Appendix A defines, the emotion words present at once in the appraisal geometry, in the VAD lexicon and in the vocabulary of every committee encoder. The 17 are amusement, anger, desire, disappointment, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, pride, relief, remorse, sadness and surprise. The eleven the vocabulary drops are admiration, annoyance, approval, caring, confusion, curiosity, disapproval, nervousness, optimism, realization and neutral, which include the corpus's largest class and its most confusable cluster (Demszky et al., 2020). The adjudicating task is therefore the corpus restricted to the labels that are emotion words. The same two filters on the test split leave 1,444 of 5,427, and Table 5 gives the three nested counts for every task and split.

The other three tasks meet only the second filter. EmpatheticDialogues and TweetEval emotion are single-label by construction, so only the vocabulary bites. The bridge from the 32 published emotion words of the former to the vocabulary is derivational and frozen before any measurement: angry to anger, terrified to terror, and so on for 19 of the 32. We drop the 13 without a derivational counterpart in the vocabulary, such as afraid, devastated and nostalgic. 20 Newsgroups enters with its published by-date split whole, 11,314 training and 7,532 test documents, with headers, footers and quoted text removed. Nothing is subsampled: the declared cap of 20,000 training texts never binds, the largest filtered training split being 12,037.

The validation splits go unread, the classes are unbalanced, and the counts are identities. *The official validation split is not used.* GoEmotions publishes one, and after the same two filters it holds 1,433 texts; no number in this paper reads it, and the same holds for the validation splits of the other two affective corpora. The temperature and the ridge penalty are fitted on the training split, and every reported bit is scored on the test split. The networks of the decoder ladder cut their own held-out tenth out of the training split rather than borrowing the published one. *The training split is unbalanced, and that is why every functional here is macro.* On GoEmotions the classes run from 39 to 1,857 training texts, a spread of nearly fifty to one; a micro-averaged score would be a report on the two largest classes. The measure of Section 2 averages the information gain over labels for that reason. It is also why an item bootstrap over test texts enters the standard error beside the Monte Carlo of the draws. A macro average over a test split whose smallest class holds two texts is itself noisy, and the two errors add in quadrature. *The counts are identities, not estimates.* Each corpus is walked through the same path the measurement itself takes. The three counts nest in every row, the two GoEmotions numbers and the 20 Newsgroups split are the ones the body cites, and no filtered training split reaches the cap.

task	split	published	single-label	in the classes	n	min	max	use
			(texts)			(texts per class)		
GoEmotions	train	43,410	36,308	11,621	17	39	1,857	train
	validation	5,426	4,548	1,433	17	6	261	unused
	test	5,427	4,590	1,444	17	2	260	test
EmpatheticDialogues	train	19,209	19,209	12,037	19	485	997	train
	validation	2,756	2,756	1,672	19	70	148	unused
	test	2,542	2,542	1,573	19	65	127	test
TweetEval emotion	train	3,257	3,257	2,963	3	708	1,400	train
	validation	374	374	346	3	89	160	unused
	test	1,421	1,421	1,298	3	358	558	test
20 Newsgroups	train	11,314	11,314	11,314	20	377	600	train
	test	7,532	7,532	7,532	20	251	399	test

Table 5: **The four tasks, split by split.** *published* is the split as the corpus distributes it; *single-label* keeps the texts carrying exactly one label, which removes nothing in three of the four tasks; *in the classes* keeps, among those, the ones whose label belongs to the task's roster, the 77-term emotion vocabulary for the three affective corpora and all twenty newsgroups for the topic one. n is the resulting number of classes, and *min*/*max* the smallest and largest class in texts. *use*: the training split fits the map and the temperature, the test split carries every reported number, and the published validation splits are not read.

The reference level belongs to the task, and it ranges elevenfold. What the measure subtracts before crediting a geometry is the level a random constellation of the same dimension already reaches, and that level is a property of the task and the decoder. On the primary encoder at $d = 4$ the reference level is 0.086 bits on the 17 emotion labels of GoEmotions, 0.638 on the 19 of EmpatheticDialogues, 0.666 on the 3 of TweetEval and 0.946 on the 20 newsgroup topics. Four tasks of the same dimension therefore start from four different zeros, which is why every level in this appendix is read inside one task. Three tables of this appendix print a reference level for that same cell, and the three differ because each is a Monte Carlo mean over its own draws: 999 in the roster table and in the sentence above, 199 paired draws in the ladder. The Monte Carlo standard error the ladder prints for it is 0.016 bits, and the three estimates agree within it.

The ridge penalty is elected once per cell and then shared by every candidate. The map from text to label coordinates is a ridge regression with intercept under squared loss, and the head is the squared Euclidean distance from the prediction to each label point. The penalty is chosen by the leave-one-out cross-validation of the standard ridge implementation, over a logarithmic grid of 13 values from 10^{-2} to 10^{4}. The text features are *not* standardised, which is the convention the earlier waves of this project fixed. Two clauses matter for reading the tables. The penalty is elected *once per task and encoder*, against a declared reference geometry: the four-dimensional label-name geometry on the affective task, and the label-name geometry at $d = n - 1$ on the topic one. It is then *shared by every candidate of that cell*, the one-hot design included. Sharing it is what makes the contrasts within a row comparable, since no candidate is given a penalty tuned to itself; and it is also not neutral, since the reference geometry is one of the candidates. Table 8 prints the elected values.

The temperature is fitted on the training split, never on what the bits score. The distance head becomes a distribution through one scalar, and the canonical convention refits it per cell by minimising the negative log-likelihood of the distance softmax *on the training split*, by bounded scalar minimisation over $\log \tau \in [-6, 6]$. Two honest clauses follow. The elected $\log \tau$ lies in the interior of that grid in every cell, which we require rather than observe. And the criterion is training likelihood, not test information, so the canonical τ is not the τ that maximises information gain on the test split. No number in this paper is calibrated on what it is scored against, and the price is that every geometry is read at a temperature slightly worse than its best.

The ladder moves the decoder and holds the features, the draws and the geometries fixed. The ladder replaces the ridge by five trained heads on the same features and the same paired draws: a linear map trained by cross-entropy, three one-hidden-layer networks of 64, 256 and 1,024 rectified units, and a two-layer net of 256 units. All five share one implementation, so the optimiser, the schedule and the stopping rule do not vary across rungs. They are Adam at a learning rate of 10^{-3}, at most 200 epochs in mini-batches of 512, early stopping with patience 15 on a stratified tenth of the training split, no dropout and no weight decay, one seed for initialisation and for the split. Two differences from the ridge are worth naming because they are the ones that could explain a level. The features *are* standardised here, by a scaler fitted on the training split alone; and the log-temperature is a free parameter of the network, trained jointly rather than fitted afterwards. That the reference level moves between the linear arms and the non-linear ones, which Section 4 reports, lives in this paragraph and not in the geometries.

The ladder saturates at 64 units, and the arm that breaks ranks is the linear one. Table 6 prints the whole ladder: the held-out information gain and its three parts, with the Monte Carlo standard error over the paired draws, for the four geometries under each of the six decoders. Three things it says. The ridge and the four non-linear rungs agree on the order of the four geometries, and the four non-linear rungs agree with each other on the level as well, so the ladder saturates at 64 units. The level itself jumps between the two linear arms and the non-linear ones, which is why it belongs to the decoder and not to the geometries. The arm that breaks ranks is the linear map trained as a network, which is blind to the spectrum, and that is why the closed forms of Section 3 describe the ridge with a fitted temperature and never "linear decoders". The ladder runs one seed per cell, so a difference between two rungs is read against the run-to-run spread that Appendix C.3 measures over six seeds.

geometry	IG	arrangement gain	assignment gain	position in the range
ridge (closed form): 3,076 parameters, M_0 = 0.068 ± 0.016				
label-name geometry	1.039	0.199 ± 0.020	0.772 ± 0.012	0.73
four-axis appraisal geometry	0.332	0.019 ± 0.022	0.245 ± 0.015	0.20
VAD lexicon	0.455	0.037 ± 0.020	0.349 ± 0.011	0.29
per-dimension oracle	1.393	0.334 ± 0.020	0.991 ± 0.011	1.00
linear map, CE: 3,076 parameters, M_0 = 0.287 ± 0.016				
label-name geometry	1.077	0.040 ± 0.020	0.751 ± 0.013	0.89
four-axis appraisal geometry	0.493	0.016 ± 0.022	0.190 ± 0.015	0.23
VAD lexicon	0.595	−0.034 ± 0.020	0.342 ± 0.012	0.35
per-dimension oracle	1.178	0.162 ± 0.019	0.730 ± 0.011	1.00
MLP-64: 49,476 parameters, M_0 = −0.354 ± 0.032				
label-name geometry	0.984	0.401 ± 0.035	0.937 ± 0.014	0.81
four-axis appraisal geometry	0.250	0.171 ± 0.037	0.434 ± 0.019	0.37
VAD lexicon	0.370	0.371 ± 0.035	0.353 ± 0.014	0.44
per-dimension oracle	1.296	0.992 ± 0.034	0.658 ± 0.011	1.00
MLP-256: 197,892 parameters, M_0 = −0.365 ± 0.033				
label-name geometry	0.966	0.422 ± 0.036	0.909 ± 0.015	0.78
four-axis appraisal geometry	0.164	0.170 ± 0.038	0.359 ± 0.019	0.31
VAD lexicon	0.470	0.388 ± 0.036	0.447 ± 0.014	0.49
per-dimension oracle	1.341	1.037 ± 0.034	0.669 ± 0.011	1.00
MLP-1024: 791,556 parameters, M_0 = −0.395 ± 0.033				
label-name geometry	0.921	0.401 ± 0.037	0.916 ± 0.016	0.76
four-axis appraisal geometry	0.170	0.159 ± 0.039	0.406 ± 0.019	0.33
VAD lexicon	0.389	0.378 ± 0.036	0.406 ± 0.014	0.46
per-dimension oracle	1.327	1.069 ± 0.035	0.654 ± 0.011	1.00
MLP-256×2: 263,684 parameters, M_0 = −0.379 ± 0.034				
label-name geometry	0.839	0.453 ± 0.037	0.765 ± 0.014	0.73
four-axis appraisal geometry	0.154	0.202 ± 0.039	0.331 ± 0.019	0.32
VAD lexicon	0.384	0.425 ± 0.037	0.338 ± 0.014	0.45
per-dimension oracle	1.299	0.987 ± 0.036	0.691 ± 0.011	1.00

Table 6: **The decoder ladder, in full: the primary encoder, $d = 4$, 199 paired draws.** Six decoders on the same features, from the ridge in closed form to a two-layer network; four geometries under each. Every entry is in bits: the held-out information gain IG, the arrangement gain and the assignment gain, each with the Monte Carlo standard error over the draws; the reference level M_0 and the parameter count belong to the decoder and head each block. The last column places IG in the range from M_0 to the per-dimension oracle, which is a lower bound on the ceiling at $d = 4$ and sits at 1 by construction. The reference level moves between the linear arms and the non-linear ones and then stays; the arrangement gain of the four-axis geometry is at zero under the two linear arms and near 0.17 from the first non-linear rung on.

Four encoders replicate every measurement, and all four are frozen. The four are mpnet at 768 dimensions, GTE-large and E5-large-v2 at 1,024, and a non-neural control. The control is TF-IDF over unigrams and bigrams, with a minimum document frequency of two and sublinear term frequency, truncated by SVD to 768 components. E5 receives the query prefix its model card requires. Freezing them is what keeps the screen applicable. With an encoder trained on the target, the ridge operator of Section 2 depends on the geometry of that target, and the prediction is no longer one product GZ. The screen of Section 3 reads c_G off that operator, so it no longer applies.

No encoder sits on both sides of the measurement. The committee of 12 sentence encoders, four static, four contextual and four contrastive, enters only to build the label-name geometry from the distances among the label names, and shares no member with the four encoders of the texts. The primary text encoder, mpnet, belongs to the committee of 15 from which this one is cut; we removed it together with the lowest-weighted member of each other class, so that no encoder sits on both sides of the measurement, and the two committees' distance matrices agree at a Spearman correlation of 0.99 over the label pairs. Table 7 lists both rosters, the weight of each member in

each, and the agreement under four measures; the geometries built by dropping one further member at a time are the leave-one-encoder-out variants of Section 4.

encoder	**class**	RV weight, 15	RV weight, 12	**status**
word2vec	static	0.076	0.093	kept
GloVe	static	0.075	0.092	kept
fastText	static	0.075	0.092	kept
Numberbatch	static	0.080	0.098	kept
GloVe-Twitter	static	0.066	—	removed: lowest weight of its class
BERT (iso)	contextual	0.058	0.071	kept
BERT (tmpl)	contextual	0.064	0.080	kept
RoBERTa (iso)	contextual	0.030	0.037	kept: declared outlier (anisotropy), skipped by the rule
RoBERTa (tmpl)	contextual	0.065	0.079	kept
DeBERTa-v3 (tmpl)	contextual	0.043	—	removed: lowest weight of its class
MiniLM	contrastive	0.070	0.085	kept
mpnet	contrastive	0.074	—	removed: primary text encoder
BGE-M3	contrastive	0.074	0.091	kept
Qwen3-Emb	contrastive	0.070	0.086	kept
OpenAI-3-large	contrastive	0.079	0.097	kept

agreement of the two label-name geometries	
Spearman over the label-pair distances	0.993
Kendall τ_b, idem	0.931
RV coefficient between cross-products	0.996
debiased CKA	0.994

Table 7: **The committee that builds the label-name geometry: 15 members, and the 12 the paper uses.** The weight is the RV weight the committee aggregate assigns to each member, re-estimated on the 12; weights sum to one within each committee. The rule removes the primary text encoder and then the lowest-weighted member of each remaining class; in the contextual class the literal lowest, RoBERTa without a template, is a declared outlier by anisotropy and a degenerate member of the weighting, so removing it would improve the committee by construction and confound the control, and the rule skips to the next. The alternative that removes it instead agrees with the 15 to the same figures. Below: the agreement between the aggregated distance matrices of the two committees over the pairs of the 77 terms.

The draw budgets are fixed, and the precision floor sits under every threshold. The draw budgets are 999 for the reference level and 999 for the total fibre, 499 for the fibre under a declared partition, 199 paired draws on the ladder and 2,000 item-bootstrap resamples. Everything runs in `float32` on one consumer GPU, and that fixes a floor. The ridge operator is formed once per encoder from a Gram matrix over $11{,}621 \times 768$ or $\times 1{,}024$ entries in that precision. Changing only the thread count moves the information gain of a cell by an amount of order 10^{-8} bits, and the drift between sessions is of the same order. The perturbation is common to every evaluation inside a cell, so it does not average down over draws; every reproduction threshold in this paper is set above it.

Every constant of the fit, in one place. Table 8 gathers them: the penalty grid and the elected penalty, the temperature grid and its objective, the rungs and the optimiser of the ladder, the draw budgets, the feature dimension of each encoder, and the training cap. Every value there is the one the fit used, tabulated rather than transcribed. Two rows repay a second look. The elected penalty is where a reader checks the sharing rule above, one value per task and encoder for every candidate geometry of that cell. Feature standardisation differs between the ridge and the ladder on purpose, and it is the one asymmetry that could move a level between the two decoder families.

stage	setting	value
ridge	penalty grid	13 values, 0.01–10,000
	elected λ, affective task	1.00 / 0.10 / 0.10 / 1.00
	elected λ, topic task	0.32 / 0.10 / 0.10 / 0.32
	feature standardisation	none
temperature	grid for $\log \tau$	-6 to 6
	objective	train NLL, refit per cell
MLP rungs	hidden units	64 / 256 / 1,024 / 256×2
	optimiser, learning rate	Adam, 0.001
	max epochs, patience	200, 15
	batch size	512
	dropout, weight decay	0, 0
	validation split	0.10 of train, stratified, seed 7
	feature standardisation	standardised
draws	reference level, fibre	999, 999
	fibre under a partition	499
	paired draws, decoder ladder	199
	item bootstrap	2,000
encoders	feature dimension	768 / 1,024 / 1,024 / 768
data	training cap, seed	20,000, 0

Table 8: **Every constant of the fit, in one place.** The four values of a row are the four encoders in the order mpnet, GTE-large, E5-large-v2, TF-IDF-SVD. The elected λ is one value per task and encoder, shared by every candidate geometry of that cell. *feature standardisation* differs between the two decoder families on purpose and is the one asymmetry that could move a level.

The second candidate set deforms one geometry in two families, and both move the spectrum. The second candidate set of Section 4 starts from the label-name geometry at $d = 4$ and deforms it in two ways, both fixed before any measurement. The first family collapses the nearest pairs of label points onto their midpoints, taken greedily in order of distance and never reusing a point, for one pair up to six: six members. The second flattens the smallest principal component of the same geometry, multiplying its singular value by seven tenths, four tenths, one fifth and one twentieth: four members. Three undeformed geometries join them, the label-name geometry itself, the appraisal geometry, and a Gaussian constellation drawn once from a declared seed and carried into the admissible set by the same routine as the others. Three plus six plus four is the 13 the body counts. Merging a pair and flattening an axis both change the eigenvalues, which is why every spectral function orders this set, and why the family below is the one that separates the two scalars.

The orientation family fixes the spectrum and checks its own order. The third candidate set of Section 4 fixes the spectrum of the label-name geometry at $d = 4$, with its observed trace, and draws 48 orthonormal frames uniformly on the Stiefel manifold of $\mathbf{1}^{\perp}$. This is the orientation null, a Haar rotation given the spectrum, of the eigenstrapping (Koussis et al., 2025) and tensor-maximum-entropy (Elsayed & Cunningham, 2017) family. Each frame realises that same spectrum with a different assignment of directions to labels, so the family varies the diagonal of K and nothing the spectrum determines. Every fibre mean is estimated from 999 relabellings, and one list of relabellings serves every geometry. The Monte Carlo error is then common to the whole family, and the order among its members is estimated far better than any single mean. Splitting that list into its odd and even halves gives two estimates of the order, and the Spearman correlation between them is the reliability the body reports. By construction the spectral scalars are constant across the draws, and every draw realises the declared spectrum and the declared trace.

C.2 Full tables, encoder by encoder

The body averages over the four encoders, and this subsection prints them one by one, in the order the body asks for them. Each float below carries the paragraph that says what it adds. Convention throughout: canonical temperature, reference A, 999 draws; the errors are Monte Carlo standard errors and none of these numbers carries an interval. The second experiments, with their two tables and their figure, follow in Appendix C.3.

The four geometries of the roster differ before any number is read. Figure 7 puts them on the same 17 classes, one frame each. What to read in it is the radii, the distances of the label points from their centroid, because that is one of the two numbers the screen of Section 3 combines. The standard design has a single radius, so its frame is a ring and its radius spread is zero by construction. The other three scatter their points off the dashed circle, each in its own dimension and on its own scale, so none of them has a spread of zero. The frames are two-dimensional projections, while every number in this paper uses the dimension printed in the frame.

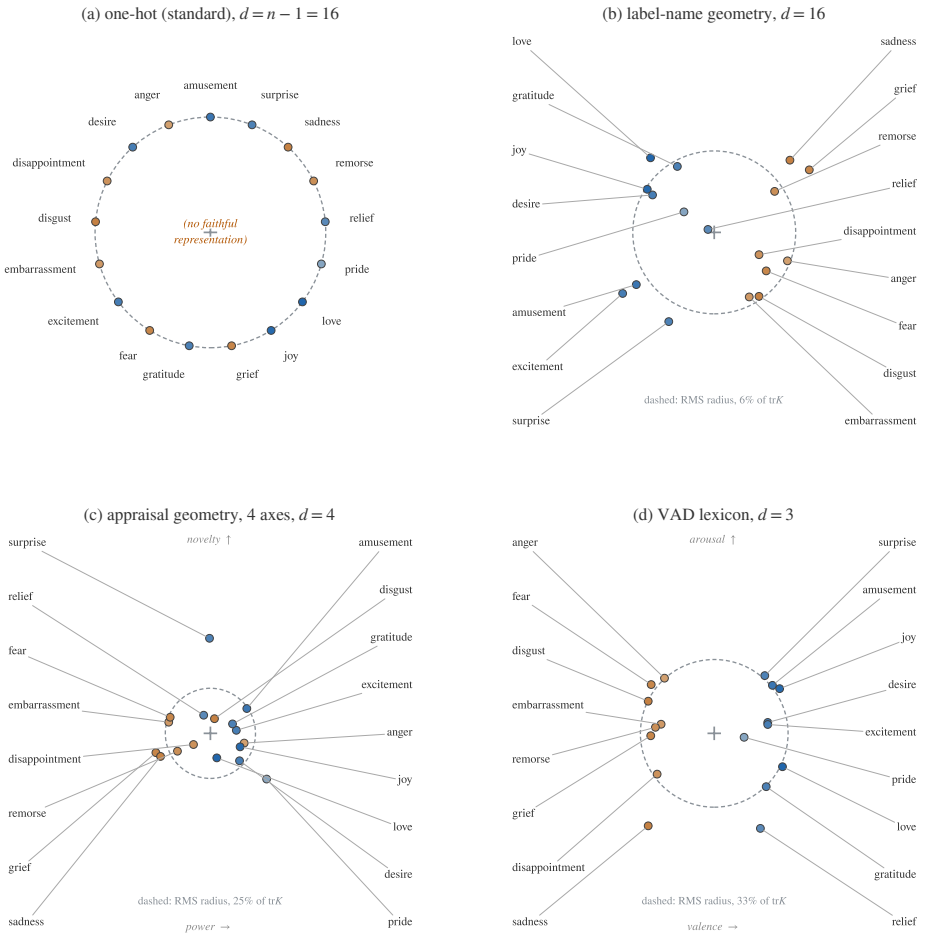


Figure 7: **The four label geometries of the affective task, on the same 17 rows.** Each frame is a two-dimensional picture of one geometry over the 17 classes, centred and scaled into the admissible set; the dashed circle is the root-mean-square radius, so frames are not on a common scale, and colour is valence in the VAD lexicon. (a) The standard design at $d = 16$, a ring because all 136 pairwise distances are equal. (b) The label-name geometry at $d = 16$, on its first two axes. (c) The appraisal geometry at $d = 4$, on its two highest-variance axes. (d) The VAD lexicon at $d = 3$, on valence and arousal. The pictures are for reading; every number uses the dimension in the frame.

Averaging over encoders hides the spread, and the full table shows it. Table 9 is Table 2 before the average: one column per encoder, four predictors per candidate set. It adds two things. The spread across encoders becomes visible, so a reader can see whether a margin in the mean rests on all four encoders or on one of them. And its third block is the 13 constructions on the topic task, which the body's table replaces with the orientation family, so each spectral predictor can be read on a second task where the constructions move the spectrum by design.

candidate set / predictor	mpnet	GTE-large	E5-large-v2	TF-IDF-SVD
committee + 12 leave-one-encoder-out subsets (GoEmotions)				
two scalars	0.956	0.934	0.863	0.758
radius spread alone	0.929	0.907	0.874	0.874
spectral concentration alone	0.170	0.176	0.165	−0.038
log-determinant	0.115	0.126	0.115	−0.110
13 constructions (GoEmotions)				
two scalars	0.940	0.890	0.890	0.978
radius spread alone	0.956	0.907	0.907	0.995
spectral concentration alone	0.775	0.802	0.802	0.769
log-determinant	0.736	0.786	0.786	0.709
13 constructions (20 Newsgroups)				
two scalars	0.874	0.907	0.984	0.940
radius spread alone	0.758	0.874	0.929	0.934
spectral concentration alone	0.863	0.764	0.819	0.863
log-determinant	0.857	0.687	0.819	0.846

Table 9: **Table 2, encoder by encoder.** Spearman correlation between each predictor and the canonical arrangement gain, oriented so that positive is the predicted order. Three candidate sets: the leave-one-encoder-out variants and the constructions on the adjudicating task, and the constructions on the topic task.

The roster table prints the two scalars of the screen, and the log-determinant beside them as the comparator. Table 10 is Table 1 per encoder, with four readings the body leaves out. The null fraction φ places each geometry against the label-permutation null, and it is undefined for the standard design, whose fibre is one point. The two scales that cross dimensions, $\mathrm{IG} - M_0$ and the budget share ω, let a geometry at $d = 4$ be set beside one at $d = 16$. The rent is printed at every dimension, an identity at $d = n - 1$ and, in italics, a prediction below it; the exchange rate η is printed only where the rent is an identity. And three columns sit beside each geometry, all computed on the admissible representative. The variance of the squared radii and $\mathrm{tr}\, K^2$ are divided by the squared trace; the log-determinant is taken on the spectrum truncated to the rank and divided by the trace, so it does not compare two geometries of different dimension. Two of those three are the scalars the screen combines, the radius spread and the spectral concentration. The third is the comparator and not a third scalar of the screen: the log-determinant is a function of the spectrum alone, so it sees the concentration and is blind to the radii. All three are printed once per geometry because they are properties of the label geometry and not of the encoder.

geometry	d	encoder	IG	M_0	arr.	asg.	φ	IG$-M_0$	ω	$\mathrm{Var}(r)$	$\mathrm{tr}\,K^2$	$\log\det$	rent	η
			(bits)					(bits)		$\times 10^3$			(bits)	
one-hot (standard)	16	mpnet	1.235	0.547	0.688	0.000	—	0.688	1.000	0.00	0.0625	−44.361	0.000	—
		GTE-large	1.464	0.756	0.709	0.000	—	0.709	1.000				0.000	—
		E5-large-v2	1.743	0.971	0.771	0.000	—	0.771	1.000				0.000	—
		TF-IDF-SVD	1.638	0.926	0.712	0.000	—	0.712	1.000				0.000	—
label-name geometry	16	mpnet	1.437	0.547	0.188	0.702	1.00	0.890	0.273	0.11	0.1327	−64.014	0.500	1.404
		GTE-large	1.701	0.756	0.198	0.747	1.00	0.945	0.280				0.510	1.463
		E5-large-v2	1.890	0.971	0.207	0.712	1.00	0.919	0.268				0.565	1.262
		TF-IDF-SVD	1.493	0.926	0.265	0.302	0.99	0.567	0.372				0.448	0.676
appraisal geometry, 4 axes	4	mpnet	0.332	0.080	0.010	0.241	0.90	0.252	0.009	1.32	0.3651	−6.740	*0.677*	—
		GTE-large	0.463	0.159	0.010	0.294	0.93	0.303	0.008				*0.699*	—
		E5-large-v2	0.381	0.252	−0.001	0.129	0.68	0.129	0.000				*0.772*	—
		TF-IDF-SVD	−0.073	0.263	0.023	−0.359	0.08	−0.336	0.017				*0.689*	—
VAD lexicon	3	mpnet	0.455	0.048	0.059	0.347	1.00	0.406	0.050	0.81	0.5268	−4.609	*0.628*	—
		GTE-large	0.641	0.107	0.064	0.470	1.00	0.534	0.047				*0.645*	—
		E5-large-v2	0.684	0.175	0.060	0.449	1.00	0.509	0.038				*0.712*	—
		TF-IDF-SVD	0.444	0.194	0.101	0.149	0.77	0.250	0.070				*0.611*	—

Table 10: **GoEmotions under the ridge decoder, encoder by encoder, with the two scalars of the screen and the log-determinant beside each geometry.** Columns as in Table 1; φ is the null fraction, undefined for the standard design; $\mathrm{IG} - M_0$ and ω are the two scales that cross dimensions; $\mathrm{Var}(r)$ is the variance of the squared radii and $\mathrm{tr}\,K^2$ the spectral concentration, the two the screen combines, both divided by the squared trace; the log-determinant, divided by the trace, is the spectral comparator and is read only within a dimension.

Six of the eight dimension-matched cells pay for the trade under the ridge. The trade of Section 4 is read on the eight dimension-matched cells of both tasks in Table 11: under the ridge the exchange rate exceeds one in six of the eight, and one of the two that fail collects a negative budget share, its arrangement gain below the reference level; nothing in the construction forbids that. The eight rows are the four encoders of each task at $d = n - 1$, the dimension in which the standard design is itself a candidate and the rent is therefore an identity. The budget and the rent are printed beside the share and the rate, so the identity of Section 3 can be checked row by row. The share and the rate answer different questions, one about the arrangement alone and one about the whole trade, and the table lets a reader see that they need not agree.

The paired contrast gives the trade an interval. The identity of Section 3 carries no interval, so Section 4 also reads the trade as a paired contrast over the held-out texts, one 95% interval per encoder that combines the Monte Carlo error with a bootstrap over texts. On GoEmotions at $d = 16$ the label-name geometry minus the one-hot code is -0.57 to -0.45 bits of arrangement gain and $+0.30$ to $+0.75$ of assignment gain across the four encoders, each excluding zero and clearing the resolution criterion of Section 2 in every one of them. The total is $+0.15$ to $+0.24$ bits, excluding zero in the three neural encoders and clearing the criterion in one of them, against -0.15 in the non-neural control, excluding zero and clearing the criterion on the negative side.

task	encoder	d	budget	rent (Jensen gap)	share ω	η
GoEmotions	mpnet	16	0.688	0.500	0.273	1.404
	GTE-large	16	0.709	0.510	0.280	1.463
	E5-large-v2	16	0.771	0.565	0.268	1.262
	TF-IDF-SVD	16	0.712	0.448	0.372	0.676
20 Newsgroups	mpnet	19	0.539	0.605	−0.123	0.974
	GTE-large	19	0.516	0.357	0.308	1.338
	E5-large-v2	19	0.597	0.404	0.323	1.247
	TF-IDF-SVD	19	0.800	0.177	0.778	1.386

Table 11: **The eight dimension-matched cells ($d = n - 1$), in which the simplex is in the admissible set, under the ridge.** Budget $= \mathrm{IG}(K_s) - M_0$; rent $= \mathrm{IG}(K_s) - M_1(\mathrm{map})$; $\omega =$ arrangement gain of the map over the budget; $\eta =$ assignment gain over rent, and the map delivers more in total than the standard design when $\eta > 1$. The geometry is the label-name geometry at $d = n - 1$.

C.3 SECOND EXPERIMENTS: A TRAINED DECODER, AND A TASK OUTSIDE AFFECT

Two experiments that vary what the body holds fixed: the decoder, and the family of the task. Neither changes the measure; both change what it returns. Convention as in Appendix C.2: canonical temperature, reference A, 999 draws for the reference level, and Monte Carlo standard errors without an interval. The fibre budget is the one each experiment declares, 199 paired draws under the network and 499 under the partition of the topic cell.

Under a two-layer network the same map no longer pays in any encoder, and the sign of the trade is what travels. The ridge is the measuring device of this paper and not a competitor: it is held fixed so that the geometry, and not the decoder, is what gets measured. A decoder free to separate what the ridge could not absorbs part of what a geometry sells. The mechanism is the one of Section 2: an assignment earns bits only where the decoder did not separate the labels, so a decoder that separates what the ridge confused leaves less for the assignment to earn, while the standard design, which has nothing to assign, loses nothing. The ridge reads the information in the geometry itself; a stronger decoder reads the geometry through its own capacity. So the question is what a stronger decoder does to the trade. Table 12 repeats the roster under a two-layer network of 256 units trained by cross-entropy with a free log-temperature, on the same features and the same paired draws. Levels do not compare across decoders, so what the two tables share is the direction of each part and the sign of the trade, never the size of a level. The exchange rate is a ratio formed inside one decoder, so it is the one column the two tables can be read side by side. Under the network the standard design's arrangement gain rises in three of the four encoders and the label-name geometry's assignment gain falls, and the exchange rate of the label-name geometry at $d = 16$ drops below one in every encoder, from 0.75 on E5-large-v2 to 0.30 on the non-neural control at the training seed of Table 12. Six training seeds give the same reading: the rate stays below one in all 24 seed-encoder cells, with means of 0.65 to 0.67 in the three neural encoders and 0.34 in the control, and a spread across seeds of about 0.1. The same map that pays for its rent under the ridge no longer does under a decoder able to separate what the ridge could not. The sign of each part and the leader in assignment travel between the two decoders, the balance does not, and "six of eight" is a statement about the ridge. The same holds across a ladder of six decoder families on the primary encoder (Appendix C.1): whoever compares candidates before training gets the leader and the sign, not the level. The body's "three of the four encoders pay" is, in the same way, a statement about the ridge.

geometry	d	**encoder**	IG	M_0	arr.	asg.	rent	η
					(bits)			
one-hot (standard)	16	mpnet	1.335	0.404	0.931	0.000	0.000	—
		GTE-large	1.678	0.824	0.854	0.000	0.000	—
		E5-large-v2	1.868	1.165	0.702	0.000	0.000	—
		TF-IDF-SVD	1.363	−0.046	1.409	0.000	0.000	—
label-name geometry	16	mpnet	1.093	0.404	0.208	0.481	0.723	0.665
		GTE-large	1.458	0.824	0.206	0.428	0.648	0.661
		E5-large-v2	1.735	1.165	0.182	0.388	0.520	0.746
		TF-IDF-SVD	0.730	−0.046	0.500	0.276	0.909	0.304
appraisal geometry, 4 axes	4	mpnet	0.130	−0.455	0.245	0.340	*1.545*	—
		GTE-large	0.502	−0.159	0.161	0.499	*1.675*	—
		E5-large-v2	0.664	0.137	0.131	0.396	*1.600*	—
		TF-IDF-SVD	−0.133	−0.139	0.151	−0.145	*1.351*	—
VAD lexicon	3	mpnet	0.371	−0.336	0.371	0.336	*1.300*	—
		GTE-large	0.801	−0.109	0.279	0.632	*1.509*	—
		E5-large-v2	0.879	0.146	0.220	0.514	*1.502*	—
		TF-IDF-SVD	0.066	−0.018	0.291	−0.206	*1.090*	—

Table 12: **The same roster under a two-layer network of 256 units, on the same features and the same paired draws.** Every column is read as in Table 1; the rent is an identity at $d = n - 1$ and a prediction below it, in italics, with the ceiling taken from the one-hot design under this same decoder. One training seed per cell.

Outside affect the parts still sum, and the predictor loses its margin. The body is affective text. The second family is a forum corpus whose taxonomy the corpus itself publishes, and it asks whether the construction depends on emotion words. Figure 8 reads it: the three parts sum to the information gain exactly, for the label-name geometry and for the published taxonomy turned into a geometry; and the two scalars keep the order here but lose their margin, because the thirteen constructions move the spectrum by design and the log-determinant orders too, the narrow regime of Section 3. One negative, pre-declared and lost: the geometry that is pure taxonomy, the blocks of the tree as points and nothing inside them, carries positive arrangement gain in every encoder. A taxonomy is not a naming device with no geometry; it also arranges the points. Panel (a) of the figure is the accounting, encoder by encoder; panel (b) is the order, the two scalars together, each of them alone and the log-determinant, over the 13 constructions of each encoder. Table 13 prints the cell: the three parts, the fibre fraction φ under the full permutation null and under the null that concedes the published prefixes, and the assignment gain that survives inside the prefixes. Both geometries of this cell live at $d = 4$, while Table 11 reads the label-name geometry of the same task at $d = n - 1$, so the two tables are two dimensions of one task and not two readings of one object.

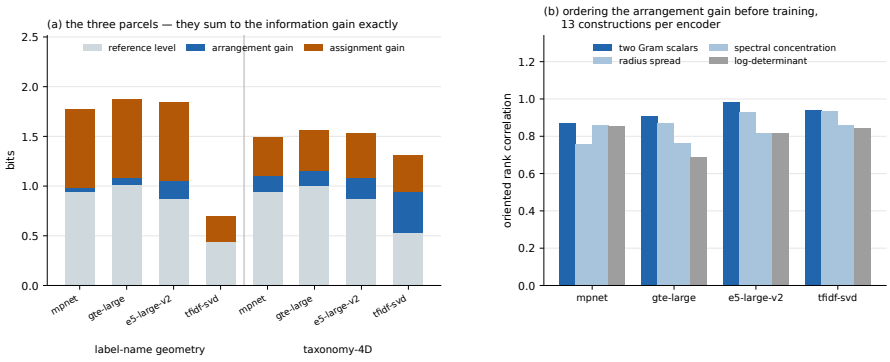


Figure 8: **The topic task, both geometries at $d = 4$.** (a) The three parts per encoder for the label-name geometry and the taxonomy geometry; they sum to the information gain exactly. (b) The oriented rank correlation of the two scalars, of each alone and of the log-determinant against the measured arrangement gain, 13 constructions per encoder.

geometry	**encoder**	IG	M_0	arr.	asg.	φ (S_n)	φ (prefixes)	asg. within prefixes
label-name	mpnet	1.774	0.946	0.039	0.789	1.00	1.00	0.267
	GTE-large	1.876	1.009	0.078	0.790	1.00	1.00	0.346
	E5-large-v2	1.844	0.868	0.188	0.788	1.00	1.00	0.334
	TF-IDF-SVD	0.697	0.525	−0.085	0.258	1.00	0.95	0.141
taxonomy	mpnet	1.494	0.946	0.156	0.392	1.00	undefined	≈ 0 (identity)
	GTE-large	1.557	1.009	0.142	0.407	1.00	undefined	≈ 0 (identity)
	E5-large-v2	1.529	0.868	0.216	0.445	1.00	undefined	≈ 0 (identity)
	TF-IDF-SVD	1.313	0.525	0.419	0.369	1.00	undefined	≈ 0 (identity)

Table 13: **The task outside affect: 20 Newsgroups, two geometries at $d = 4$, four encoders.** The label-name geometry is built from the names of the 20 classes by the same committee and cut to $d = 4$; the taxonomy geometry places the seven published prefixes as points and nothing inside them. IG, M_0 and the two gains are in bits and sum exactly. φ is the fraction of the fibre the geometry beats: under the full permutation null (S_n), and under the null that permutes names only within a prefix, which concedes the taxonomy. The last column is the assignment gain measured against that conceding null: what naming earns once the tree is given. For the taxonomy geometry that quantity is an identity, zero by construction, since within-prefix permutations fix it.

D RELATED WORK: FOUR FAMILIES THAT PUT INFORMATION INTO THE LABELS

The families named in Section 1 all put information into the labels of a classifier. What none of them has is a scale with a declared zero, a null for each part, and the separation of arrangement from assignment inside one experiment, which is what Section 2 builds. The first three are about the output side of a classifier, where a geometry is chosen or arrived at by training. The fourth and fifth randomise the labels, the manoeuvre closest to ours; the sixth and seventh are what the measure inherits, and the eighth is the scope. Three words recur: the *fibre* of an arrangement is the set of its relabellings, *rent* and *budget* are the two sides of the identity of Section 3, and a *cell* is one task and one encoder. The psychometric lineage of the two published geometries of the roster is Appendix A.

Label embedding and zero-shot classification swap the output space, and do not need to separate the two sources. Palatucci et al. (2009) encode classes as semantic output codes, so that a class never seen in training stays predictable. Akata et al. (2013) embed the classes in a space of human-defined attributes and learn a compatibility with the image features. Akata et al. (2015) hold the image encoder fixed and vary only the output geometry, comparing attributes, word vectors and the WordNet hierarchy, and Xian et al. (2019) standardise the protocol for the whole family. That is the ancestor of the design used here: one encoder, several label geometries, one task. What the measure says is that swapping the space moves where the points lie and which class sits where at once, so the two sources arrive together in every cell. Their rulers are accuracy and its variants, and none of them carries a declared zero for a label geometry.

Fixed classifiers report one number, and all but one make a single decision. Mensink et al. (2013) and Snell et al. (2017) classify by distance to a class prototype, the decoder of Section 2 without a temperature. Hoffer et al. (2018) replace the learned prototype matrix by a fixed one and measure the cost, while Yang et al. (2022) and Li et al. (2023) freeze the head at the equiangular frame; those three let the assignment fall where the class order left it. Mettes et al. (2019) place prototypes at maximal angular separation and add a ranking term that makes the prototype similarities follow the order of the class-name similarities, the only place in this family where both decisions are named in one objective. Saadabadi et al. (2024) instead optimise the label-to-prototype assignment with the arrangement held fixed, and each reports the outcome as a single accuracy, so the two decisions arrive summed. The measure separates them: a head frozen at the equiangular frame collects no assignment gain, by construction and not by measurement, because its fibre is a single point (Proposition 2). Prototypes placed by class names pay rent in order to name, and a dynamic assignment buys back what the others leave at the fibre mean.

Neural collapse characterises the arrangement that training reaches, and the measure prices it. Papyan et al. (2020) report that the last-layer class means of a trained network converge to a simplex equiangular tight frame, and Zhu et al. (2021) and Zhou et al. (2022) recover it as the global solution of an unconstrained-features problem, under cross-entropy and under squared error. Han et al. (2022, Prop. 1) follow the squared-error case along a central path on which the least-squares head refits to any permutation of the class means. The term of their analysis comparing those means to the one-hot target is therefore a function of one spectrum, blind to which class sits at which vertex. Massion et al. (2026, Prop. IV.3) carry it to an arbitrary encoding of the classes. With unconstrained features and balanced classes, exchanging which class receives which code only permutes blocks of the optimum, so the naming is worth nothing whatever the encoding; Wu & Mondelli (2025) study collapse beyond that model, proving that within-class variability vanishes in a three-layer mean-field network rather than under free features. What the unit adds is not which arrangement appears but how arrangements are ordered, and what part the naming contributes, which with unconstrained features is nothing. The same geometry appears in this paper as the fixed point of the label-permutation null (Proposition 2), which Section 5 reads. And the nearest-class-mean rule this literature uses to test its own predictions is the decoder of Section 2 at zero temperature.

Declared taxonomies inject a tree, and one of them randomises it as a condition rather than as a reference distribution. Bertinetto et al. (2020) take a declared hierarchy and charge an error by the height of the lowest common ancestor, and Kosmopoulos et al. (2015) unify such measures, which score distance in the tree and not the information the tree puts into the labels. Brust & Denzler

(2020) ask it in its barest form, how much a domain hierarchy adds to a deep classifier, which is what a part answers in bits against a declared zero. Wu et al. (2019) exhibit a case where a hierarchical loss on a non-hierarchical problem degrades it. The randomised tree is theirs and not ours: Bertinetto et al. (2020) randomise the nodes of the hierarchies, repeat the experiments, and the tree-distance measures deteriorate sharply. For several values of their two loss parameters the top-1 error under the random hierarchy is consistently lower than under the real one: the declared tree costs accuracy and buys error severity. They exchange one tree for another as an experimental condition, whereas the null that concedes a taxonomy permutes labels only inside declared blocks, a reference distribution whose moments are closed forms, and reports bits. The easy reading is that a tree only names, and it is wrong: in bits a declared taxonomy delivers both parts, which Appendix C.3 reports on a corpus that publishes its own taxonomy.

Erasing or swapping the assignment is the closest family, and it is the external control of the assignment gain. Dietterich & Bakiri (1995) reassign codewords to classes at random and find no appreciable variation, and Evron et al. (2023) fix the codebook, vary only the codeword-to-class assignment, and find that it matters. The measure predicts the pair: under a code drawn without structure the assignment gain is zero, and under a structured code it is not, which is what the second finds. Min et al. (2022) replace the labels of in-context demonstrations by labels drawn at random from the same space, separating it from the input-to-label pairing. Hewitt & Liang (2019) give a probe a control task with behaviour assigned at random, a null over the naming with no arrangement to preserve. Levy et al. (2015) make the same move on the input side, splitting the vocabulary between training and test, and find the credit in the term. Where the assignment is arbitrary the assignment gain should be zero, and this is where that case has already been exercised.

The form of the decomposition is a conditional score decomposition, and the unit is the ignorance score. The functional of Section 2 is the logarithmic score of Good (1952), read against the uniform forecast and expressed in bits, which is the ignorance score of Roulston & Smith (2002). Bröcker (2009) decomposes a proper score and ties it to sufficiency, and Allen et al. (2023) condition a proper score on auxiliary information with the same telescoping. They leave the order of the increments open; we settle it by precedence in the sense of Faigle & Kern (1992), because "naming without arrangement" has no instance to evaluate. The scheme is standard in global sensitivity analysis, where Owen (2014) applies the value of Shapley (1953) to conditional expectations and Song et al. (2016) estimate the result. In all of them the conditioning variable lives on the input side: variables of a model, states of a forecasting situation, or a representation of the text (Hewitt et al., 2021; Xu et al., 2020; Ethayarajh et al., 2022). Here it takes three values, degrees of knowledge about the geometry of the output. What the change buys is a null for each increment: knowing the arrangement and knowing which label sits where are conditions a reference distribution realises by drawing, which an input representation is not.

Nulls that preserve a statistic have a literature, and so does the null that concedes structure. The label-permutation null is the quadratic-assignment lineage of Mantel (1967) and Hubert & Schultz (1976), with Ojala & Garriga (2010) for the argument that different randomisations answer different questions. The law on the fibre is uniform because the reference is exchangeable, and the orientation null, a rotation drawn given the spectrum, is eigenstrapping (Koussis et al., 2025) and the tensor-maximum-entropy family (Elsayed & Cunningham, 2017). Averaging a Gram matrix over the orbit of a finite subgroup is the orthogonal projection on the fixed-point subspace of Shah & Chandrasekaran (2012, Lemma 2.2), who also apply Jensen to a group mean. The variance reduction of an orbit mean is Chen et al. (2020), and the basis of the linear maps that commute with every permutation, indexed by partitions, is due to Maron et al. (2019), published at this venue. Winkler et al. (2015) permute only inside blocks exchangeable under the null, a requirement of validity and not a concession of structure, and Alexander-Bloch et al. (2018) give the field precedent for conceding structure on purpose. What is left to us is one claim in two halves, and neither is the algebra. A Young subgroup is the null that a taxonomy declared from outside calls for, where that literature permutes inside blocks for validity. And the second moment of this functional under that subgroup closes where the group is too large to enumerate, which makes the Jensen gap a quantity to report.

The measure applies wherever there is a label Gram matrix and a frozen representation. The input is always the same: a Gram matrix over the classes, a representation that does not move with

the labelling, and a declared partition when the null that concedes a taxonomy is used. In-context label spaces (Min et al., 2022), probes read against a declared reference (Hewitt & Liang, 2019; Hewitt et al., 2021) and fixed classification heads (Hoffer et al., 2018; Mettes et al., 2019) all supply that input. The output is an order, or a part read against a declared zero, and never a level with the reference and the dimension left unsaid. That is what this appendix owes the families above: a declared zero, a null for each part, and arrangement separated from assignment in one experiment. The regime the construction does not cover is end-to-end training, where the representation moves with the geometry and the objection of Appendix C.3 returns in full. We have not run these three cases; they are the input the construction takes, not results.