SELF-SUPERVISED ENCODERS LEARN A TASK PRIOR IN THEIR SPECTRUM THAT ISOTROPY ERASES

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised objectives increasingly push representations toward an isotropic Gaussian, which LeJEPA proves optimal for worst-case linear probes. We show that once tasks come from a distribution, a representation's spectrum is a label-free prior over them, and isotropy is the uniform prior. A one-penalty ridge probe on spectrum s is generalised ridge on the latents with per-direction penalty λ/s_i, and is the Bayes (Wiener) estimator iff $s \propto p$, the prior's variances, for any sample size, dimension and design that excites every direction (Lindley–Smith read from the representation's side), and a power-law prior recovers the spectral-bias learning-curve rate. Theorem A prices a spectrum in labels, and Theorem B shows that a C-class task's prior has rank $C-1$, so isotropy flattens exactly the directions that carry the task. Under a protocol on 23 encoders and 15 vision and text tasks (169 cells), isotropy costs a median 21 accuracy points at 200 labels (range 0.5–64), and recovering the discarded prior takes a median 500 labels, a price predicted ($\rho = 0.89$) by the cost of detecting the weakest task direction, found post hoc on 20 cells, then replicated on the other 149. The prior comes back without labels. Averaged over *other* tasks, it recovers a median 89 % of the oracle gain, and training toward a non-flat spectral target instead of isotropy raises 200-label accuracy from 60 % to 80 %, with the isotropy-trained encoder still worse even given the oracle prior. A supervised control pays the same price, so the effect is a property of linear probing itself.

1 INTRODUCTION

Isotropy has become a training target. LeJEPA (Balestriero & LeCun, 2025) proves that, among embedding distributions with a fixed covariance budget, the isotropic Gaussian uniquely minimises the worst-case risk of a linear probe, and enforces it with SIGReg. Concurrent preprints refine the same target (Nicollier et al., 2026; Alvarez, 2026; Le, 2026). None puts a *distribution* over the downstream tasks. This paper does. Put a prior $w \sim (0, P)$ on the linear read-outs a representation will be asked for. A single-penalty ridge probe on spectrum s is then generalised ridge on the latents with per-direction penalty λ/s_i, and it is the Bayes (Wiener) estimator exactly when $s \propto p$, the prior's variances (Theorem 1; Lindley & Smith 1972 read from the representation's side). The spectrum is the only channel through which a one-penalty probe can express a non-uniform prior. Isotropy is the uniform-prior case. It is correct precisely on LeJEPA's hypothesis and wrong as soon as the prior has any spread, by a cost that is a Jensen gap (Corollary 4; Figure 1). We make three contributions.

1. **An exact theorem and two new results** (§3). Theorem 1 holds for any n and d, including $n < d$, for any fixed design without a zero column, and for a prior diagonal in the representation's coordinates. *Theorem A* prices a spectrum in labels. Estimating the prior from k labels costs $(v\sigma^2/nk)\, N_{\text{eff}}$ in excess risk, counting only directions near the probe's noise floor, and $k_{\text{det}} = \tau n/\sigma^2$ labels, linear in the probe's own budget, suffice to detect every direction a probe can use. *Theorem B* shows that a single C-class task has a prior covariance of rank $C-1$. In its eigenbasis the prior is $C-1$ spikes, and in any other basis the diagonal is majorised by them (Schur–Horn), so the power-law body of a measured single-task prior is leakage from its head. The relation $\alpha^\star = \beta$ is therefore about task *families*, and the best single power law overshoots the fitted body slope (Proposition B3, under a unimodality condition that holds in every case we checked).

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.

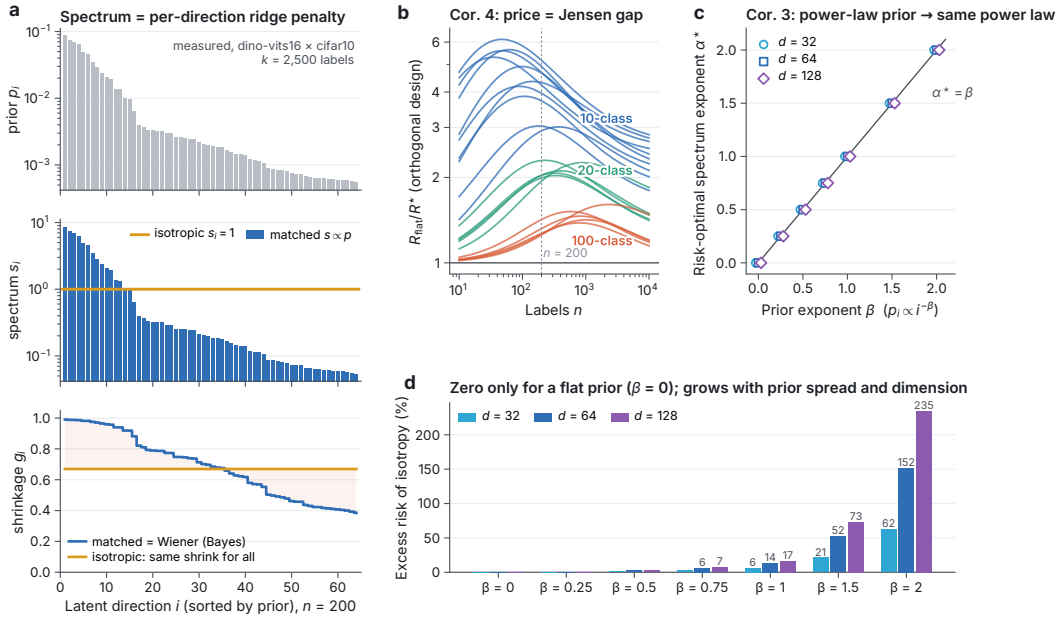


Figure 1: **The spectrum as a prior and the Jensen gap of isotropy.** (a) One measured prior (DINO ViT-S/16 × CIFAR-10, $k = 2500$ labels, top 64 whitened directions), the matched spectrum $s \propto p$ against the isotropic one, and the per-direction shrinkage g_i of a single-penalty ridge probe at $n = 200$ under the orthogonal design. On the matched spectrum it is the Wiener shrinkage $np_i/(np_i + \sigma^2)$ (Theorem 1), and isotropy shrinks every direction alike. (b) $R_{\text{flat}}/R^\star$ for the measured priors of all 20 cells against n (Corollary 4; colour = class count). (c) Risk-optimal power-law exponent against the prior's, $\alpha^\star = \beta$ (Corollary 3; numerical optimisation, $\sigma^2 = 0.25$). (d) Excess risk of isotropy, $R_{\text{flat}}/R^\star - 1$, which is zero only for a flat prior.

2. **The price of isotropy in labels across methods** (§4–§5). On 20 encoder×task cells, isotropising costs a median 21 accuracy points at $n = 200$. Recovering 90 % of the oracle prior's advantage takes a median 500 labels, and below about 100 labels a plug-in prior is *worse* than isotropy in 15 of 20 cells. The accuracy crossover is predicted by $k_{\text{head}} = \tau/p_{(C-1)}$, the cost of seeing the weakest of the $C - 1$ task directions. The predictor was found post hoc on the 20 cells and then replicated on 149 further cells of vision and text.
3. **Recovering the prior without labels and training toward it** (§6, §5.5). A prior averaged over *other* tasks in a shared whitened basis, applied without labels to a held-out task, recovers a median 89 % of what the task's own prior gives and beats the native spectrum in 74 % of 121 cells, and fails informatively on fine-grained variant tasks. Training a VICReg ResNet-18 toward such a population prior (the mean of the 20 measured priors of §4, five of them CIFAR-10 cells) instead of toward isotropy raises 200-label probe accuracy from 60 % to 80–81 % on all three seeds, and a label-free i^{-1} target does the same. The isotropy-trained encoder stays worse even with the oracle prior applied afterwards.

Scope of the claims. Theorem 1 is known mathematics with a new reading, while Theorems A and B are new. Every measurement on a trained encoder rescales its spectrum after the fact. Only §5.5 trains at controlled targets, on one architecture and one dataset. The supervised control (DeiT-S, ViT-B, ResNet-50) pays the same price as the self-supervised encoders. What isotropy buys in return, resistance to unsupervised alignment, is an open prediction (§7).

2 Related Work

Isotropy as a target. LeJEPA (Balestriero & LeCun, 2025) shows that the isotropic Gaussian uniquely minimises the integrated square bias of k-NN and kernel probes under a covariance budget (its Theorem 1), and that for linear probes any anisotropy admits a task on which it has higher bias

(its Lemmas 1–2). SIGReg enforces the target with an Epps–Pulley test on Cramér–Wold projections. SPHERE-JEPA (Nicollier et al., 2026) restates the same result as a minimax over $\|w\| \leq 1$, the uniform-task assumption made explicit. HamJEPA (Alvarez, 2026) minimises the worst-case task variance $\sup_{w^\top H^{-1} w \leq 1} w^\top \Sigma w$ under $\mathrm{tr}(H\Sigma) = c$ and obtains $\Sigma^\star = (c/d)H^{-1}$. For the same task ellipsoid our answer is $\Sigma \propto P = H$, reciprocal because their criterion equalises an energy while ours preserves a signal (§3.6). UR-JEPA (Le, 2026) reports that LeJEPA's projector spectrum is nearly flat (top-to-bottom ratio ≤ 3.6), and we measure 1.7 on the official recipe's projector, with a backbone exponent of 1.1–1.2 behind it (§5.3). Rectified LpJEPA (Kuang et al., 2026) moves the target toward sparsity, where §5.5 moves it toward a task prior. Klindt et al. (2026) prove Gaussian latents optimal for linear identifiability, the criterion under which isotropy wins. Betser et al. (2026) show that InfoNCE representations are asymptotically Gaussian in projection, and VISReg and Weak-SIGReg (Wu et al., 2026; Akbar, 2026) are covariance variants of SIGReg. None of these derives a target from a task prior.

Related mathematics and priors over tasks. The posterior mean under a Gaussian prior is generalised ridge with penalty $\sigma^2 P^{-1}$ (Lindley & Smith, 1972). Proposition 1b uses the deterministic equivalent of ridge risk of Dobriban & Wager (2018) and Hastie et al. (2022). Bordelon et al. (2020) and Canatar et al. (2021) analyse spectral bias and task–model alignment in kernel regression; we recover their $n^{1/\beta-1}$ rate from the representation's side. Han & Kim (2026) find the same prior-set diagonal regulariser rediscovered by learning in inverse problems. Empirical-Bayes plug-in costs of order (dimension)$/k$ are classical (Efron, 2010; Johnstone & Silverman, 2004), and Theorem A says which dimensions count and adds the linear-in-n detection cost. Neural collapse (Papyan et al., 2020) is the simplex geometry of Lemma 7, emerging there from training and imposed here by whitening under a stated hypothesis. Stringer et al. (2019) and α-ReQ (Agrawal et al., 2022) are the $\beta = 1$ case of Corollary 3. Task Priors (Patel & Balestriero, 2025) put a Gibbs prior over label graphs at the kernel level to evaluate a model over the space of tasks. Our family prior P_{fam} is the read-out-side counterpart, estimated from tasks, and yields the optimal spectrum. Dhinagar et al. (2026) and Gupta (2026) give sample thresholds for recoverable covariance modes and an effective rank that predicts when few-shot labels stop helping, the closest label-count statements to Theorem A. Njaradi et al. (2026) analyse pretrain-then-probe in proportional asymptotics. vec2vec and mini-vec2vec (Jha et al., 2025; Dar, 2025) align embedding spaces without pairs, and our alignment test (§7) uses the latter's quadratic-assignment centroid step. RAE (Zheng et al., 2025) finds that a diffusion denoiser's width must match the token dimension, and Proposition 9 refines this to the number of modes above the noise floor.

3 THEORY

Proofs are in Appendix A.

3.1 SETTING

Let $u \in \mathbb{R}^d$ have independent unit-variance coordinates. These are the latent directions an encoder could represent; allocating variance $s_i > 0$ to direction i gives $z = \mathrm{diag}(\sqrt{s})\, u$ under LeJEPA's budget $\sum_i s_i = d$. A downstream task is a linear functional of the *latents*, $y = w \cdot u + \varepsilon$ with $\varepsilon \sim (0, \sigma^2)$, and its weights carry a prior $w \sim (0, P)$ with $P = \mathrm{diag}(p)$, $p \succ 0$, diagonal in the representation's coordinates. Writing the task on u rather than on z is what separates this setting from LeJEPA's. The task is fixed before the encoder chooses what to keep, so shrinking a direction loses its signal, whereas a task family defined on z (LeJEPA's $\|w\| \leq 1$ on the embedding) is rescaled with it, which is why isotropy is optimal there. LeJEPA's worst case is the uniform prior. The probe is ridge regression with one penalty on n labelled points $Z \in \mathbb{R}^{n \times d}$, $\hat{\beta} = (Z^\top Z + \lambda I)^{-1} Z^\top y$, and its prediction risk on a fresh isotropic u is $\mathbb{E}\|\hat{w} - w\|^2$ with $\hat{w} = \mathrm{diag}(\sqrt{s})\hat{\beta}$.

3.2 THE SPECTRUM AS THE PROBE'S ONLY CHANNEL FOR A PRIOR

The theorem is Lindley & Smith (1972) read from the representation's side. The posterior mean under a Gaussian prior is generalised ridge with penalty $\sigma^2 P^{-1}$.

Theorem 1 (optimal allocation, exact)**.** *Fix $n \geq 1$, $\sigma^2 > 0$, $p \succ 0$, and let the design $U \in \mathbb{R}^{n \times d}$ (rows $u^\top$) be any fixed matrix with no identically-zero column (any rank, including $n < d$), or any*

random design that almost surely has no zero column (e.g. one with a density). Among single-penalty ridge probes on $z = \mathrm{diag}(\sqrt{s})\, u$ with $s \succ 0$, $\lambda > 0$, the Bayes risk $\mathbb{E}\|\hat{w} - w\|^2$ is minimised uniquely, up to the common scale of (s, λ), at $s \propto p$ with $\lambda = \sigma^2 s_i / p_i$ (the same constant for every i), and the minimum equals the Bayes risk of the posterior mean $\mathbb{E}[w \mid U, y]$.

Sketch. With $S = \mathrm{diag}(\sqrt{s})$ and $Z = US$, $\hat{w} = S(SU^\top US + \lambda I)^{-1} SU^\top y = (U^\top U + \lambda\, \mathrm{diag}(1/s))^{-1} U^\top y$, so a single-$\lambda$ ridge on spectrum s is generalised ridge on u with per-direction penalty λ/s_i, exactly. The posterior mean $(U^\top U + \sigma^2 P^{-1})^{-1} U^\top y$ is the unique minimiser of the Bayes risk, and equality with it forces column j of the design to vanish wherever $\lambda/s_j \neq \sigma^2/p_j$ (Appendix A.1). Only second moments are used, and the budget is inert because the risk depends on (s, λ) only through λ/s_i. □

Under an orthogonal design ($Z^\top Z = n\ \mathrm{diag}(s)$) the risk separates over directions and is minimised at $\lambda/s_i = \sigma^2/p_i$, with value $R^\star = \sum_i \sigma^2 p_i/(np_i + \sigma^2)$ (Proposition 1a). In the proportional regime the deterministic equivalent of ridge risk (Dobriban & Wager, 2018; Hastie et al., 2022) is again minimised, uniquely and globally, at $s \propto p$, with the noise inflated to $\sigma^2 + F^\star$ (Proposition 1b, Appendix A.2).

3.3 Power-law priors and the cost of isotropy

Corollary 3. *If $p_i \propto i^{-\beta}$ then $s_i \propto i^{-\beta}$: $\alpha^\star = \beta$; $\beta = 0$ is isotropy, $\beta = 1$ the $1/n$ spectrum of Stringer et al. (2019).*

Corollary 4 (isotropy is optimal iff the prior is uniform)**.** *Let $f(q) = \sigma^2 q/(nq + \sigma^2)$ and $\bar{p} = \frac{1}{d}\sum_i p_i$. Under the orthogonal design, $R^\star = d \cdot \mathrm{mean}_i f(p_i)$ and $R(\mathrm{flat}) = d\, f(\bar{p})$, so $R(\mathrm{flat}) - R^\star$ is the Jensen gap of the strictly concave f: non-negative, zero iff p is constant.*

The "iff" is exact by Theorem 1's uniqueness. With $p_i = Ai^{-\beta}$, $\beta > 1$, $R^\star = \Theta(n^{1/\beta - 1})$, the spectral-bias rate of Bordelon et al. (2020) and Canatar et al. (2021).

3.4 Task families and single tasks

A prior is a distribution over tasks, so it belongs to a *family* of read-outs, whose second moment $P_{\mathrm{fam}} = \mathbb{E}_{\mathrm{tasks}}[ww^\top]$ is what Theorem 1 matches when it is diagonal in the encoder's coordinates. Corollaries 3–4 are statements about that diagonal. Generically P_{fam} is not diagonal there, and then no single-λ diagonal spectrum attains the Bayes estimator. What survives is weaker, namely that $s \propto \mathrm{diag}(P_{\mathrm{fam}})$ is the optimal diagonal spectrum under the orthogonal-design risk, in which only the diagonal enters, and that is the prescription the experiments test (§4).

Theorem B1. *Let u be whitened, $y \in \{1, \ldots, C\}$ with $P(y = c) = \pi_c$, $Y_c = \mathbf{1}\{y = c\} - \pi_c$, $\mu_c = \mathbb{E}[u \mid y = c]$, $\bar{\mu} = \sum_c \pi_c \mu_c$. The prior covariance of the one-vs-rest family, $P_{\mathrm{full}} = \sum_c \mathrm{Cov}(u, Y_c)\mathrm{Cov}(u, Y_c)^\top = \sum_c \pi_c^2 (\mu_c - \bar{\mu})(\mu_c - \bar{\mu})^\top$, has rank at most $C - 1$, with equality iff the class means are affinely independent; its range is the between-class (Fisher) subspace.*

Corollary B2. *In any orthonormal basis the per-coordinate prior $p_i = (P_{\mathrm{full}})_{ii}$ is the energy of $C - 1$ directions projected onto coordinate i; in the eigenbasis it is $(\lambda_1, \ldots, \lambda_{C-1}, 0, \ldots, 0)$, and by Schur–Horn (Schur, 1923; Horn, 1954) the diagonal in any basis is majorised by that vector. So the "body" of a measured single-task prior beyond its top $C - 1$ coordinates is leakage of the head, not an independent tail.*

So $\alpha^\star = \beta$ concerns families with full-rank P_{fam} (what §6 estimates), not single tasks, whose head of $C - 1$ directions is steeper than any power law fitted through the body. Proposition B3 (Appendix A.4) says which way the best power law then errs. For a flat head $p_i = h$ on $i \leq m$ above the extrapolation of a body $t\, i^{-\beta}$, the body is under-penalised at $\alpha = \beta$ and, provided the first coordinate is not under-penalised, β is not optimal and, when the risk of $s_i = i^{-\alpha}$ is unimodal in α (as in every family we checked numerically), $\alpha^\star > \beta$ (mirror statement for a head below the extrapolation; for a pure power law the $n \to \infty$ limit is β). On the *measured* priors the risk prefers a single power law with exponent $1.26\hat{\beta} + 0.16$ at $n = 200$ ($1.45\hat{\beta} - 0.18$ at $n = 100$) rather than $\hat{\beta}$. The steeper-than-$\hat{\beta}$ optimum the experiments find (§4) is what head-heavy priors predict.

3.5 THEOREM A AND THE PRICE OF A SPECTRUM IN LABELS

A native spectrum is a prior the encoder learned for free. Isotropising discards it, and getting a prior back means estimating p from labels. We ask how many labels this takes. Let $g_i = np_i/(np_i + \sigma^2)$, $R^\star = (\sigma^2/n)\sum_i g_i$, $D(s\|p) = R(s) - R^\star \geq 0$ (orthogonal-design risk), and from k labels form $\hat{p}_i = p_i(1 + \eta_i)$ with $\mathbb{E}\eta_i = 0$, $\mathbb{E}\eta_i^2 = v_i/k$, $\eta_i = o(1)$, using $\hat{s} = \hat{p}$ with $\lambda = \sigma^2$ (tuning λ only lowers the risk).

Proposition A1 (estimation cost)**.** $\mathbb{E}[R(\hat{s}_k)] - R^\star = \frac{\sigma^2}{nk}\sum_i v_i\, g_i^2(1 - g_i) + o(1/k)$*; with a common* v, $\approx \frac{v\sigma^2}{nk} N_{\text{eff}}(p, n)$, $N_{\text{eff}} = \sum_i g_i^2(1 - g_i)$.

Since $g^2(1 - g) \leq 4/27$, attained at $np_i = 2\sigma^2$, only directions within a constant factor of the noise floor σ^2/n need their prior estimated (Corollary A2). Below that regime a direction is invisible. The label-variance estimator of §4 has $\mathbb{E}e_i = kp_i + V$ under the alternative and mean V with 99th-percentile excess τ under the permutation null, so direction i is distinguishable from the floor iff $kp_i \geq \tau$.

Proposition A3 (detection cost scales with the probe's own budget)**.** *Assume* $\mathbb{E}e_i = kp_i + V$ *with a null spread* τ *that does not depend on* k *(first order in the central limit theorem; Appendix A.5). The probe at budget* n *is sensitive to directions with* $g_i \geq \frac{1}{2}$*, i.e.* $p_i \geq \sigma^2/n$*. Detecting every such direction from labels is guaranteed once* $k \geq k_{\text{det}} := \tau n/\sigma^2$*, and this many are needed when the weakest usable direction sits at* $p_i = \sigma^2/n$*: the labels sufficient to* see *the part of the prior a probe with* n *labels can* use *grow linearly in* n*, with constant* τ/σ^2 *set by the label noise.*

Corollary A4 (Appendix A.5) combines the two regimes. After isotropising, the plug-in matches the native spectrum at $k^\star \approx \max\{k_{\text{det}},\, v\sigma^2 N_{\text{eff}}/(n\, D(s_{\text{nat}}\|p))\}$, so the better aligned the native spectrum, the more labels it was worth. k_{det} is a per-direction threshold that bounds the cost of recovering the *whole* usable prior. The data (§5.1) later showed that the accuracy crossover of a C-class task is governed instead by detecting its weakest head direction, $k_{\text{head}} = \tau/p_{(C-1)}$, where Theorem B1 meets the detection rule.

3.6 ALIGNMENT, DENOISING AND HAMJEPA

Alignment. An observer of unpaired samples from μ and its rotation $Q_\#\mu$ identifies $Q \in O(d)$ exactly up to the coset $Q \cdot \text{Stab}(\mu)$ of the stabiliser, which for $\mu = N(0, I_d)$ is all of $O(d)$ (Lemma 5, Corollary 6). SIGReg's fixed point is thus the maximally unalignable distribution, a privacy benefit Theorem 1 says is paid for in accuracy. Whitening alone already blunts the centroid-matching stage of practical aligners. Under H1–H3 (affinely independent centroids, whitening, within-cluster scatter isotropic on the centroid span *in the whitened coordinates*) an equal-weight mixture's centroids form an exact regular simplex and a centroid matcher faces $K!$ equally good configurations (Lemma 7, Corollary 8). H3 is strong, and with isotropic raw scatter the simplex is only approximate. **Denoising.** For $z \sim N(0, \text{diag}\, s)$ in noise σ^2, a flat spectrum needs $\lceil(1 - \epsilon)d\rceil$ modes to capture a $1 - \epsilon$ fraction of the Bayes denoiser's improvement, a power law with $\alpha > 1$ needs $O(1)$ (Proposition 9, Appendix A.7). **HamJEPA** (Alvarez, 2026) is the reciprocal case. Its task ellipsoid $\{w^\top H^{-1} w \leq 1\}$ is a level set of $N(0, H)$, its worst-case criterion gives $\Sigma^\star \propto H^{-1}$, and Theorem 1 gives $\Sigma \propto H$ (Appendix A.8).

4 MEASURING THE PRIOR

Estimator (Figure 2). We whiten the encoder's principal components on 5000 unlabelled training images and, on a disjoint 2500-image *prior split*, compute for each whitened component i the centred one-hot label variance it explains, $e_i = \sum_c (Y_c^\top Z_i)^2/N$. A 200-shuffle permutation null gives a floor c_0 (its median) and a detection window q_{99} (its 99th percentile). The estimated prior is $\hat{p}_i = \max(e_i - c_0, 0)$, and $\hat{\beta}$ is its log-log slope over the K directions above q_{99}, sorted by $\hat{p}$ (the theorem's coordinates). Probes draw n labels from a disjoint 2500-image *probe split*, fit scikit-learn's `RidgeClassifierCV` (Pedregosa et al., 2011) over 13 penalties, and are scored on 2000 held-out images. Every spectrum is a unit-mean variance multiplier on the whitened coordinates, and all share the same five probe seeds. The 20 primary cells pair DINO ViT-S/16 (Caron et al., 2021), DINOv2-S (Oquab et al., 2024), MAE-B (He et al., 2022), DeiT-S (the supervised control; Touvron

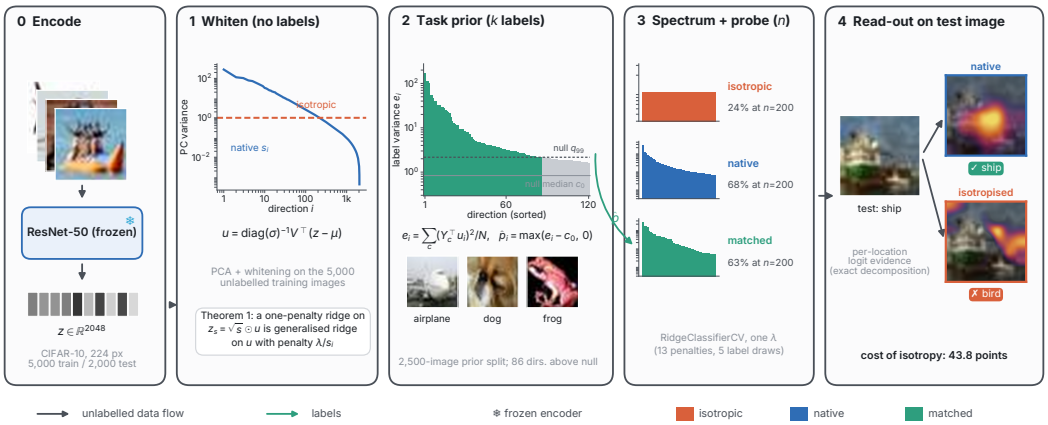


Figure 2: **The measurement pipeline on real data.** (0) A frozen encoder maps CIFAR-10 images to features z (here a supervised ResNet-50, $d = 2048$). (1) Whitening on unlabelled images gives coordinates u. The native spectrum is the PC variances, and isotropy sets them all to 1. (2) On the prior split, the label variance e_i of each whitened direction against a 200-shuffle permutation null (c_0, q_{99}) gives $\hat{p}$. (3) A spectrum s (isotropic, native or matched $s \propto \hat{p}$) rescales u, and a one-penalty ridge probe is fitted on n labels. By Theorem 1 this is generalised ridge on u with penalty λ / s_i. (4) On a test image, the per-location evidence for the true class (exact decomposition of the probe logit) with the native and the isotropised spectrum. Accuracies at $n = 200$ and the cost of isotropy are from this encoder (Appendix G).

et al., 2021) and a third-party LeJEPA ViT-S/16 reproduction with CIFAR-10, CIFAR-100 fine and coarse (Krizhevsky, 2009), and STL-10 (Coates et al., 2011), at 5000/2000 images and 224 px. Re-extracting the features on a second device reproduces every accuracy of the original run to within 0.004, and the best-α argmax flips in two cells on one-prediction ties (Appendix B).

The direct test. Setting $s \propto \hat{p}$, the orthogonal-design prescription of §3.4 with no free parameter, lands within one standard deviation of the best swept power law (α up to 4, either ordering) in **16/20** cells at $n = 100$ and **14/20** at $n = 200$. The matched spectrum beats the best power law in all five 100-class cells and trails it by at most 1.3 points in the ten 10-class cells (Table 3, Figure 4).

Head-heaviness (Theorem B1 in the data). A power law fits the measured priors in the theorem's coordinates ($R^2 \geq 0.84$ in every cell), yet the accuracy-optimal exponent runs steeper, with $\alpha^\star = 1.40\hat{\beta} - 0.05$ ($r = 0.79$) at $n = 200$ and $2.47\hat{\beta} - 0.97$ ($r = 0.89$) at $n = 100$. The theorem applied to each cell's measured prior predicts $1.26\hat{\beta} + 0.16$ and $1.45\hat{\beta} - 0.18$, and tracks the measured optimum cell by cell at $r = 0.86$–0.88. Scoring the same probes by one-hot squared error instead of accuracy moves the $n = 100$ slope from 2.47 to 1.83 ($r = 0.97$ against the theory's per-cell prediction; Figure 10) and the $n = 200$ slope to 1.25 against the predicted 1.26. The $n = 100$ residual stays open.

The cost of isotropy. Isotropising costs a median 21 accuracy points at $n = 200$ over the 20 cells, range 4–42 (20–42 on the 10-class tasks, 17–22 on the 20-class task, and 4–9 on CIFAR-100 fine, where every spectrum is poor). The supervised DeiT-S pays 5–26. Over the 169 cells of §5.4 the median is again 21 points and the range 0.5–64 (Figure 5). Appendix G shows on real images what the discarded spectrum carried.

5 The Price in Labels

5.1 Recovering a Discarded Prior

We isotropise each of the 20 cells, estimate the prior from $k \in \{10, \ldots, 2500\}$ prior-split labels (five draws per k; Appendix B), and probe at $n \in \{100, 200, 500\}$. We call k_{90} the smallest k at which the plug-in prior recovers 90 % of the oracle prior's advantage over isotropy, and $k^\star_{\mathrm{nat}}$ the smallest k

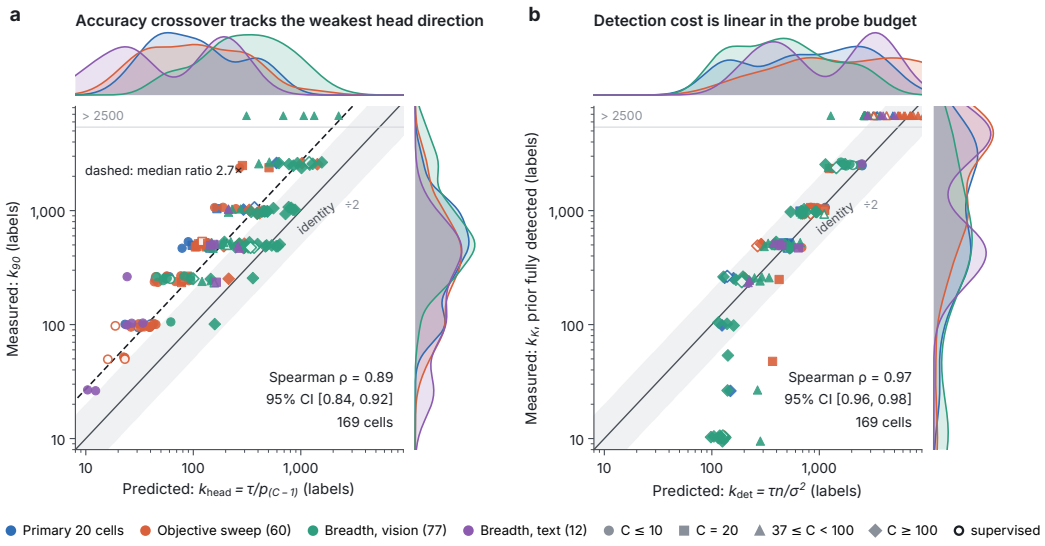


Figure 3: **Theorem A on 169 cells.** (a) k_{90} against $k_{\text{head}} = \tau/p_{(C-1)}$ ($\rho = 0.89$); (b) k_K against $k_{\text{det}} = \tau n/\sigma^2$ ($\rho = 0.97$); bootstrap CIs in brackets and in §5.4. Colour = cell set, marker = class count, hollow = supervised; identity solid, factor-2 band shaded, dashed in (a) the median ratio; crossings beyond the grid drawn at the top (> 2500); margins: densities per cell set.

at which it matches the encoder's own spectrum (Figure 6; Table 4). Undefined crossings enter the medians as 5000, so the medians are lower bounds (bootstrap CIs in brackets). At $n = 200$ the median k_{90} is **500** $[250, 750]$, and the median $k^{\star}_{\text{nat}}$ is 500 $[250, 500]$. The median k_{90} is 375 at $n = 100$ and 1000 at $n = 500$, so the price rises with the probe's budget. In 15 of 20 cells the plug-in prior is worse than isotropy at some $k \leq 100$. Proposition A1's median measured-to-predicted excess is 1.41, 1.31 and 0.97 at $k = 250$, 500 and 1000. Proposition A3 gives Spearman$(k_{90}, k_{\text{det}}) = -0.89$.

Reading. The price grows with n, as Proposition A3 predicts. The estimation formula has the right scale but depends on the class count. Per-cell ratios span 0.13–15, with 0.4–0.5 on the 10-class tasks (Proposition A1's expansion needs $k \gg v$ and fails on Corollary B2's leakage directions) and 2.9–3.6 on the 100-class tasks (the expansion omits the hard detection threshold). The detection test compared a per-direction threshold, $k_{\text{det}} \propto n/\sigma^2$, which *shrinks* with the class count because the summed residual variance σ^2 grows, with a whole-prior recovery time, which grows with C. The negative sign is the class count showing through both. Tested on what Proposition A3 says, it holds. The label count k_K at which the detected-direction count reaches the probe-usable count $\#\{g_i \geq \frac{1}{2}\}$ tracks k_{det} at $\rho = 0.97$ (19/20 cells within a factor 2). The accuracy crossover is predicted instead by $k_{\text{head}} = \tau/p_{(C-1)}$, the detection cost of the weakest of Theorem B1's $C - 1$ directions, at $\rho = 0.97$ with $k_{90} \approx 4k_{\text{head}}$ (range 2.2–6.4). Both predictors were post hoc on these 20 cells, and we then evaluated them on the cells below.

5.2 Steep reconstruction and flat invariance objectives

Fifteen further encoders, each given an isotropy-pressure rank (0 reconstruction, 1 distillation/clustering, 2 contrastive, 3 decorrelation, 4 SIGReg; Table 6), ran on the four tasks. Reconstruction objectives are steep (α_{nat} 1.33–2.00) and every invariance-based objective is flat (0.61–0.99), the supervised encoders among them (Figure 7). The **Theorem-A replication** also holds on the 60 fresh cells, with Spearman$(k_{\text{det}}, k_K) = +0.92$ and Spearman$(k_{\text{head}}, k_{90}) = +0.94$, median $k_{90}/k_{\text{head}} = 3.1$ (per encoder, Table 6). The finer hypothesis, that α_{nat} and the median price order with the pressure rank within both the ViT-B/16 and the RN50 family, gives Spearman correlations (ViT-B/16, $n = 8$: $\rho = -0.37$ and -0.56; RN50, $n = 5$: -0.63 and undefined, every encoder at 375). The price measure is confounded with the oracle gap, as Appendix C explains.

Table 1: Results by cell set, $n = 200$. Calibration: accuracy-optimal exponent on $\hat{\beta}$; direct test: matched within 1 sd of the best power law; Thm. A: $\rho(k_{\mathrm{det}}, k_K)$ / $\rho(k_{\mathrm{head}}, k_{90})$.

cell set	n	calibration $\alpha^\star = b\hat{\beta} + a$ (r)	direct test	median k_{90}	Thm. A
primary	20	$1.40\hat{\beta} - 0.05$ (0.79)	70 %	500	0.97 / 0.97
objective sweep	60	$1.50\hat{\beta} + 0.10$ (0.83)	40 %	375	0.92 / 0.94
breadth, vision	77	$1.20\hat{\beta} + 0.35$ (0.53)	**83 %**	500	0.97 / 0.83
breadth, text	12	$0.85\hat{\beta} + 0.63$ (0.66)	58 %	250	0.94 / 0.86
all	**169**	$1.26\hat{\beta} + 0.28$ (0.72)	64.5 %	500	**0.97 / 0.89**

5.3 Measuring the official LeJEPA recipe where SIGReg acts

The 20-cell LeJEPA measurement probes a third-party checkpoint at the backbone, while SIGReg acts on the projector. The official repository ships no checkpoint, so we trained its documented recipe unchanged (ViT-S/8, 16-d projector under SIGReg, 800 epochs; Appendix B) on CIFAR-10 at 64 px (kNN 90.9 %) and probed three layers (Table 8). SIGReg works as designed. The 16-d projector is isotropic (top-to-bottom eigenvalue ratio 1.7, against UR-JEPA's reported ≤ 3.6) and has nothing to recover, because the whole task fits in it. The backbone behind it is as steep as DINO's (α_{nat} 1.09 at the 512-d head, 1.21 at the CLS token), and at the head, which is what a user probes, the price is 0.28 at 200 labels with $k_{90} = 100$. The third-party backbone (cost 0.17–0.26) was representative.

5.4 Breadth over 169 vision and text cells

We add seven vision datasets $\times$ 11 encoders and four text tasks $\times$ three sentence encoders (Appendix B) through the same machinery, and apply the calibration, the direct test and the label-price headline per cell set and pooled.

The defensible statement stays "monotone, slope above 1, explained by head-heavy priors" (§4). The direct test is weak on the RN50-heavy objective sweep, where 2048-dimensional matched spectra lose to power laws. The Theorem-A replication holds in every set, vision and text alike, with pooled $\rho(k_{\mathrm{det}}, k_K) = 0.97$ (bootstrap 95 % CI $[0.96, 0.98]$) and $\rho(k_{\mathrm{head}}, k_{90}) = 0.89$ (CI $[0.84, 0.92]$), with $k_{90} \approx 2.4 k_{\mathrm{head}}$ on the breadth cells (Figure 3). Per-cell results are in Table 10.

5.5 Training toward a prior instead of isotropy

Everything above rescales encoders that were already trained. Here we train VICReg ResNet-18 on CIFAR-10 for 200 epochs with an added spectral regulariser, $\mu \cdot \mathrm{mean}_i(\lambda_i/\bar{\lambda} - s_i/\bar{s})^2$ on the eigenvalues of the 512-d backbone covariance, toward an isotropic, an i^{-1}, or a population-prior target in the sense of §6 (the mean of the 20 measured priors of §4, resampled to 512 ranks). A pilot rule set $\mu = 1$ for every arm, with three seeds each.

Population minus isotropic is $+0.208$, $+0.209$ and $+0.199$ on the three seeds. The native-spectrum probe reaches 0.60–0.61 on the isotropic arm, 0.80 on the i^{-1} arm and 0.80–0.81 on the population arm (kNN 0.83–0.84 against 0.86–0.87; spectra within 0.07–0.28 of target, 0.64–0.69 for the isotropic arm; Figure 9, Table 9). First, the isotropic arm went flat on about 470 of 512 directions and collapsed the rest, and its probe with the oracle prior applied afterwards reaches only 0.65–0.67 against 0.80 for the other arms, so the target destroyed information, not only its scaling. Second, the power-law and population targets are indistinguishable, which matters for the population arm's provenance, because its target came from the 20 cells of §4, five of them CIFAR-10 cells, so 12 500 CIFAR-10 prior-split labels shaped a target then used to train a CIFAR-10 encoder. The target is a sorted, basis-free vector averaged over 20 tasks, and the label-free i^{-1} target gains the same $+0.18$ to $+0.20$. That a non-flat target beats isotropy holds without qualification, while "the population prior" carries this caveat until a replication builds the target from non-CIFAR-10 cells.

6 RECOVERING THE PRIOR WITHOUT LABELS

An objective could instead target a family prior estimated from *other* tasks. For each of 11 encoders, we fit one shared whitened basis on 2000 unlabelled images from each of 11 vision datasets (the four above plus the seven of §5.4), measure every task's prior in that basis, and form the population prior $\bar{p}_{-t}$ as the mean of the other tasks' normalised priors. We then apply $\bar{p}_{-t}$, with no label from task t, to that task at $n = 200$ and compare it with the isotropic, native and oracle (own-prior) spectra in 121 cells. The population prior recovers at least half of the oracle advantage in **63.6 %** of cells (bootstrap CI $[0.55, 0.72]$) and beats the native spectrum in **74.4 %** (CI $[0.67, 0.82]$), with a median recovery of 0.89 over the 93 cells whose oracle prior beats isotropy (Figure 8). Leaving the task out of the average costs nothing ($+0.0008$ from leaving it in), so what transfers is a family prior in the sense of §3.4, not leakage. The family has a boundary. The median recovery is 0.87–0.99 on each of the seven natural-category tasks and 0.72 on CIFAR-100 fine, but 0.64, 0.32 and -2.3 on Food-101, Flowers-102 and FGVC-Aircraft. Fine-grained variant recognition lives in different directions, outside the family. On Aircraft even the task's own prior, from 2500 labels of 100 classes, is worse than isotropy for 8 of the 11 encoders (28 of the 121 cells have oracle $\leq$ isotropic), the "bad prior" regime of §5.1 entered by class count. CLIP transfers best (91 % of its cells recover) and data2vec worst (27 %; Table 7).

7 ISOTROPY AND UNPAIRED ALIGNMENT

Corollary 6 says an isotropic Gaussian embedding is the one distribution an unpaired aligner can learn nothing from, and Lemma 7 predicts a two-stage collapse of alignment, whitening then Gaussianisation. This is theory plus synthetic evidence and two real-data tests, so it is discussion, not a contribution. We aligned three pairs of sentence encoders on 20 000 unpaired sentences with mini-vec2vec's pipeline (quadratic-assignment centroid matching, Procrustes, nearest-neighbour refinement) in six conditions (raw, whitened, one to eight Gaussianisation rounds), with the Epps–Pulley statistic (0.52–1.64) as the Gaussianity axis. Raw top-10 retrieval is low, as in an earlier attempt (0.122, 0.047, 0.001), while the supervised Procrustes ceiling is 0.86–0.97 on every pair. The spaces are alignable and the unsupervised matcher is what falls short, so the alignment prediction is untested rather than refuted. One transition is only suggestive. From raw to whitened, top-10 retrieval falls from 0.122 to 0.018 and from 0.047 to 0.008 on two pairs while the ceiling moves by at most 0.04, which is the first-stage signature the lemma predicts.

8 LIMITATIONS

1. **Scope.** Every measurement on a trained encoder is post hoc rescaling. Only §5.5 trains at controlled targets (one architecture, one dataset, 200 epochs). We use linear probes only, and the nonlinear analogue is open, because for a kernel probe the eigenbasis moves with the covariance. We use 5000/2000 images per dataset at one resolution. LeJEPA is measured on the official recipe at CIFAR-10, 64 px, not ImageNet scale, and in the 20 cells through a third-party checkpoint whose backbone spectrum the official run matches. The alignment prediction is untested, because both attempts were limited by the aligner. The supervised control pays the same price, so the price is a property of regularised linear probing, not of self-supervision.
2. **Open results.** The direct test at $n = 200$ is 15/20 in one run and 14/20 in another (16/20 at $n = 100$ in both). k_{head} was post hoc before it was replicated on 149 cells. The estimation formula's per-cell ratios span 0.13–15, and the $n = 100$ calibration residual is open. The pressure hypothesis's price measure is confounded with the oracle gap.
3. **Hypotheses.** Theorem 1 assumes a prior diagonal in the representation's coordinates. A real task's P_{full} is not, so the prescription we test, $s \propto \mathrm{diag}(\hat{p})$, is the optimal diagonal spectrum under the orthogonal-design risk, not the exact Bayes statement. Lemma 7 is exact only under whitened isotropy on the centroid span, and with isotropic raw scatter the simplex is approximate. Proposition A3 assumes a null spread independent of k, true on real features to 5 % once $k \geq 10C$, so k_{det} and k_{head} understate the cost at a few labels per class. Proposition B3's sign conclusions assume the risk is unimodal in α, true in every case checked but not proved.

AI USE STATEMENT

In this work we used a generative AI tool, an LLM-based coding agent, for the following tasks: to help develop the theoretical model and conceptual framework (the spectrum-as-prior reading and the setting of §3.1), to formulate mathematical claims (the statements of Theorems 1, A and B and their corollaries), to provide ingredients for proofs (Appendix A), to help design some of the experiments, and to write the code for the experiments, analyses and figures. We also used it to polish the writing of the paper and to help find related work. We did not use generative AI tools for any other purpose in this work. We have reviewed all AI-assisted work. Every theorem, proposition and lemma was re-derived line by line and checked numerically, and every claim, proof step and number in the paper was checked by the authors. We take responsibility for the final content of this work. Appendix E gives the disclosure in full.

ETHICS STATEMENT

This work uses only public datasets under their published terms and involves no human subjects. Two points deserve mention. First, the alignment consequence of §7 has a privacy reading. An isotropic Gaussian embedding is the one distribution an unpaired aligner can learn nothing from, so the isotropy that costs accuracy in §5 may buy resistance to embedding-inversion attacks of the vec2vec kind. The paper offers this as a prediction and provides no new attack. Our alignment test re-implements a published unsupervised aligner on public sentence encoders. Second, as a matter of research integrity, AI assistance was used in developing the theory, in designing some experiments and in writing the code (disclosed above and in Appendix E).

REPRODUCIBILITY STATEMENT

An anonymised repository, to be shared with reviewers through OpenReview when the discussion period opens, contains every script (feature extraction, the estimator and probes, every experiment, the theory numerics and the verification), the raw result files and per-cell tables (Appendix C), and the verification scripts and their outputs (Appendix D). Appendix B gives the estimator, splits, probes, encoders (with the pressure ranks), datasets and the training recipes of §5.3 and §5.5. Feature caches are regenerated by one command per experiment. Every number in the paper is traceable to a named result file and key (the repository's `NUMBERS.md`).

REFERENCES

Kumar Krishna Agrawal, Arnab Kumar Mondal, Arna Ghosh, and Blake Richards. α-ReQ: Assessing representation quality in self-supervised learning by measuring eigenspectrum decay. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17626–17638, 2022.

Habibullah Akbar. Weak-SIGReg: Covariance regularization for stable deep learning. *arXiv preprint arXiv:2603.05924*, 2026.

Robert Jenkinson Alvarez. Beyond isotropy in JEPAs: Hamiltonian geometry and symplectic prediction. *arXiv preprint arXiv:2605.20107*, 2026.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, 2022.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.

Roy Betser, Eyal Gofer, Meir Yossef Levi, and Guy Gilboa. InfoNCE induces Gaussian distribution. *arXiv preprint arXiv:2602.24012*, 2026. ICLR 2026.

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1024–1034, 2020.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature Communications*, 12:2914, 2021.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision*, 2021.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. In *Workshop on Natural Language Processing for Conversational AI*, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 2020a.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision*, 2021.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *International Conference on Artificial Intelligence and Statistics*, 2011.

Guy Dar. mini-vec2vec: Scaling universal geometry alignment with linear transformations. *arXiv preprint arXiv:2510.02348*, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

Nikhil J. Dhinagar, Vidhi Chhatbar, Chirag Jagad, Pavithra Senthilkumar, Sophia I. Thomopoulos, Mahir H. Khan, Sook-Lei Liew, Paul M. Thompson, et al. Anchoring the eigengap: Cross-modal spectral stabilization for sample-efficient representation learning. *arXiv preprint arXiv:2605.08764*, 2026.

Werner Dinkelbach. On nonlinear fractional programming. *Management Science*, 13(7):492–498, 1967.

Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: ridge regression and classification. *Annals of Statistics*, 46(1):247–279, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Bradley Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press, 2010.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. In *CVPR Workshop on Generative-Model Based Vision*, 2004.

Arnav Gupta. A spectral phase diagram for binary few-shot classification: Intrinsic dimensionality, geometric saturation, and representational diagnosis. *arXiv preprint arXiv:2606.24903*, 2026.

Jeahn Han and Pyojin Kim. Why learning rediscovers the closed-form diagonal regularizer. *arXiv preprint arXiv:2609.09656*, 2026.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of Statistics*, 50(2):949–986, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Alfred Horn. Doubly stochastic matrices and the diagonal of a rotation matrix. *American Journal of Mathematics*, 76:620–630, 1954.

Rishi Jha, Collin Zhang, Vitaly Shmatikov, and John X. Morris. Harnessing the universal geometry of embeddings. *arXiv preprint arXiv:2505.12540*, 2025.

Iain M. Johnstone. *Gaussian Estimation: Sequence and Wavelet Models*. Book draft, Stanford University (version of August 9, 2017), 2017.

Iain M. Johnstone and Bernard W. Silverman. Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *Annals of Statistics*, 32(4):1594–1649, 2004.

David Klindt, Yann LeCun, and Randall Balestriero. When does LeJEPA learn a world model? *arXiv preprint arXiv:2605.26379*, 2026.

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

Yilun Kuang, Yash Dagade, Tim G. J. Rudner, Randall Balestriero, and Yann LeCun. Rectified Lp-JEPA: Joint-embedding predictive architectures with sparse and maximum-entropy representations. *arXiv preprint arXiv:2602.01456*, 2026.

Ken Lang. NewsWeeder: Learning to filter netnews. In *International Conference on Machine Learning*, 1995.

Triet M. Le. UR-JEPA: Uniform rectifiability as a regularizer for joint-embedding predictive architectures. *arXiv preprint arXiv:2606.01443*, 2026.

Dennis V. Lindley and Adrian F. M. Smith. Bayes estimates for the linear model. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(1):1–18, 1972.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Léo Nicollier, Max Dunitz, Marc Pic, Pablo Musé, Enric Meinhardt-Llopis, and Gabriele Facciolo. SPHERE-JEPA: Spherical prediction with homogeneous embeddings. *arXiv preprint arXiv:2605.26900*, 2026.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008.

Valentina Njaradi, Rachel Swanson, Clémentine Dominé, Marco Mondelli, and Andrew Saxe. Optimal representation size: High-dimensional analysis of pretraining and linear probing. *arXiv preprint arXiv:2605.20105*, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

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.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 2019.

Niket Patel and Randall Balestriero. Task priors: Enhancing model evaluation by considering the entire space of downstream tasks. *arXiv preprint arXiv:2507.09871*, 2025.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Mark S. Pinsker. Optimal filtering of square-integrable signals in Gaussian noise. *Problems of Information Transmission*, 16(2):120–133, 1980.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Conference on Empirical Methods in Natural Language Processing*, 2019.

Issai Schur. Über eine Klasse von Mittelbildungen mit Anwendungen auf die Determinantentheorie. *Sitzungsberichte der Berliner Mathematischen Gesellschaft*, 22:9–20, 1923.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Conference on Empirical Methods in Natural Language Processing*, 2013.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, 2020.

Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Matteo Carandini, and Kenneth D. Harris. High-dimensional geometry of population responses in visual cortex. *Nature*, 571:361–365, 2019. doi:10.1038/s41586-019-1346-5.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, 2021.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, 2020.

Ross Wightman. PyTorch image models. `https://github.com/huggingface/pytorch-image-models`, 2019.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020.

Haiyu Wu, Randall Balestriero, and Morgan Levine. VISReg: Variance-invariance-sketching regularization for JEPA training. *arXiv preprint arXiv:2606.02572*, 2026.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-Pack: Packaged resources to advance general Chinese embedding. *arXiv preprint arXiv:2309.07597*, 2023.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, 2021.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, 2015.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025.

A PROOFS

Statements are restated with the hypotheses their proofs use (Appendix D lists how each was checked). Notation as in §3; $\varphi_i(\rho) = (\rho^2 p_i + n\sigma^2)/(n+\rho)^2$, $g_i = np_i/(np_i+\sigma^2)$, $f(q) = \sigma^2 q/(nq+\sigma^2)$.

A.1 THEOREM 1

Proof. Write $S = \operatorname{diag}(\sqrt{s})$ and $Z = US$. The ridge probe read back in latent coordinates is

$$\begin{aligned}\hat{w}_{s,\lambda} &= S(SU^\top US + \lambda I)^{-1} SU^\top y = S\big[S(U^\top U + \lambda S^{-2})S\big]^{-1} SU^\top y \\ &= (U^\top U + \Lambda)^{-1} U^\top y, \qquad \Lambda := \lambda \operatorname{diag}(1/s),\end{aligned}$$

using only $s_i > 0$: a single-λ ridge on spectrum s is generalised ridge on u with the diagonal penalty $\Lambda_i = \lambda/s_i$. Under $w \sim N(0, P)$ and $\varepsilon \sim N(0, \sigma^2 I)$ the posterior of w given (U, y) is Gaussian with mean $\hat{w}_B = (U^\top U + \sigma^2 P^{-1})^{-1} U^\top y$ (complete the square in $\|y - Uw\|^2/\sigma^2 + w^\top P^{-1} w$), and for every fixed U the posterior mean is the almost-surely unique minimiser of $\mathbb{E}[\|\hat{w} - w\|^2 \mid U]$ over all measurable estimators, because squared loss is strictly convex. Prediction risk on a fresh $u \sim (0, I)$ independent of everything is $\mathbb{E}[(u^\top(\hat{w} - w))^2] = \mathbb{E}\|\hat{w} - w\|^2$ since $\mathbb{E}uu^\top = I$. Hence $R(s, \lambda \mid U) \geq R_B(U)$ for every U and, averaging over a random design, $R(s, \lambda) \geq R_B$.

Equality holds iff $\hat{w}_{s,\lambda} = \hat{w}_B$ almost surely. Both are linear in y, and $y = Uw + \varepsilon$ has covariance $UPU^\top + \sigma^2 I \succ 0$, so $(M - M_B)y = 0$ a.s. forces $M = M_B$ (the second-moment version of the full-support argument). With $A = U^\top U + \Lambda$ and $B = U^\top U + \sigma^2 P^{-1}$, both positive definite, $A^{-1}U^\top = B^{-1}U^\top$ is $(B - A)B^{-1}U^\top = 0$, i.e.

$$(\sigma^2 P^{-1} - \Lambda)\, B^{-1} U^\top = 0. \tag{1}$$

Case $\operatorname{rank} U = d$. $B^{-1}U^\top$ has rank d, so it has no zero row, and (1) forces every diagonal entry of $D := \sigma^2 P^{-1} - \Lambda$ to vanish: $\lambda/s_i = \sigma^2/p_i$ for all i, i.e. $s \propto p$.

Case $\operatorname{rank} U < d$ *(in particular* $n < d$*), any fixed design with no zero column.* Suppose $D_{jj} \neq 0$. Then (1) forces row j of $B^{-1}U^\top$ to vanish: $e_j^\top B^{-1} U^\top = 0$, i.e. $Uv = 0$ for $v := B^{-1} e_j$ (B is

symmetric). But $Bv = e_j$ reads $U^\top Uv + \sigma^2 P^{-1} v = e_j$, and with $Uv = 0$ this is $\sigma^2 P^{-1} v = e_j$, so $v = (p_j/\sigma^2)e_j$; then $Uv = 0$ says $Ue_j = 0$: column j of the design is identically zero. Contrapositive: for every fixed design with no zero column, $D = 0$ and $s \propto p$. Conversely, if $Ue_j = 0$ then $(U^\top U)_{j\cdot} = 0$, B is block-diagonal in coordinate j, row j of $B^{-1}U^\top$ is $(p_j/\sigma^2)(U^\top)_{j\cdot} = 0$ and Λ_{jj} is free: the probe never sees direction j, $\hat{w}_j \equiv 0$ whatever s_j, and the tie is exactly there. A random design with a density has no zero column almost surely, which recovers the genericity argument as a special case.

Uniqueness is of the shape: $(s, \lambda) \mapsto (cs, c\lambda)$ leaves Λ unchanged, so the minimiser is the ray $s \propto p$ with $\lambda = \sigma^2 s_i/p_i$, and the budget $\sum_i s_i = d$ picks the point $s_i = dp_i/\sum_j p_j$, $\lambda^\star = d\sigma^2/\sum_j p_j$. If w, ε are not Gaussian, $\hat{w}_B$ is still the linear minimum-mean-square-error (Wiener) estimator, the unique minimiser of $\mathbb{E}\|My - w\|^2$ over linear M given $\mathbb{E}[ww^\top] = P$, $\mathbb{E}[\varepsilon\varepsilon^\top] = \sigma^2 I$, $w \perp \varepsilon$; ridge probes are linear in y and the equality step used only $\mathbb{E}[yy^\top] \succ 0$, so the theorem holds under second-moment assumptions alone. □

The zero-column argument covers every fixed design, including rank-deficient ones with $n < d$: spectra that differ only on a zero column of the design tie, and no others do. The statement was checked numerically on rank-12 duplicate-row designs at $n = 24 < d = 48$ (Appendix D).

A.2 Propositions 1a and 1b: the value of the risk

Proposition 1a (value, orthogonal design)**.** *If $Z^\top Z = n\,\mathrm{diag}(s)$ exactly, $R(s, \lambda) = \sum_i \varphi_i(\lambda/s_i)$ with $\varphi_i(\rho) = (\rho^2 p_i + n\sigma^2)/(n + \rho)^2$, minimised at $\rho_i = \sigma^2/p_i$ with $R^\star = \sum_i \sigma^2 p_i/(np_i + \sigma^2)$.*

Proposition 1b (value, proportional regime)**.** *Let $n, d \to \infty$ with $d/n \to \gamma$ and U with i.i.d. standard entries. With $\mu = \lambda/n$ and $\kappa \geq 0$ solving $\mu = \kappa(1 - \frac{1}{n}\sum_i s_i/(s_i + \kappa))$, the Bayes prediction risk converges to $R_\infty = \big[\kappa^2 \sum_i p_i/(s_i + \kappa)^2 + \frac{\sigma^2}{n}\sum_i s_i^2/(s_i + \kappa)^2\big]\big/\big(1 - \frac{1}{n}\sum_i s_i^2/(s_i + \kappa)^2\big)$. In the gains $g_i = s_i/(s_i + \kappa)$ its unique global minimiser over admissible g is $g_i^\star = p_i/(p_i + (\sigma^2 + F^\star)/n)$, i.e. $s \propto p$, where $F^\star$ is the unique positive root of $F^\star = \sum_i (\sigma^2 + F^\star)p_i/(np_i + \sigma^2 + F^\star)$.*

Proof of Proposition 1a. Under $Z^\top Z = n\,\mathrm{diag}(s)$, $\hat{\beta}_i = (ns_i\beta_i^* + (Z^\top\varepsilon)_i)/(ns_i + \lambda)$ with $\beta_i^* = w_i/\sqrt{s_i}$, $\mathbb{E}(Z^\top\varepsilon)_i = 0$ and $\mathrm{Var}(Z^\top\varepsilon)_i = ns_i\sigma^2$. Bias $-\lambda\beta_i^*/(ns_i + \lambda)$, variance $ns_i\sigma^2/(ns_i + \lambda)^2$; the risk is $\sum_i s_i \mathbb{E}(\hat{\beta}_i - \beta_i^*)^2$ because $\hat{w}_i - w_i = \sqrt{s_i}(\hat{\beta}_i - \beta_i^*)$, and $\mathbb{E}_w[s_i\beta_i^{*2}] = \mathbb{E}w_i^2 = p_i$. Substituting $ns_i + \lambda = s_i(n + \rho_i)$ with $\rho_i = \lambda/s_i$ cancels every s_i and gives $R = \sum_i \varphi_i(\rho_i)$. $\varphi_i'(\rho) = 2n(\rho p_i - \sigma^2)/(n + \rho)^3$ is negative below and positive above $\rho = \sigma^2/p_i$, a unique minimum with value $\sigma^2 p_i/(np_i + \sigma^2)$. For a design with $U^\top U = nI$ (which needs $n \geq d$) this coincides with the exact value $\mathrm{tr}((nI/\sigma^2 + P^{-1})^{-1})$; for $n < d$ Proposition 1a is a closed-form heuristic and the exact value is $R_B(U)$ or the proportional-regime value. □

Proof of Proposition 1b. The deterministic equivalent is the standard one for ridge on a design with covariance $\mathrm{diag}(s)$, prediction risk in the $\mathrm{diag}(s)$-norm, with the bias term averaged over $\mathbb{E}[\beta^*\beta^{*\top}] = \mathrm{diag}(p/s)$ (Hastie et al., 2022, Theorem 5): bias numerator $\kappa^2 \sum_i s_i(p_i/s_i)/(s_i + \kappa)^2 = \kappa^2 \sum_i p_i/(s_i + \kappa)^2$, variance $(\sigma^2/n)\sum_i s_i^2/(s_i + \kappa)^2$, common denominator $1 - \frac{1}{n}\sum_i s_i^2/(s_i + \kappa)^2$; $\mu \mapsto \kappa$ is a bijection onto $[\kappa_0, \infty)$ since $\frac{d}{d\kappa}[\kappa(1 - \mathrm{df}_1(\kappa)/n)] \geq 0$. In the gains $g_i = s_i/(s_i + \kappa)$: $(1 - g_i)^2 = \kappa^2/(s_i + \kappa)^2$, $\mathrm{df}_2 = \sum_i g_i^2$, admissibility ($\mu \geq 0$) is $\sum_i g_i \leq n$, and every $g \in (0, 1)^d$ with $\sum_i g_i \leq n$ is realised by some (s, κ), so the problem is exactly

$$\min_g \; R_\infty(g) = \frac{N(g)}{D(g)}, \qquad N = \sum_i p_i(1 - g_i)^2 + \frac{\sigma^2}{n}\sum_i g_i^2, \qquad D = 1 - \frac{1}{n}\sum_i g_i^2. \quad (2)$$

Stationarity, $\partial_{g_i} N = R_\infty\, \partial_{g_i} D$, reads $-2p_i(1 - g_i) + 2(\sigma^2/n)g_i = -2R_\infty g_i/n$, i.e. $g_i^\star = p_i/(p_i + (\sigma^2 + F^\star)/n)$ with $F^\star = R_\infty(g^\star)$. Since $s_i = \kappa g_i/(1 - g_i) = \kappa np_i/(\sigma^2 + F^\star)$, the stationary point is $s \propto p$. Substituting into (2) with $c = (\sigma^2 + F^\star)/n$ and the identity $p_i c^2/(p_i + c)^2 + cp_i^2/(p_i + c)^2 = cp_i/(p_i + c)$ gives $N = F^\star D$ iff $c\sum_i p_i/(p_i + c) = F^\star$, which is the fixed point $F^\star = \sum_i (\sigma^2 + F^\star)p_i/(np_i + \sigma^2 + F^\star)$: Proposition 1a's value with σ^2 replaced by the effective noise $\sigma^2 + F^\star$.

Global optimality (Dinkelbach). N is strictly convex (a sum of quadratics with positive coefficients) and D is concave and positive on the admissible set ($\sum_i g_i^2 < \sum_i g_i \le n$). Let $g^\star$ be stationary with value $t = N(g^\star)/D(g^\star)$. Then $\nabla N(g^\star) = t\nabla D(g^\star)$, so $\phi_t := N - tD$ is strictly convex with $\nabla\phi_t(g^\star) = 0$: $g^\star$ is its unique global minimiser and $\phi_t(g^\star) = 0$, hence $N(g) > tD(g)$, i.e. $R_\infty(g) > t$, for every admissible $g \neq g^\star$ (Dinkelbach, 1967). Every stationary point is therefore the unique global minimiser, the fixed-point equation has exactly one positive root, and that root is $\min R_\infty$. Admissibility is slack there: from the fixed point, $\sum_i g_i^\star = nF^\star/(\sigma^2 + F^\star) < n$, so $\mu^\star = \kappa^\star(1 - \sum_i g_i^\star/n) > 0$. □

A.3 Corollaries 3 and 4, and the learning curve

Corollary 3 is Theorem 1 with $p_i \propto i^{-\beta}$. For Corollary 4: a flat spectrum has a common ρ, so $R(\text{flat}) = \min_\rho(\rho^2 \sum_i p_i + dn\sigma^2)/(n+\rho)^2$; the derivative $2n(\rho\sum_i p_i - d\sigma^2)/(n+\rho)^3$ vanishes at $\bar\rho = d\sigma^2/\sum_i p_i = \sigma^2/\bar p$ with value $d\sigma^2\bar p/(n\bar p + \sigma^2) = d\,f(\bar p)$, while $R^\star = \sum_i f(p_i)$ by Proposition 1a; $f''(q) = -2n\sigma^4/(nq+\sigma^2)^3 < 0$, so Jensen gives $\text{mean}_i f(p_i) \le f(\bar p)$ with equality iff all p_i are equal. The "iff" does not need the orthogonal design: by Theorem 1 the flat spectrum is optimal iff $\mathbf{1} \propto p$. Numerically, at $d = 64$, $n = 128$, $\sigma^2 = 0.25$ and with p normalised to $\sum_i p_i = 1$, the cost of isotropy is 0.5 % at $\beta = 0.25$, 13.6 % at $\beta = 1$ and 151.8 % at $\beta = 2$. The normalisation matters, because f depends on p_i only through np_i/σ^2: under the mean-one normalisation $\sum_i p_i = d$ the same formulas give 0.01 %, 0.27 % and 6.4 %. Learning-curve form: with $p_i = Ai^{-\beta}$, $np_i = \sigma^2$ at $i_c = (An/\sigma^2)^{1/\beta}$; $(1/p_i + n/\sigma^2)^{-1} \approx \sigma^2/n$ for $i < i_c$ and $\approx p_i$ beyond, and $\sum_{i>i_c} Ai^{-\beta} \approx Ai_c^{1-\beta}/(\beta-1)$ for $\beta > 1$, so $R^\star \approx (\sigma^2 i_c/n)\,\beta/(\beta-1) = \Theta(n^{1/\beta-1})$ (the standard integral approximation, $\beta > 1$, $d \gg i_c$).

A.4 Theorem B1, Corollary B2, Proposition B3

Proof of Theorem B1. $\text{Cov}(u, Y_c) = \mathbb{E}[u\,\mathbf{1}\{y = c\}] - \pi_c\mathbb{E}[u] = \pi_c\mu_c - \pi_c\bar\mu = \pi_c(\mu_c - \bar\mu)$, and $\sum_c \pi_c(\mu_c - \bar\mu) = \bar\mu - \bar\mu = 0$: one linear relation among the C vectors, so their span has dimension at most $C-1$. $P_{\text{full}} = \sum_c v_c v_c^\top$ has range $\text{span}\{v_c\} = \text{span}\{\mu_c - \bar\mu\}$, the direction space of the affine hull of the class means, which is $(C-1)$-dimensional iff the means are affinely independent. □

Proof of Corollary B2. The formula for p_i and the eigenbasis statement are immediate. By the Schur–Horn theorem the diagonals of P_{full} over all orthonormal bases are exactly the vectors majorised by its eigenvalue vector $\lambda = (\lambda_1, \dots, \lambda_{C-1}, 0, \dots, 0)$; a d-entry power law $q_i \propto i^{-\beta}$ with $\sum q = \sum \lambda$ is majorised by λ iff $\sum_{i\le k} q_{(i)} \le \sum_{i\le k}\lambda_i$ for every $k < C-1$, which holds whenever the eigenvalue head is at least as heavy as the power law's head. So a power-law diagonal can exist in some basis, and where it does it is the head's mass spread out; its body carries no information beyond the $C-1$ eigen-directions. Theorem 1 needs a diagonal prior ($p \succ 0$; zero eigenvalues are the limit $p_i \downarrow 0$); for a non-diagonal P_{full} the Bayes estimator $(U^\top U + \sigma^2 P_{\text{full}}^{-1})^{-1}U^\top$ is not a diagonal-penalty ridge and no single-λ diagonal spectrum attains it, while under the orthogonal-design risk of Proposition 1a the cross terms of P_{full} drop out of $\sum_i s_i\mathbb{E}(\hat\beta_i - \beta_i^*)^2$, so $s \propto \text{diag}(P_{\text{full}})$ is the optimal diagonal spectrum there. This is the prescription tested in §4–§6. □

Remark. For non-diagonal P_{full}, $s \propto \text{diag}(P_{\text{full}})$ is again the optimal diagonal spectrum under the orthogonal-design risk; in the eigenbasis it reduces to LDA.

Proposition B3 (the best power law overshoots the body)**.** *Let the prior, in its sorted coordinates, be $p_i = h$ for $i \le m$ and $p_i = t\,i^{-\beta}$ for $i > m$, and let $R(\alpha) = \min_\lambda \sum_i \varphi_i(\lambda i^\alpha)$ be the orthogonal-design risk of the power-law spectrum $s_i = i^{-\alpha}$ with tuned λ. Write $\rho_i = \lambda_\beta i^\beta$ for the penalties at $\alpha = \beta$ (λ_β the tuned penalty there) and $e_i = \rho_i p_i - \sigma^2$ for the excess penalty of coordinate i, zero at the Wiener optimum. (a) If $h > t$, the body is under-penalised at $\alpha = \beta$ ($e_i = \lambda_\beta t - \sigma^2 < 0$ for all $i > m$) and, provided $e_1 = \lambda_\beta h - \sigma^2 \ge 0$, $R'(\beta) < 0$: β is not optimal and, if R is unimodal in α (true numerically in every family checked, Appendix D), $\alpha^\star > \beta$. (b) If $h < t\,m^{-\beta}$, the body is over-penalised at $\alpha = \beta$ and, provided $e_m = \lambda_\beta h m^\beta - \sigma^2 \le 0$, $R'(\beta) > 0$, so β is not optimal and, under the same unimodality, $\alpha^\star < \beta$. (c) As $n \to \infty$ with d fixed ($n \gg \sigma^2/\min_i p_i$), $(\alpha^\star, \lambda^\star)$ converge to the minimiser of the weighted least-squares problem $\min_{\alpha,\lambda}\sum_i p_i(\lambda i^\alpha - \sigma^2/p_i)^2$. For a pure power law its exponent is $\alpha_\infty = \beta$; for the two-part prior with $h > t$ the argument of (a) with*

weights i^β gives $\alpha_\infty > \beta$ under the analogous side condition $\lambda_\infty h - \sigma^2 \geq 0$ and unimodality, and $\alpha_\infty \to \beta$ as the body lengthens ($d \to \infty$). Throughout, the tuned λ_α is assumed to be the unique minimiser in λ at fixed α; it is interior automatically.

Proof of Proposition B3. By the envelope theorem, valid because λ_α is the unique minimiser, $R'(\alpha) = \sum_i \varphi_i'(\rho_i)\rho_i \ln i$ at $\lambda = \lambda_\alpha$, and stationarity in λ reads $\sum_i \varphi_i'(\rho_i)\rho_i = 0$. With $a_i := \varphi_i'(\rho_i)\rho_i = w_i e_i$, where $w_i := 2n\rho_i/(n+\rho_i)^3 > 0$, these read $\sum_i a_i = 0$ and $R'(\beta) = \sum_i a_i \ln i = \sum_i a_i \ln(i/m)$. At $\alpha = \beta$ the body excesses coincide, $e_i = \delta := \lambda_\beta t - \sigma^2$ for $i > m$, and the head excesses $e_i = \lambda_\beta h i^\beta - \sigma^2$ increase in i. *(a)* If $h > t$ then $hi^\beta > t$ for every head i, so $e_i > \delta$ on the head; $\sum_i w_i e_i = 0$ with every head term above δ and every body term equal to δ forces $\delta < 0$. If moreover $e_1 \geq 0$, every head excess is ≥ 0, so on the head $a_i \geq 0$ and $\ln(i/m) \leq 0$, while on the body $a_i < 0$ and $\ln(i/m) > 0$: every term of $R'(\beta)$ is ≤ 0 and the body terms are < 0. *(b)* is the mirror image: $h < tm^{-\beta}$ gives $hi^\beta < t$ for $i \leq m$, so the head excesses lie below δ, forcing $\delta > 0$; with $e_m \leq 0$ every term of $R'(\beta)$ is ≥ 0 and the body terms are > 0. *(c)* For $\rho \ll n$, $\varphi_i(\rho) = \sigma^2/n + (\rho^2 p_i - 2\rho\sigma^2)/n^2 + O(\rho^3/n^3)$, so $n^2(R - d\sigma^2/n) \to \sum_i(\rho_i^2 p_i - 2\rho_i\sigma^2) = \sum_i p_i(\rho_i - \sigma^2/p_i)^2 - \sum_i \sigma^4/p_i$. The Wiener penalties σ^2/p_i are $O(1)$ in n (up to σ^2/p_d in d), so for fixed d and $n \gg \sigma^2/\min_i p_i$ the expansion is uniform on a neighbourhood of the optimum and the power-law optimum converges to the weighted-least-squares minimiser; for a power-law prior the residual vanishes at $\alpha = \beta$ and nowhere else. □

Remarks. The side conditions $e_1 \geq 0$ in (a) and $e_m \leq 0$ in (b) say that at $\alpha = \beta$ the head is not mis-penalised in the direction that moving α would worsen; they are sufficient, not necessary. Between the two regimes ($tm^{-\beta} < h < t$) the sign of $R'(\beta)$ is not determined by h alone. By (c), $\alpha^\star(n) - \beta$ need not vanish as n grows (in the checked family $\alpha_\infty = 1.00, 1.00, 1.02, 1.04, 1.14$ for head excess $1, 2, 5, 10, 50$, against $\alpha^\star = 0.82, 1.00, 1.30, 1.58, 2.50$ at $n = 100$).

A.5 Theorem A: Propositions A1, A3 and Corollaries A2, A4

Proof of Proposition A1. With $\rho_i = \sigma^2/\hat{p}_i = \rho_i^\star/(1+\eta_i)$, $\rho_i^\star = \sigma^2/p_i$, expand $\Delta\rho_i = -\rho_i^\star\eta_i + O(\eta_i^2)$. Since $\varphi_i'(\rho_i^\star) = 0$, $\mathbb{E}[\varphi_i(\rho_i) - \varphi_i(\rho_i^\star)] = \frac{1}{2}\varphi_i''(\rho_i^\star)\rho_i^{\star 2}\mathbb{E}\eta_i^2 + o(1/k)$ (the first-order term vanishes although $\mathbb{E}\Delta\rho_i \neq 0$; higher moments of η are $o(1/k)$). From $\varphi_i'(\rho) = 2n(\rho p_i - \sigma^2)/(n+\rho)^3$, $\varphi_i''(\rho) = 2n[p_i(n+\rho) - 3(\rho p_i - \sigma^2)]/(n+\rho)^4$, so at $\rho_i^\star$ (where $\rho p_i = \sigma^2$) $\varphi_i'' = 2np_i^4/(np_i+\sigma^2)^3$ and $\frac{1}{2}\varphi_i''\rho_i^{\star 2} = np_i^2\sigma^4/(np_i+\sigma^2)^3 = (\sigma^2/n)g_i^2(1-g_i)$, using $np_i = g_i\sigma^2/(1-g_i)$. Sum over i. Hypotheses: orthogonal-design risk; $\mathbb{E}\eta_i = 0$, $\mathbb{E}\eta_i^2 = v_i/k$; $\lambda = \sigma^2$ fixed (tuning lowers the risk, so the formula bounds the tuned plug-in). The formula was verified by simulation to 1 % for $k \geq 200$, and with heteroscedastic v_i spanning a factor 16 to 0.7 % for $k \geq 1000$ (Appendix D). □

Corollary A2 (only the boundary needs a prior)**.** *$g^2(1-g) \leq 4/27$, attained at $g = 2/3$, i.e. at $np_i = 2\sigma^2$: directions with $g_i \to 1$ and $g_i \to 0$ contribute nothing to the estimation cost, and $N_{\text{eff}} \leq \frac{4}{27}\#\{\epsilon < g_i < 1-\epsilon\} + O(\epsilon d)$.*

Proof of Corollary A2. $\frac{d}{dg}[g^2(1-g)] = g(2-3g) = 0$ at $g = 2/3$ with value $4/27$, and $g = 2/3 \iff np = 2\sigma^2$. Contributions from $g \leq \epsilon$ are $\leq \epsilon^2$ each and from $g \geq 1-\epsilon$ are $\leq \epsilon$ each, so $N_{\text{eff}} \leq \frac{4}{27}\#\{\epsilon < g_i < 1-\epsilon\} + \epsilon\,\#\{g_i \geq 1-\epsilon\} + \epsilon^2\#\{g_i \leq \epsilon\}$. □

Derivation behind Proposition A3. For i.i.d. (y_j, z_j) the label-variance statistic is $e_i = \sum_c S_c^2/k$ with $S_c = \sum_j Y_c(j)Z_i(j)$; $\mathbb{E}S_c = k\,\mathrm{Cov}(Y_c, Z_i)$ and $\mathrm{Var}\,S_c = k\,\mathrm{Var}(Y_c Z_i)$, so $\mathbb{E}e_i = k\sum_c \mathrm{Cov}(Y_c, Z_i)^2 + \sum_c \mathrm{Var}(Y_c Z_i) = kp_i + V_i$ with $V_i \approx \sum_c \pi_c(1-\pi_c)$, to first order (empirical centring of Y adds $O(1/k)$). Under the permutation null $S_c/\sqrt{k} \to N(0, \pi_c(1-\pi_c))$, so the null's quantiles, hence $\tau = q_{99} - c_0$, are k-independent to first order by the central limit theorem, with finite-k corrections at small k. Direction i is distinguishable from the floor iff $kp_i \geq \tau$. The probe uses $g_i \geq \frac{1}{2} \iff p_i \geq \sigma^2/n$; $kp_i \geq \tau$ for all such i holds once $k \geq \tau/\min\{p_i : p_i \geq \sigma^2/n\}$, and $\tau/\min\{\cdot\} \leq \tau n/\sigma^2$, with equality when the weakest usable direction sits at the threshold $p_i = \sigma^2/n$. So k_{det} is sufficient, and necessary only in that case. On real features (Appendix D, T6) the linear mean holds to 2 % in the slope, and $\tau(k)/\tau(2500)$ is within 5 % for $k \geq 10C$ and 12 % for $k \geq 5C$ but 48 % higher at one label per class, so k_{det} and k_{head} understate the cost at $k \lesssim 5C$. □

Corollary A4 (the price of isotropy in labels)**.** *Isotropise a representation with native spectrum s_{nat}, then recover the prior from k labels. The recovered representation matches the native one in risk at $k^\star \approx \max\{k_{\mathrm{det}},\ v\sigma^2 N_{\mathrm{eff}}/(n\,D(s_{\mathrm{nat}}\|p))\}$ and comes within a fraction ϵ of the oracle prior's advantage over isotropy at $k_\epsilon \approx v\sigma^2 N_{\mathrm{eff}}/(n\,\epsilon\,D(\mathbf{1}\|p))$: the better aligned the native spectrum, the more labels it was worth.*

Setting the excess of Proposition A1, $(v\sigma^2/nk)N_{\mathrm{eff}}$, equal to $D(s_{\mathrm{nat}}\|p)$ gives $k = v\sigma^2 N_{\mathrm{eff}}/(nD)$; setting it equal to $\epsilon D(\mathbf{1}\|p)$ gives k_ϵ; the $\max\{k_{\mathrm{det}}, \cdot\}$ combines the detection and estimation regimes heuristically. v and τ are estimator constants, measured, not assumed; the plug-in spectrum in practice clips below-floor directions, which Proposition A1 treats as estimated (conservative).

A.6 Alignment: Lemma 5, Corollaries 6, 8, Lemma 7

Lemma 5. *Let μ be a Borel probability measure on $\mathbb{R}^d$, $Q \in O(d)$, $\nu = Q_{\#}\mu$ and $\mathrm{Stab}(\mu) = \{S \in O(d) : S_{\#}\mu = \mu\}$. From the marginals (μ, ν) alone, Q is identifiable exactly up to the left coset $Q \cdot \mathrm{Stab}(\mu)$: $\{Q' \in O(d) : Q'_{\#}\mu = \nu\} = Q \cdot \mathrm{Stab}(\mu)$.*

Proof. $Q'_{\#}\mu = \nu = Q_{\#}\mu \iff (Q^{-1}Q')_{\#}\mu = \mu \iff Q^{-1}Q' \in \mathrm{Stab}(\mu) \iff Q' \in Q \cdot \mathrm{Stab}(\mu)$; the first equivalence uses that Q is invertible and pushforward is functorial. $\mathrm{Stab}(\mu)$ is a closed subgroup of $O(d)$. □

Corollary 6. *If $\mu = N(0, I_d)$ then $\mathrm{Stab}(\mu) = O(d)$: every orthogonal map is consistent with the observation, and an unpaired aligner's estimate carries no information about Q.*

Proof. $N(0, I)$ is rotation-invariant, so $O(d) \subseteq \mathrm{Stab}(\mu) \subseteq O(d)$. □

Lemma 7 (whitened centroids form a regular simplex)**.** *Let μ be an equal-weight mixture of K components with means $m_1, \dots, m_K \in \mathbb{R}^d$, centred so that $\sum_k m_k = 0$, and common within-component covariance W; let $B = \frac{1}{K}\sum_k m_k m_k^\top$ and $V = \mathrm{span}\{m_k\}$. Assume (H1) the centroids are affinely independent, $\dim V = K - 1$; (H2) the representation is whitened, $B + W = I_d$; (H3) the within scatter is isotropic on the centroid span* in the whitened coordinates*, $W|_V = c\,I_V$ for some $c \in [0, 1)$. Then the centroid Gram matrix is $G = (m_j \cdot m_k)_{jk} = K(1-c)(I_K - \frac{1}{K}\mathbf{1}\mathbf{1}^\top)$: $\|m_k\|^2 = (K-1)(1-c)$ for every k, $m_j \cdot m_k = -(1-c)$ for $j \neq k$, and the centroids are the vertices of a regular $(K-1)$-simplex with pairwise cosine $-1/(K-1)$.*

Proof. Let $M = [m_1 \cdots m_K]$. On V, $B = \frac{1}{K}MM^\top = I_V - W|_V = (1-c)I_V$, so $MM^\top = K(1-c)P_V$ with P_V the orthogonal projector onto V, and $G = M^\top M$ has the same non-zero eigenvalues as $MM^\top$: $K(1-c)$ with multiplicity $\dim V$. Centring gives $G\mathbf{1} = 0$; by (H1) $\mathrm{rank}\, G = K - 1$, so $\mathbf{1}$ spans the kernel and G has eigenvalue $K(1-c)$ with multiplicity $K-1$ on $\mathbf{1}^\perp$; without (H1) the kernel is larger and the conclusion fails. A symmetric $K \times K$ matrix with those eigenvalues and kernel $\mathrm{span}\{\mathbf{1}\}$ is $K(1-c)(I_K - \mathbf{1}\mathbf{1}^\top/K)$. □

(H3) is a strong condition. If the raw mixture has between-covariance B_0 and within-covariance W_0, whitening with $\Sigma = B_0 + W_0$ gives $W|_V = cI_V$ iff $W_0|_V \propto B_0|_V$: the within scatter must mirror the between scatter on V. For isotropic raw scatter $W_0 = wI$ and raw Gram $G_0 = M^\top M$ with non-zero eigenvalues $\gamma_1, \dots, \gamma_{K-1}$, the whitened Gram is $G = G_0(G_0/K + wI)^{-1}$ (push-through identity), with non-zero eigenvalues $K\gamma_j/(\gamma_j + Kw)$: a regular simplex iff $w = 0$ or the γ_j are all equal, and an approximate one when $\gamma_{\min} \gg Kw$, the deviation of the cosines from $-1/(K-1)$ controlled by the spread of those eigenvalues. The "regular simplex in the limit" of the informal statement is the exact case $c \to 0$.

Corollary 8 (cluster identity is unidentifiable up to permutation)**.** *Let μ be as in Lemma 7 with Gaussian (or elliptically symmetric) components, and let $\pi \in S_K$. There is $S_\pi \in O(d)$ with $S_\pi m_k = m_{\pi(k)}$ for all k and $S_\pi W S_\pi^\top = W$, so $(S_\pi)_{\#}\mu = \mu$ and $S_K \hookrightarrow \mathrm{Stab}(\mu)$: a centroid-matching aligner faces $K!$ configurations that are exactly equally good.*

Proof. By Lemma 7 the Gram matrix is invariant under simultaneous permutation of rows and columns, and two configurations with the same Gram matrix differ by an orthogonal map, so there is $S_\pi \in O(d)$ carrying $\{m_k\}$ to $\{m_{\pi(k)}\}$ on V; extend it by the identity on $V^\perp$. S_π preserves V and

$V^\perp$, and under (H2)–(H3) $W = cI_V \oplus I_{V^\perp}$ (B is supported on V, so $W = I - B$ is the identity on $V^\perp$ with no cross terms), hence $S_\pi W S_\pi^\top = W$. With Gaussian components $\mu = \frac{1}{K}\sum_k N(m_k, W)$ is determined by the centroid set and W, both preserved, so $(S_\pi)_\# \mu = \mu$; apply Lemma 5. No "isotropic W" hypothesis is needed, and none is possible when $K - 1 < d$ (W has eigenvalue 1 on $V^\perp$ and $c < 1$ on V). □

A.7 Denoiser width: Proposition 9

Proposition 9. *Let $z \sim N(0, \operatorname{diag} s)$ be observed as $z_t = z + \sigma\xi$, $\xi \sim N(0, I)$. (i) The Bayes denoiser is linear and diagonal in the eigenbasis, $\hat{z}_i = g_i z_{t,i}$ with $g_i = s_i/(s_i + \sigma^2)$, and achieves $\mathrm{MSE}^\star = \sigma^2 \sum_i g_i$; the trivial denoiser achieves $\sum_i s_i$, the identity $d\sigma^2$. (ii) The best diagonal denoiser of rank r keeps the r largest s_i, achieving $\sigma^2 \sum_{i \le r} g_i + \sum_{i > r} s_i$. (iii) The fraction of the achievable improvement over the trivial denoiser captured by rank r is $\kappa(r) = \sum_{i \le r} s_i g_i / \sum_i s_i g_i$, $s_i g_i = s_i^2/(s_i + \sigma^2)$. (iv) For a flat spectrum $\kappa(r) = r/d$, so a $1 - \epsilon$ fraction needs $\lceil (1-\epsilon)d \rceil$ modes at every noise level; for $s_i \propto i^{-\alpha}$ with $\alpha > 1$, $r(\epsilon) = O(1)$ independent of d, and the count of modes above the noise floor $\#\{i : g_i > \epsilon\}$ is d or 0 for a flat spectrum and $\Theta((\sigma^2\epsilon)^{-1/\alpha})$ for the power law.*

Proof. (i) For jointly Gaussian (z, z_t) with independent coordinates, $\mathbb{E}[z_i \mid z_t] = s_i z_{t,i}/(s_i + \sigma^2)$; for a diagonal linear map a, $\mathbb{E}[(a_i z_{t,i} - z_i)^2] = (a_i - 1)^2 s_i + a_i^2 \sigma^2$ is minimised at $a_i = g_i$ with value $\sigma^2 g_i$; $a_i = 0$ gives s_i and $a_i = 1$ gives σ^2. (ii) Given a support S the optimum on it is $a_i = g_i$, so the MSE is $\sum_i s_i - \sum_{i \in S}(s_i - \sigma^2 g_i)$ and $s_i - \sigma^2 g_i = s_i g_i$ is increasing in s_i: the optimal support is the top r. (iii) Immediate. (iv) For $s_i \equiv \bar{s}$ every $s_i g_i$ is equal; for $\alpha > 1$, $\sum_i s_i g_i \le \sum_i s_i < \infty$ and the partial sums reach any fixed fraction at an index independent of d; the mode count follows from $s_i/(s_i + \sigma^2) > \epsilon$. □

The proposition is about representability (rank); the experiment that measured learnability (the bottleneck width an MLP denoiser trained by Adam needs to beat its baseline, on DINO ViT-S/16 CIFAR-10 features) found width 128 for the whitened space against 8 for the native spectrum, a factor of 16 where the proposition predicts a $d/O(1)$ gap in rank; the proposition is the mechanism and the measurement the empirical statement, and a learnability version is open. This sharpens RAE's finding (Zheng et al., 2025) that a diffusion denoiser's width must match the token dimension: what must be matched is the number of modes above the noise floor, and only for a flat spectrum is that the ambient dimension.

A.8 HamJEPA and the minimax–Bayes relation

HamJEPA's Theorem 3.2 (Alvarez, 2026), checked against the paper's own statements: task set $\mathcal{W}_H = \{w : w^\top H^{-1} w \le 1\}$, worst-case task variance $V(\Sigma) = \sup_{w \in \mathcal{W}_H} w^\top \Sigma w = \lambda_{\max}(H^{1/2}\Sigma H^{1/2})$, minimised over $\Sigma \succeq 0$ under $\operatorname{tr}(H\Sigma) = c$, unique minimiser $\Sigma^\star = (c/d)H^{-1}$ with value c/d; their Theorem 3.3 is the maximum-entropy statement under the same budget. With $\tilde{\Sigma} = H^{1/2}\Sigma H^{1/2}$ the budget fixes $\operatorname{tr}\tilde{\Sigma} = c$ and the objective is $\lambda_{\max}(\tilde{\Sigma})$, minimised at $\tilde{\Sigma} = (c/d)I$: LeJEPA's "minimise the top eigenvalue at fixed trace" conjugated by the metric H, with no sample size, noise or estimator. An ellipsoid and a Gaussian prior describe the same task ensemble when the ellipsoid is a level set of the prior density; the level sets of $N(0, P)$ are $\{w^\top P^{-1} w \le \text{const}\}$, so $\mathcal{W}_H$ corresponds to $P = H$ and Theorem 1 gives $\Sigma \propto H$ against their H^{-1}. The two criteria pull in opposite directions by construction: minimising $\sup_w w^\top \Sigma w$ rewards putting little variance where the task set is wide (a spread criterion), while minimising estimation error rewards putting more variance where the prior has power (signal is lost when s_i shrinks). They coincide only under the opposite ellipsoid convention $\{w^\top H w \le 1\}$. Where an estimation risk is put on the minimax side, the minimax risk over an ellipsoid equals the Bayes risk at its least favourable prior (Pinsker, 1980; Johnstone, 2017), which for $\sum_i a_i^2 w_i^2 \le C$ is Gaussian with $P_i \propto (\mu/a_i - 1)_+$; Theorem 1 then gives $s_i \propto (\mu/a_i - 1)_+$, which truncates: directions harder than a threshold that moves with n get zero variance, something no n-free minimax statement can express, and the same truncation that appears as the cutoff i_c in the learning curve of Appendix A.3.

B ESTIMATOR, PROTOCOL AND ENCODER REGISTRY

Features and splits. 5000 training and 2000 test images per dataset, drawn once with a fixed seed, at 224 px (bicubic), features pooled per the registry (CLS token for the ViT-S/B encoders that have one, patch mean for MAE, BEiT, data2vec, I-JEPA and SigLIP, global average pool for the ResNet-50s), extracted in fp32. The checkpoints (Table 2) are MAE (He et al., 2022), BEiT (Bao et al., 2022), data2vec (Baevski et al., 2022), I-JEPA (Assran et al., 2023), DINO (Caron et al., 2021), DINOv2 (Oquab et al., 2024), CLIP (Radford et al., 2021), SigLIP (Zhai et al., 2023), MoCo v3 (Chen et al., 2021), DeiT (Touvron et al., 2021) and a supervised ViT-B/16 (Dosovitskiy et al., 2021) on ViT backbones; SwAV (Caron et al., 2020), VICReg (Bardes et al., 2022), Barlow Twins (Zbontar et al., 2021), MoCo v2 (Chen et al., 2020b), SimCLR (Chen et al., 2020a) and a supervised model on ResNet-50 (He et al., 2016); and the third-party LeJEPA ViT-S/16 (Balestriero & LeCun, 2025). The three supervised checkpoints are ImageNet (Deng et al., 2009) classifiers. They are loaded with timm (Wightman, 2019), Hugging Face Transformers (Wolf et al., 2020), torchvision and `torch.hub` in PyTorch (Paszke et al., 2019). The primary datasets are CIFAR-10, CIFAR-100 (Krizhevsky, 2009) and STL-10 (Coates et al., 2011). The 5000 training images are split once, with a fixed permutation, into a *prior split* (2500) used only to estimate $\hat{p}$, q_{99}, c_0 and $\hat{\beta}$, and a *probe split* (2500) from which the probe's n labelled points are drawn (five seeds, shared by every spectrum). Datasets with fewer than 5000 training images (Flowers-102: 2040, Oxford-IIIT Pets: 3680, DTD: 3760) use the first $\min(2500, N/2)$ rows of the permutation as the prior split and the rest as the probe split (the label-price grid is capped at the prior-split size). PCA and whitening use all training images without labels; the test set is never fitted.

The estimator and the spectrum sweep. $e_i = \sum_c (Y_c^\top Z_i)^2/N$ on the whitened scores Z of the prior split with centred one-hot targets Y_c; under independence $\mathbb{E}e_i = 1 - \sum_c \pi_c^2$, but the floor is taken from a permutation null (200 label shuffles, all $200 \times d$ values pooled): c_0 is its median, q_{99} its 99th percentile, $\tau = q_{99} - c_0$. The window is $K = \#\{i : e_i > q_{99}\}$ (undefined if $K < 8$), $\hat{p}_i = \max(e_i - c_0, 0)$, and $\hat{\beta}$ is minus the OLS slope of $\log e_{(i)}$ on $\log i$ over $i \le K$ in the chosen ordering (PC order, by encoder variance; or prior order, by $\hat{p}$ descending). Estimator B (a two-component nonlinear fit) is reported as a robustness check and never selected by outcome. Spectra probed, all at unit mean on the whitened coordinates: $i^{-\alpha}$ in PC order and in prior order for $\alpha \in \{0, .25, .5, .75, 1, 1.5, 2, 2.5, 3, 4\}$; *matched* $s_i \propto \max(\hat{p}_i,\ 0.01 \max_j \hat{p}_j)$; native; isotropic. Probe: scikit-learn's `RidgeClassifierCV` (Pedregosa et al., 2011) over 13 log-spaced penalties, accuracy on the 2000 test images.

The label price. Plug-in spectra from $k \in \{10, 25, 50, 100, 250, 500, 1000, 2500\}$ prior-split labels, five draws per k with fixed seeds, a 200-shuffle null per draw (q_{99}, c_0 as above, with the task's full class count), $\hat{p}_i = \max(e_i - c_0, 0)$, detected set $\{i : e_i > q_{99}\}$, floor $= 0.01 \max_{\text{detected}} \hat{p}_i$; detected directions get $\max(\hat{p}_i, \text{floor})$, undetected ones exactly the floor, the spectrum normalised to unit mean; if nothing is detected the spectrum is isotropic. $k^\star_{\text{nat}}$ is the smallest k with mean plug-in accuracy $\ge$ native $-$ 1 sd (over the five native probe seeds); k_{90} the smallest k with plug-in $\ge$ isotropic $+0.9\,$(oracle $-$ isotropic); undefined crossings enter medians and rank correlations as 5000. Theory quantities per cell: $p_i = \max(e_i - c_0, 0)/n_{\text{prior}}$ from the full prior split; σ^2 the OLS residual of the centred one-hot labels on all d whitened components, summed over classes; $\tau = q_{99} - c_0$ from the full-split null; $g_i = np_i/(np_i + \sigma^2)$; $N_{\text{eff}} = \sum_i g_i^2(1 - g_i)$; $k_{\text{det}} = \tau n/\sigma^2$; $\hat{v}(k) = k \cdot \text{median}_i \text{Var}_{\text{draws}}(\hat{p}_i/k)/p_i^2$ over the boundary directions $0.2 < g_i < 0.8$ (unbiased variance over the five draws); predicted excess $\hat{v}\sigma^2 N_{\text{eff}}/(nk)$ against the measured excess one-hot squared error of plug-in over oracle at $n = 200$. The Theorem-A replication: k_K the smallest grid k with mean detected count $K(k) \ge n_{\text{use}} = \#\{i : g_i(200) \ge \frac{1}{2}\}$, $k_{\text{head}} = \tau/p_{(C-1)}$.

The objective sweep. Every encoder of the registry (Table 2; a DINOv3 checkpoint could not be downloaded and is omitted), on the four tasks; families ViT-B/16 and RN50 pooled, ViT-S and ViT-H reported but not pooled; α_{nat} the power-law fit of the PC variances over the top 64 components; supervised encoders plotted and excluded from the correlations.

Breadth and the population prior. Vision: Flowers-102 (Nilsback & Zisserman, 2008) (train+val/test), Oxford-IIIT Pets (Parkhi et al., 2012) (trainval/test), DTD (Cimpoi et al., 2014) (split 1), EuroSAT (Helber et al., 2019) (deterministic 80/20), FGVC-Aircraft (Maji et al., 2013)

(trainval/test, variant level), Food-101 (Bossard et al., 2014), Caltech-101 (Fei-Fei et al., 2004) (background class dropped, deterministic 80/20); text: 20 Newsgroups (Lang, 1995) (headers, footers and quotes removed), AG News (Zhang et al., 2015), SST-2 (Socher et al., 2013) (validation as test), Banking77 (Casanueva et al., 2020), with all-MiniLM-L6-v2 (Wang et al., 2020), bge-small-en-v1.5 (Xiao et al., 2023) and all-mpnet-base-v2 (Song et al., 2020) embeddings from Sentence-Transformers (Reimers & Gurevych, 2019), raw. Population prior: one shared basis per encoder, PCA and whitening on the union of 2000 unlabelled training images from each of the encoder's ten dataset files (CIFAR-100 fine and coarse share one file); each task's prior measured in that basis on its prior split and normalised to sum one; $\bar{p}_{-t}$ the mean over the other tasks, excluding the same-image twin; spectra at $n = 200$ with the 1 % floor; `population(all)` and `nearest-task` are diagnostics.

The official LeJEPA recipe and training with targets. Official recipe: the LeJEPA repository's documented minimal example (ViT-S/8, projector $512 \to 2048 \to 2048 \to 16$ with BatchNorm, SIGReg on the projector output, AdamW, lr 2×10^{-3}, batch 256, four views, 800 epochs) with its loss, projector, optimiser and schedule unchanged; deviations: CIFAR-10 train at 64 px, bf16 autocast, one GPU. Features at 64 px for the same 5000/2000 subset at three points: backbone CLS (384-d), backbone head (512-d, what the example returns), projector (16-d); α_{nat} over the top 64 components (all 16 for the projector). The online probe in the example acts on a detached embedding, so no label gradient reaches the backbone. Training with targets: VICReg (Bardes et al., 2022) (CIFAR ResNet-18 (He et al., 2016) with 3×3 stem, projector $512 \to 2048 \to 2048 \to 2048$, loss weights 25/25/1, SGD with momentum, lr $0.3\times$batch/256 with 10 warm-up epochs and cosine decay, weight decay 10^{-4}, batch 256, two views with the standard CIFAR augmentation, 200 epochs, labels never read) plus $L_{\mathrm{spec}} = \sum_i (\lambda_i/\mathrm{mean}(\lambda) - s_i/\mathrm{mean}(s))^2/512$ on the eigenvalues of the pooled two-view backbone covariance, weight μ chosen per target by a pilot rule (five-epoch pilots at $\mu \in \{1, 10, 100\}$; the smallest μ halving L_{spec} by epoch 5 while keeping the VICReg loss within 10 % of the $\mu = 0$ pilot; $\mu = 1$ for all three targets). Targets: $s_i = 1$; $s_i \propto i^{-1}$; the population target = mean over the 20 primary cells of each cell's sorted, floor-subtracted, normalised $\hat{p}$ (2500 labels), rank-resampled to 512 entries, mean 1. Runs with kNN accuracy below 0.80 count as failed (none did). Evaluation: backbone features for the same CIFAR-10 5000/2000 subset, then the spectrum sweep and the label-price protocol; the arm's accuracy is the native-spectrum probe.

Table 2: Encoder registry with the isotropy-pressure ranks (0 reconstruction, 1 distillation/clustering/target regression, 2 contrastive, 3 decorrelation, 4 SIGReg; sup. = supervised, excluded from the objective-sweep correlations). "family" is the objective family.

name	arch	objective family	pressure	dim	pooling
mae-b	vit-b	masked-recon	0	768	patchmean
beit-b-pt22k	vit-b	masked-token	0	768	patchmean
data2vec-b	vit-b	target-regression	1	768	patchmean
ijepa-h14	vit-h	jepa	0	1280	patchmean
dino-s16	vit-s	distillation	1	384	cls
dino-b16	vit-b	distillation	1	768	cls
dinov2-s	vit-s	distillation	1	384	cls
dinov2-b	vit-b	distillation	1	768	cls
clip-b16	vit-b	contrastive-language	2	768	cls
siglip-b16	vit-b	contrastive-language(sigmoid)	2	768	patchmean
mocov3-vitb	vit-b	contrastive	2	768	cls
deit-s	vit-s	supervised	sup.	384	cls
vit-b-sup	vit-b	supervised	sup.	768	cls
lejepa-s16-3p	vit-s	sigreg (3rd-party)	4	384	cls
swav-rn50	rn50	clustering	1	2048	avgpool
vicreg-rn50	rn50	variance-covariance	3	2048	avgpool
barlow-rn50	rn50	cross-correlation	3	2048	avgpool
mocov2-rn50	rn50	contrastive	2	2048	avgpool
simclr-rn50	rn50	contrastive	2	2048	avgpool
rn50-sup	rn50	supervised	sup.	2048	avgpool

C PER-CELL TABLES

Price and pressure rank. The price correlation has the wrong sign because k_{90} measures a *fraction* of the oracle-over-isotropy gap, and the reconstruction encoders have the smallest gaps (cost 0.10–0.29, against 0.37–0.39 for the invariance-based ViT-Bs) together with the largest k_{90} (750–2500).

Table 3: The spectrum sweep at $n = 200$ on the 20 primary cells. K = directions above the null's window; $\hat{\beta}$ (R^2) in prior order; best α in prior order with its accuracy; matched $= s \propto \hat{p}$; cost of isotropy = best label-informed spectrum − isotropic.

encoder	task	C	K	$\hat{\beta}$ (R^2)	best α	acc	matched	native	isotropic	cost
deit-small	cifar10	10	51	1.63 (0.92)	2	0.889	0.878	0.878	0.631	+0.258
deit-small	cifar100-coarse	20	147	0.98 (0.97)	1.5	0.614	0.614	0.605	0.435	+0.179
deit-small	cifar100-fine	100	282	0.68 (0.92)	0.5	0.349	0.351	0.344	0.300	+0.051
deit-small	stl10	10	53	1.66 (0.89)	2	0.967	0.960	0.954	0.764	+0.203
dino-vits16	cifar10	10	65	1.45 (0.95)	1.5	0.895	0.886	0.886	0.604	+0.290
dino-vits16	cifar100-coarse	20	155	0.96 (0.97)	1.5	0.634	0.635	0.631	0.433	+0.202
dino-vits16	cifar100-fine	100	304	0.64 (0.95)	0.5	0.312	0.325	0.312	0.263	+0.061
dino-vits16	stl10	10	47	1.68 (0.89)	2	0.971	0.967	0.965	0.737	+0.234
dinov2-small	cifar10	10	43	1.75 (0.90)	2	0.951	0.946	0.941	0.686	+0.265
dinov2-small	cifar100-coarse	20	140	1.07 (0.96)	2.5	0.725	0.722	0.708	0.515	+0.210
dinov2-small	cifar100-fine	100	286	0.76 (0.91)	0.75	0.454	0.464	0.465	0.378	+0.087
dinov2-small	stl10	10	50	1.62 (0.92)	3	0.985	0.980	0.976	0.769	+0.215
lejepa-vits16	cifar10	10	81	1.27 (0.96)	1.5	0.793	0.780	0.785	0.536	+0.257
lejepa-vits16	cifar100-coarse	20	142	0.95 (0.97)	1.5	0.560	0.558	0.551	0.387	+0.173
lejepa-vits16	cifar100-fine	100	277	0.62 (0.96)	0.5	0.246	0.252	0.220	0.209	+0.043
lejepa-vits16	stl10	10	59	1.43 (0.94)	2	0.912	0.908	0.909	0.690	+0.222
vit-mae-base	cifar10	10	118	1.04 (0.94)	1.5	0.725	0.721	0.712	0.371	+0.354
vit-mae-base	cifar100-coarse	20	177	0.83 (0.95)	1	0.469	0.469	0.448	0.250	+0.219
vit-mae-base	cifar100-fine	100	326	0.50 (0.97)	0.75	0.212	0.220	0.155	0.147	+0.073
vit-mae-base	stl10	10	87	1.22 (0.84)	2.5	0.920	0.915	0.911	0.497	+0.423

Table 4: The label price per cell at $n = 200$: plug-in accuracy against the prior-label budget k (mean over five draws × five probe seeds), with the isotropic, native and oracle references and the two crossings.

encoder	task	C	iso	native	oracle	k=10	25	50	100	250	500	1000	2500	$k^\star_{\mathrm{nat}}$	k_{90}
deit-small	cifar10	10	0.631	0.878	0.878	0.630	0.722	0.800	0.862	0.870	0.876	0.878	0.878	500	100
deit-small	cifar100-coarse	20	0.435	0.605	0.614	0.413	0.423	0.456	0.522	0.583	0.604	0.612	0.614	500	500
deit-small	cifar100-fine	100	0.300	0.344	0.351	0.275	0.277	0.275	0.233	0.310	0.339	0.347	0.351	500	1000
deit-small	stl10	10	0.764	0.954	0.960	0.793	0.869	0.925	0.942	0.956	0.958	0.959	0.960	250	100
dino-vits16	cifar10	10	0.604	0.886	0.886	0.605	0.707	0.813	0.854	0.871	0.880	0.883	0.886	500	250
dino-vits16	cifar100-coarse	20	0.433	0.631	0.635	0.420	0.429	0.446	0.521	0.591	0.617	0.629	0.635	1000	500
dino-vits16	cifar100-fine	100	0.263	0.312	0.325	0.252	0.252	0.248	0.205	0.278	0.313	0.322	0.325	500	1000
dino-vits16	stl10	10	0.737	0.965	0.967	0.802	0.882	0.937	0.959	0.961	0.963	0.966	0.966	250	100
dinov2-small	cifar10	10	0.686	0.941	0.946	0.682	0.813	0.887	0.930	0.940	0.944	0.945	0.946	250	100
dinov2-small	cifar100-coarse	20	0.515	0.708	0.722	0.506	0.516	0.569	0.659	0.695	0.712	0.718	0.722	250	500
dinov2-small	cifar100-fine	100	0.378	0.465	0.464	0.347	0.344	0.339	0.330	0.416	0.445	0.458	0.460	1000	1000
dinov2-small	stl10	10	0.769	0.976	0.980	0.798	0.863	0.912	0.956	0.971	0.977	0.979	0.980	250	250
lejepa-vits16	cifar10	10	0.536	0.785	0.780	0.518	0.589	0.651	0.705	0.754	0.770	0.777	0.780	1000	500
lejepa-vits16	cifar100-coarse	20	0.387	0.551	0.558	0.366	0.376	0.397	0.476	0.528	0.544	0.554	0.558	500	500
lejepa-vits16	cifar100-fine	100	0.209	0.220	0.252	0.169	0.166	0.163	0.153	0.216	0.242	0.250	0.251	250	1000
lejepa-vits16	stl10	10	0.690	0.909	0.908	0.707	0.800	0.859	0.875	0.898	0.905	0.907	0.908	500	250
vit-mae-base	cifar10	10	0.371	0.712	0.721	0.340	0.368	0.444	0.553	0.658	0.698	0.714	0.720	1000	500
vit-mae-base	cifar100-coarse	20	0.250	0.448	0.469	0.213	0.199	0.226	0.273	0.401	0.441	0.461	0.470	500	1000
vit-mae-base	cifar100-fine	100	0.147	0.155	0.220	0.105	0.097	0.107	0.104	0.155	0.196	0.213	0.220	250	2500
vit-mae-base	stl10	10	0.497	0.911	0.915	0.487	0.534	0.647	0.778	0.890	0.907	0.913	0.915	500	250

The measure is therefore confounded with linear probeability, and the hypothesis stays open until a gap-normalised price and a finer k grid exist.

Table 10: The breadth cells at $n = 200$ (77 vision + 12 text). $\hat{\beta}$ in prior order; best α (PC or prior order, whichever is higher) with its accuracy; matched $= s \propto \hat{p}$; cost of isotropy = best α's accuracy − isotropic; k_{90} and $k^\star_{\mathrm{nat}}$ from the label-price protocol on the same cell.

encoder	dataset	C	d	$\hat{\beta}$	best α (acc)	matched	native	isotropic	cost	k_{90}	$k^\star_{\mathrm{nat}}$
beit-b-pt22k	aircraft	100	768	0.19	0.5 (0.044)	0.042	0.035	0.039	+0.005	>2500	100
beit-b-pt22k	caltech101	101	768	0.64	0.5 (0.353)	0.356	0.332	0.263	+0.090	1000	500

encoder	dataset	C	d	$\hat{\beta}$	best α (acc)	matched	native	isotropic	cost	k_{90}	$k^{\star}_{\mathrm{nat}}$
beit-b-pt22k	dtd	47	768	0.40	0.75 (0.210)	0.214	0.192	0.165	+0.045	2500	500
beit-b-pt22k	eurosat	10	768	1.23	1.5 (0.883)	0.854	0.871	0.471	+0.412	250	>2500
beit-b-pt22k	flowers102	102	768	0.31	0.5 (0.255)	0.261	0.212	0.194	+0.061	1000	500
beit-b-pt22k	food101	101	768	0.29	0.75 (0.068)	0.066	0.059	0.048	+0.021	2500	1000
beit-b-pt22k	pets	37	768	0.31	0.75 (0.111)	0.110	0.095	0.080	+0.031	>2500	1000
bge-small	20ng	20	384	1.10	2.5 (0.605)	0.604	0.595	0.378	+0.227	500	500
bge-small	agnews	4	384	1.75	3 (0.882)	0.867	0.869	0.602	+0.280	100	500
bge-small	banking77	77	384	0.78	1.5 (0.534)	0.539	0.531	0.449	+0.085	500	250
bge-small	sst2	2	384	2.35	4 (0.864)	0.861	0.855	0.674	+0.190	25	25
clip-b16	aircraft	100	768	0.49	0.5 (0.208)	0.206	0.195	0.146	+0.061	1000	1000
clip-b16	caltech101	101	768	0.90	2 (0.739)	0.728	0.731	0.470	+0.269	500	1000
clip-b16	dtd	47	768	0.80	1.5 (0.532)	0.531	0.532	0.315	+0.217	500	500
clip-b16	eurosat	10	768	1.35	2 (0.899)	0.879	0.894	0.486	+0.414	250	>2500
clip-b16	flowers102	102	768	0.51	1.5 (0.600)	0.585	0.570	0.389	+0.211	500	500
clip-b16	food101	101	768	0.70	0.75 (0.453)	0.471	0.448	0.302	+0.151	500	500
clip-b16	pets	37	768	0.85	1.5 (0.644)	0.643	0.640	0.333	+0.311	500	500
data2vec-b	aircraft	100	768	0.25	0.5 (0.083)	0.083	0.054	0.071	+0.012	>2500	10
data2vec-b	caltech101	101	768	0.73	0.75 (0.482)	0.473	0.437	0.356	+0.126	1000	500
data2vec-b	dtd	47	768	0.52	0.75 (0.266)	0.268	0.259	0.188	+0.078	1000	500
data2vec-b	eurosat	10	768	1.07	1.5 (0.783)	0.755	0.762	0.454	+0.329	250	2500
data2vec-b	flowers102	102	768	0.34	0.5 (0.318)	0.326	0.262	0.240	+0.078	1000	250
data2vec-b	food101	101	768	0.31	0.75 (0.084)	0.084	0.072	0.065	+0.019	2500	1000
data2vec-b	pets	37	768	0.45	0.5 (0.203)	0.205	0.171	0.149	+0.054	2500	500
dino-b16	aircraft	100	768	0.42	0.75 (0.189)	0.192	0.170	0.132	+0.057	2500	1000
dino-b16	caltech101	101	768	0.93	1 (0.702)	0.692	0.708	0.464	+0.238	500	500
dino-b16	dtd	47	768	0.79	1.5 (0.494)	0.498	0.496	0.286	+0.208	500	500
dino-b16	eurosat	10	768	1.45	1.5 (0.938)	0.918	0.928	0.489	+0.448	100	>2500
dino-b16	flowers102	102	768	0.48	1 (0.554)	0.549	0.550	0.357	+0.197	500	500
dino-b16	food101	101	768	0.54	0.75 (0.232)	0.239	0.237	0.137	+0.095	1000	1000
dino-b16	pets	37	768	0.98	1.5 (0.736)	0.731	0.730	0.379	+0.357	250	250
dinov2-b	aircraft	100	768	0.55	0.5 (0.283)	0.281	0.184	0.229	+0.054	1000	10
dinov2-b	caltech101	101	768	0.89	1.5 (0.773)	0.760	0.772	0.532	+0.242	250	>2500
dinov2-b	dtd	47	768	0.92	1.5 (0.618)	0.618	0.609	0.386	+0.232	500	250
dinov2-b	eurosat	10	768	1.38	2 (0.896)	0.875	0.886	0.523	+0.373	250	1000
dinov2-b	flowers102	102	768	0.44	4 (0.890)	0.890	0.886	0.596	+0.294	100	100
dinov2-b	food101	101	768	0.76	1 (0.499)	0.495	0.479	0.369	+0.130	500	250
dinov2-b	pets	37	768	1.15	1.5 (0.879)	0.872	0.876	0.520	+0.359	250	1000
dinov2-s	aircraft	100	384	0.50	0.5 (0.237)	0.231	0.157	0.206	+0.031	2500	10
dinov2-s	caltech101	101	384	0.96	1.5 (0.742)	0.739	0.739	0.648	+0.094	500	500
dinov2-s	dtd	47	384	0.88	1.5 (0.561)	0.566	0.559	0.439	+0.122	500	250
dinov2-s	eurosat	10	384	1.31	1.5 (0.882)	0.872	0.879	0.641	+0.242	250	1000
dinov2-s	flowers102	102	384	0.59	1.5 (0.799)	0.795	0.806	0.693	+0.106	500	500
dinov2-s	food101	101	384	0.69	0.5 (0.347)	0.358	0.338	0.286	+0.061	1000	500
dinov2-s	pets	37	384	0.98	2 (0.773)	0.770	0.767	0.561	+0.213	500	250
lejepa-s16-3p	aircraft	100	384	0.34	0.5 (0.101)	0.101	0.061	0.089	+0.012	>2500	10
lejepa-s16-3p	caltech101	101	384	0.89	0.75 (0.596)	0.593	0.576	0.525	+0.071	500	250
lejepa-s16-3p	dtd	47	384	0.75	1 (0.449)	0.441	0.444	0.351	+0.098	1000	500
lejepa-s16-3p	eurosat	10	384	1.34	2 (0.924)	0.909	0.919	0.655	+0.268	250	>2500
lejepa-s16-3p	flowers102	102	384	0.41	0.5 (0.444)	0.442	0.386	0.392	+0.052	>2500	10
lejepa-s16-3p	food101	101	384	0.49	0.5 (0.157)	0.163	0.142	0.125	+0.032	2500	500
lejepa-s16-3p	pets	37	384	0.76	1.5 (0.456)	0.464	0.432	0.334	+0.122	1000	250
mae-b	aircraft	100	768	0.26	0.5 (0.085)	0.084	0.030	0.070	+0.015	2500	10
mae-b	caltech101	101	768	0.82	0.75 (0.600)	0.610	0.599	0.409	+0.191	500	500
mae-b	dtd	47	768	0.61	0.75 (0.370)	0.387	0.354	0.239	+0.131	1000	500
mae-b	eurosat	10	768	1.18	1.5 (0.867)	0.842	0.862	0.470	+0.398	250	>2500
mae-b	flowers102	102	768	0.35	0.75 (0.354)	0.372	0.275	0.250	+0.104	1000	250
mae-b	food101	101	768	0.36	0.75 (0.118)	0.120	0.052	0.075	+0.042	2500	10
mae-b	pets	37	768	0.58	0.75 (0.382)	0.390	0.314	0.217	+0.165	1000	250
minilm	20ng	20	384	1.14	2.5 (0.567)	0.565	0.540	0.335	+0.232	500	250
minilm	agnews	4	384	1.61	1.5 (0.864)	0.850	0.846	0.596	+0.268	100	250
minilm	banking77	77	384	0.77	0.75 (0.511)	0.520	0.511	0.446	+0.065	500	250
minilm	sst2	2	384	1.55	2 (0.769)	0.748	0.736	0.601	+0.168	250	250
mocov3-vitb	aircraft	100	768	0.38	0.75 (0.148)	0.151	0.117	0.114	+0.034	2500	500
mocov3-vitb	caltech101	101	768	0.87	1 (0.675)	0.663	0.679	0.464	+0.211	500	1000
mocov3-vitb	dtd	47	768	0.77	1 (0.494)	0.497	0.491	0.291	+0.203	500	500
mocov3-vitb	eurosat	10	768	1.39	2 (0.925)	0.900	0.917	0.482	+0.443	250	>2500
mocov3-vitb	flowers102	102	768	0.45	1 (0.498)	0.494	0.475	0.330	+0.168	500	500
mocov3-vitb	food101	101	768	0.54	0.75 (0.206)	0.211	0.204	0.132	+0.075	1000	1000
mocov3-vitb	pets	37	768	0.92	1.5 (0.707)	0.703	0.698	0.368	+0.339	500	500
mpnet	20ng	20	768	1.15	1.5 (0.626)	0.607	0.598	0.292	+0.334	250	500
mpnet	agnews	4	768	1.70	1.5 (0.871)	0.846	0.853	0.502	+0.369	100	1000
mpnet	banking77	77	768	0.74	1.5 (0.558)	0.557	0.547	0.412	+0.146	1000	500
mpnet	sst2	2	768	2.08	1.5 (0.879)	0.876	0.858	0.640	+0.239	25	25
siglip-b16	aircraft	100	768	0.60	0.75 (0.313)	0.320	0.289	0.219	+0.094	1000	500
siglip-b16	caltech101	101	768	0.88	1.5 (0.741)	0.732	0.744	0.502	+0.239	500	1000
siglip-b16	dtd	47	768	0.91	1.5 (0.654)	0.656	0.653	0.399	+0.255	500	250
siglip-b16	eurosat	10	768	1.38	3 (0.914)	0.887	0.906	0.484	+0.430	250	>2500
siglip-b16	flowers102	102	768	0.53	1.5 (0.792)	0.800	0.799	0.507	+0.285	250	500
siglip-b16	food101	101	768	0.69	1 (0.484)	0.496	0.466	0.342	+0.142	500	250

encoder	dataset	C	d	$\hat{\beta}$	best α (acc)	matched	native	isotropic	cost	k_{90}	$k^\star_{\mathrm{nat}}$
siglip-b16	pets	37	768	0.84	1.5 (0.717)	0.713	0.699	0.392	+0.324	500	500
vit-b-sup	aircraft	100	768	0.43	0.5 (0.130)	0.130	0.119	0.092	+0.038	2500	500
vit-b-sup	caltech101	101	768	0.88	2 (0.765)	0.753	0.764	0.535	+0.230	500	>2500
vit-b-sup	dtd	47	768	0.82	1 (0.494)	0.497	0.491	0.298	+0.195	500	500
vit-b-sup	eurosat	10	768	1.26	1.5 (0.870)	0.841	0.861	0.481	+0.389	250	>2500
vit-b-sup	flowers102	102	768	0.41	4 (0.855)	0.853	0.846	0.567	+0.288	250	250
vit-b-sup	food101	101	768	0.74	0.75 (0.414)	0.446	0.424	0.289	+0.125	500	250
vit-b-sup	pets	37	768	1.21	2.5 (0.844)	0.835	0.839	0.509	+0.335	250	500

D VERIFICATION OF THE THEORY AND THE NUMBERS

Before the paper was written, every theorem, proposition and lemma was re-derived line by line, fresh numerics were run against the corrected statements, and every headline number was recomputed from the raw result files, with a file-and-line leakage audit of the code and cell-bootstrap confidence intervals. The scripts and their outputs will be in the anonymised repository shared during discussion (Reproducibility statement).

Re-derivation. Verified without change: Proposition 1a, Corollary 3, the learning-curve form, Theorem B1, Proposition A1, Corollaries A2 and A4, Lemma 5, Corollary 6, Proposition 9, and the HamJEPA characterisation (checked against the paper's own Lemma 3.1 and Theorems 3.2–3.3). Gaps or errors found and fixed: Theorem 1's $n < d$ case (now the zero-column argument); Proposition 1b's global optimality (now the Dinkelbach argument); Corollary B2's "no power law in any basis" (false by Schur–Horn) and the diagonal-prior caveat; Lemma 7's affine-independence and whitened-isotropy hypotheses, the vacuous mean-cosine check, and Corollary 8's antecedent. Two corrections made earlier, during the experiments, concern Proposition B3: its large-n limit is $\alpha_\infty \geq \beta$, not β, and the head excess in its numerical checks had been mislabelled. Two further findings, on how the statements are worded, are incorporated in them: Proposition A3's bound $k \geq \tau n/\sigma^2$ is sufficient (it is necessary only when a usable direction sits at the probe's threshold $p_i = \sigma^2/n$), and Proposition B3's sign conclusions use the unimodality of $R(\alpha)$ and the uniqueness of the tuned penalty, both true numerically in every family checked but not proved, so both are stated as hypotheses.

Fresh numerics (21 of 24 pass, the three "failures" being specification artefacts and one real limit). T1, Theorem 1 on fixed Gaussian, Rademacher, $t(3)$ and rank-12 duplicate-row designs at $n = 24 < d = 48$: the tuned-λ risk of $s = p$ equals $\mathrm{tr}((U^\top U/\sigma^2 + P^{-1})^{-1})$ to 4×10^{-18} with $\lambda^\star = \sigma^2$; all 416 competing spectra are strictly worse (smallest excess 5×10^{-4}); a spectrum differing from p only in a zero design column ties exactly and one differing elsewhere loses by 4×10^{-3}; Monte Carlo within three standard errors. T2, Proposition 1b: $\sum_i g_i^\star = nF^\star/(\sigma^2 + F^\star)$ to 7×10^{-12}, $R_\infty(s \propto p) = F^\star$ to 3×10^{-14}, 200 random perturbations of $s \propto p$ all worse (smallest excess $+9.9 \times 10^{-5}$) at $n/d \in \{0.26, 0.52, 2.6\}$. T3, Proposition B3 in a new family ($m = 5$, $\beta = 1.5$, $t = 0.05$, $\sigma^2 = 0.3$, $d = 256$, 66 cells): the envelope identity equals a finite difference to 4×10^{-9}; every sign prediction holds; $R(\alpha)$ has one local minimum in every cell; $\alpha_\infty = 1.51$–$1.57 > \beta$ for $h > t$. T3b, the Corollary 4 percentages reproduce to 0.03 points under the normalisation $\sum_i p_i = 1$. T4, Proposition A1 with heteroscedastic $v_i \in [0.5, 8]$: measured/predicted 0.993–1.001 for $k \geq 1000$ (1.08 at $k = 100$); tuned λ never worse than fixed; the 5 % clip never binds for $k \geq 200$. T5, Lemma 7: exact simplex at $w = 0$; eigenvalues $K\gamma_j/(\gamma_j + Kw)$ to 2×10^{-14} with a spread up to 2.06 at $w = 3$; affinely dependent centroids break the simplex by 0.52 in cosine; the old mean-cosine statistic equals $-1/(K-1)$ to 3×10^{-17} in every case; Corollary 8's permutation isometry exists to 4×10^{-15} under (H3) and does not exist for isotropic raw scatter. T6, the detection model on real features (DINO ViT-S/16, CIFAR-10 and CIFAR-100): the linear mean $\mathbb{E}e_{k,i} = V + kp_i$ holds with slope/p_i in $[0.993, 1.019]$ ($R^2 \geq 0.998$) and the floor tracks $1 - \sum_c \pi_c^2$ within 8 %, but $\tau(k)/\tau(2500)$ is 1.04/1.02/1.00 on CIFAR-10 ($k = 100/250/500$), 1.12/1.06/1.02 on CIFAR-100 coarse and 1.48/1.25/1.11/1.05 on CIFAR-100 fine ($k = 100/250/500/1000$): τ is constant to 5 % once $k \geq 10C$ and inflated by half at one label per class, so k_{det} and k_{head} understate the cost at $k \lesssim 5C$.

Table 5: Theorem-A quantities per cell at $n = 200$: σ^2, τ, $k_{\mathrm{det}} = \tau n/\sigma^2$, $n_{\mathrm{use}} = \#\{g_i \geq \frac{1}{2}\}$, k_K (labels until $K(k) \geq n_{\mathrm{use}}$), $k_{\mathrm{head}} = \tau/p_{(C-1)}$ and the measured k_{90}.

encoder	task	C	σ^2	τ	k_{det}	n_{use}	k_K	k_{head}	k_{90}
dino-vits16	cifar10	10	0.172	1.33	1545	45	2500	56	250
dino-vits16	cifar100-fine	100	0.552	0.36	132	22	250	378	1000
dino-vits16	cifar100-coarse	20	0.349	0.90	514	61	500	134	500
dino-vits16	stl10	10	0.078	1.32	3382	55	>2500	23	100
dinov2-small	cifar10	10	0.107	1.34	2500	43	2500	42	100
dinov2-small	cifar100-fine	100	0.463	0.37	159	59	250	293	1000
dinov2-small	cifar100-coarse	20	0.269	0.89	661	71	1000	98	500
dinov2-small	stl10	10	0.055	1.34	4832	77	>2500	57	250
lejepa-vits16	cifar10	10	0.232	1.32	1135	48	1000	90	500
lejepa-vits16	cifar100-fine	100	0.596	0.37	124	18	100	463	1000
lejepa-vits16	cifar100-coarse	20	0.387	0.90	465	47	500	128	500
lejepa-vits16	stl10	10	0.116	1.35	2334	56	2500	44	250
deit-small	cifar10	10	0.161	1.34	1661	36	2500	40	100
deit-small	cifar100-fine	100	0.530	0.37	139	40	250	371	1000
deit-small	cifar100-coarse	20	0.352	0.89	506	55	500	143	500
deit-small	stl10	10	0.077	1.34	3468	64	>2500	31	100
vit-mae-base	cifar10	10	0.203	1.33	1309	76	2500	79	500
vit-mae-base	cifar100-fine	100	0.498	0.37	149	10	25	595	2500
vit-mae-base	cifar100-coarse	20	0.340	0.89	526	60	500	166	1000
vit-mae-base	stl10	10	0.084	1.33	3154	105	>2500	60	250

Table 6: The objective sweep per encoder (medians over the four tasks at $n = 200$). sup. = supervised, excluded from the correlations. Undefined k_{90} crossings enter the medians as 5000.

encoder	family	pressure	α_{nat}	median k_{90}	cost of isotropy
mae-b	vit-b	0	1.79	750	0.286
beit-b-pt22k	vit-b	0	1.33	1750	0.142
data2vec-b	vit-b	1	2.00	2500	0.097
dino-b16	vit-b	1	0.74	300	0.380
dinov2-b	vit-b	1	0.61	250	0.393
clip-b16	vit-b	2	0.88	175	0.376
siglip-b16	vit-b	2	0.92	300	0.385
mocov3-vitb	vit-b	2	0.75	375	0.367
vit-b-sup	vit-b	sup.	0.78	175	0.387
swav-rn50	rn50	1	0.92	375	0.479
mocov2-rn50	rn50	2	0.99	375	0.469
simclr-rn50	rn50	2	0.93	375	0.427
vicreg-rn50	rn50	3	0.78	375	0.454
barlow-rn50	rn50	3	0.77	375	0.451
rn50-sup	rn50	sup.	0.91	375	0.466
ijepa-h14	other	0	0.99	375	0.387
dino-s16	other	1	0.73	375	0.218
dinov2-s	other	1	0.61	375	0.212
deit-s	other	sup.	0.71	300	0.191
lejepa-s16-3p	other	4	1.12	500	0.198

Recomputation from the raw files. 71 of 75 headline quantities agree with the reports to better than 0.005: the calibration regressions on every cell set; the direct-test counts (14/2 of 20; 24 of 60; 64 of 77; 7 of 12; 109/42 of 169); the label-price medians (500/500, five undefined crossings over 169 cells); the 20-cell Theorem-A correlations (-0.889, $+0.972$, $+0.971$); the objective-sweep family correlations and all 15 α_{nat} medians; the label-free transfer fractions (0.636, 0.744, median 0.892, leave-in gain $+0.0008$, 28 undefined cells); all six training-target gains. The four disagreements were the breadth-vision and pooled Theorem-A correlations, traced to a hard-coded prior-split size of 2500 in the post-hoc reader where 33 small-dataset cells have 1020, 1840 or 1880; the raw files carry the per-cell denominator, the reader was fixed, and both correlations rose ($0.949 \to 0.971$ and $0.824 \to 0.832$ on breadth vision; $0.965 \to 0.972$ and $0.864 \to 0.886$ pooled). Internal cross-checks: recomputed k_{90}, $k^{\star}_{\mathrm{nat}}$ and k_{det} match the stored values on all 169 cells; the isotropic accuracy stored by the spectrum sweep and by the label-price run agree to 0 on every cell; nine of 338 (cell, ordering) best-α pairs move under the opposite tie-break, which changes the pooled calibration line to $1.27\hat{\beta} + 0.28$, $r = 0.707$.

Table 7: Label-free transfer per encoder (11 held-out tasks each): fraction of cells recovering $\geq 50\,\%$ of the oracle advantage, fraction beating the native spectrum, and the median recovery over cells with a defined one.

encoder	tasks	frac. recovering $\geq 50\,\%$	frac. beating native	median recovery
beit-b-pt22k	11	0.55	0.82	1.26
clip-b16	11	0.91	0.91	0.92
data2vec-b	11	0.27	0.45	0.83
dino-b16	11	0.82	0.82	0.79
dinov2-b	11	0.73	0.82	0.93
dinov2-s	11	0.64	0.82	1.22
lejepa-s16-3p	11	0.45	0.45	0.78
mae-b	11	0.55	0.73	0.87
mocov3-vitb	11	0.73	0.82	0.85
siglip-b16	11	0.64	0.73	0.87
vit-b-sup	11	0.73	0.82	0.91

Table 8: The official LeJEPA recipe: the three probed layers at $n = 200$ (α_{nat} fitted over the window in parentheses; PR = participation ratio; $\hat{\beta}$ and best α in prior order).

layer	d	α_{nat}	R^2	top/bottom	PR	$\hat{\beta}$	best α	matched	native	isotropic	cost	k_{90}	$k^{\star}_{\mathrm{nat}}$
backbone-cls	384	1.21 (64)	0.96	1.5×10^8	21.6	1.35	3	0.876	0.878	0.655	+0.225	250	500
backbone-head	512	1.09 (64)	0.94	7.6×10^9	26.6	1.39	2.5	0.873	0.877	0.598	+0.283	100	500
projector	16	0.20 (16)	0.93	1.7	15.6	0.44	0.25	0.832	0.833	0.833	+0.000	10	10

Leakage audit. The prior and probe splits are the two halves of one fixed permutation; PCA and whitening are fitted on the training rows only and never read labels; the probe is fitted on its n training points (penalty chosen by leave-one-out) and scored on the test set, which is never fitted; the label-price plug-in labels come from the prior split only and its null from shuffles of those labels; the label-free transfer's shared basis is fitted on pooled unlabelled training rows and $\bar{p}_{-t}$ excludes the held-out task and its same-image twin; the LeJEPA recipe's online probe acts on a detached embedding and the training-target loop discards labels, both using labels only for the post-hoc kNN sanity check. Label provenance of the training target: the population target averages 20 cells of which five are CIFAR-10 cells, i.e. 12 500 CIFAR-10 prior-split labels shaped a sorted, basis-free target; the label-free i^{-1} arm gains $+0.201/+0.205/+0.184$ against $+0.208/+0.209/+0.199$.

Cell bootstraps (2000 resamples over cells, 95 % percentile intervals): median k_{90} $[250, 750]$ and median $k^{\star}_{\mathrm{nat}}$ $[250, 500]$ on the 20 cells; $\rho(k_{\mathrm{head}}, k_{90})$ $[0.92, 0.98]$ (20), $[0.89, 0.97]$ (60), $[0.77, 0.91]$ (89), $[0.84, 0.92]$ (169); $\rho(k_{\mathrm{det}}, k_K)$ $[0.91, 0.99]$, $[0.85, 0.95]$, $[0.95, 0.98]$, $[0.96, 0.98]$; label-free transfer: fraction recovering $[0.55, 0.72]$ and fraction beating native $[0.67, 0.82]$; pooled calibration slope $[1.04, 1.48]$, intercept $[0.07, 0.53]$, r $[0.57, 0.84]$. Training-target gains are over three training seeds (min/mean/max $+0.199/+0.205/+0.209$; per-arm probe sd ≤ 0.023).

E AI-USE DISCLOSURE IN FULL

This appendix expands the required AI use statement. The human authors set the research question and the organising claim, chose which experiments to run and in what order, and are responsible for everything below.

What the AI tool was used for. An LLM-based coding agent was used (i) to help develop the theoretical model and conceptual framework: the spectrum-as-prior reading of a single-penalty ridge probe and the setting of §3.1; (ii) to formulate mathematical claims: the statements of Theorem 1, Propositions 1a–1b, Corollaries 3–4, Theorem B (B1–B3), Theorem A (A1–A4), Lemmas 5 and 7, Corollaries 6 and 8, and Proposition 9; (iii) to provide ingredients for their proofs (Appendix A); (iv) to help design some of the experiments; and (v) to write the code for the experiments, analyses and figures. We also used it to polish the writing of the paper and to help find related work. It was not used for any other purpose.

Table 9: Training with spectral targets, per run: native-spectrum probe accuracy at three budgets, the matched probe (oracle prior applied afterwards) and the isotropic probe at $n = 200$, and k_{90}.

target	seed	n=100	n=200	n=500	matched@200	isotropic@200	k_{90}
isotropic	0	0.467	0.597	0.721	0.652	0.459	1000
isotropic	1	0.477	0.596	0.721	0.653	0.442	1000
isotropic	2	0.492	0.612	0.725	0.672	0.457	500
powerlaw	0	0.756	0.798	0.842	0.792	0.486	250
powerlaw	1	0.753	0.800	0.835	0.797	0.493	250
powerlaw	2	0.753	0.797	0.836	0.793	0.497	250
population	0	0.764	0.805	0.843	0.800	0.486	250
population	1	0.760	0.804	0.839	0.799	0.483	250
population	2	0.772	0.811	0.845	0.802	0.494	250

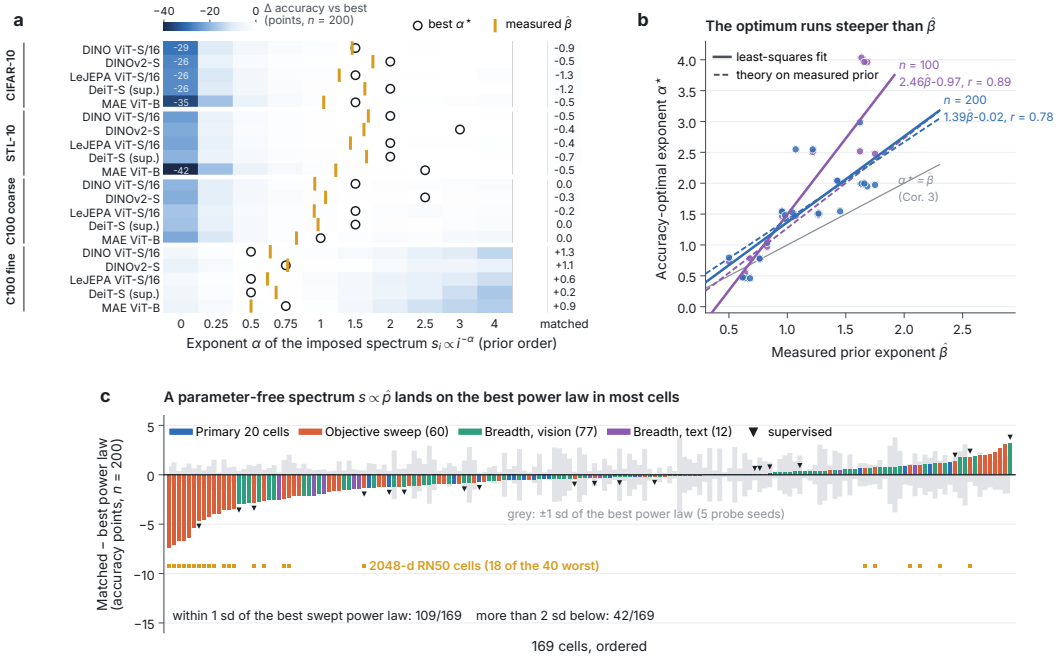


Figure 4: **Theorem 1 in the data.** (a) Accuracy at $n = 200$ minus the best swept power law ($s_i \propto i^{-\alpha}$ in prior order) for the 20 primary cells; circles: accuracy-optimal $\alpha^\star$; ticks: measured $\hat{\beta}$; right column: the matched spectrum $s \propto \hat{p}$. (b) $\alpha^\star$ against $\hat{\beta}$ with least-squares fits ($n = 100$: $2.47\hat{\beta} - 0.97$, $r = 0.89$; $n = 200$: $1.40\hat{\beta} - 0.05$, $r = 0.79$) and the theory applied to each cell's measured prior (dashed); grey: $\alpha^\star = \hat{\beta}$. (c) The direct test on 169 cells: matched minus the best power law in either order at $n = 200$, cells sorted; grey: ± 1 standard deviation of the best power law over its five probe seeds; triangles: supervised encoders; ochre: 2048-d RN50 cells. Within 1 sd: 109 of 169; more than 2 sd below: 42.

How it was checked. Every theorem was re-derived by hand in the verification, which found and fixed six gaps or errors (Appendix D). The authors checked every claim, proof step and number in the paper.

F Additional figures

Figures 5–10 belong to §4–§6; they are collected here to keep the main text within the page limit.

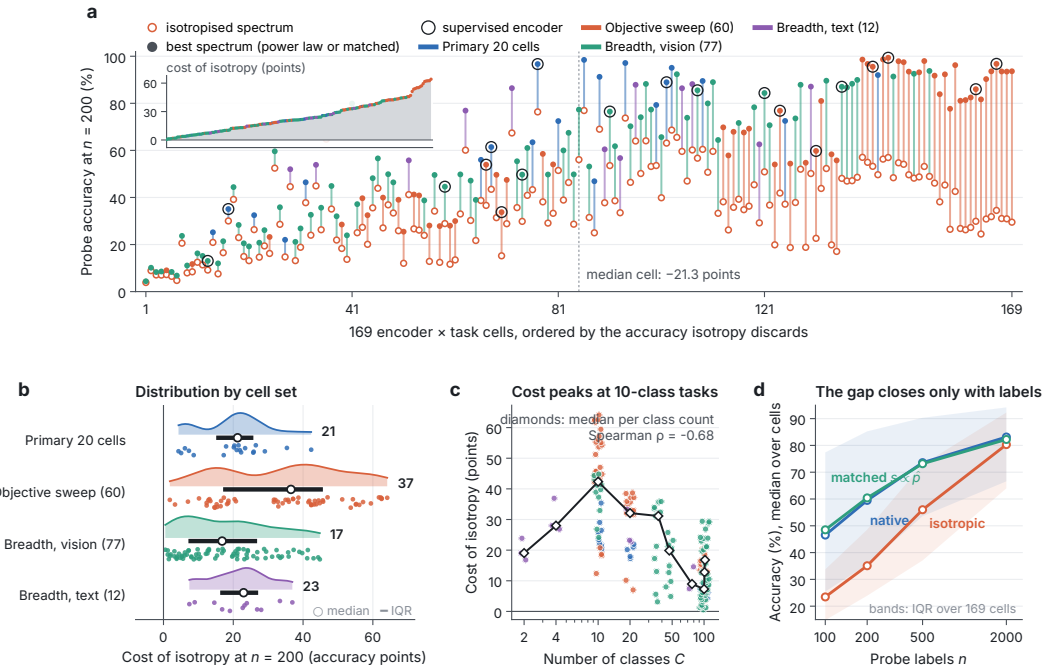


Figure 5: **The cost of isotropy on 169 cells.** (a) Accuracy at $n = 200$ of the isotropised spectrum (open) and of the best spectrum (best power law in either order, or matched; filled), cells ordered by the difference; ringed: supervised encoders; inset: the ordered cost; median 21 points. (b) Cost by cell set (density, cells, median and interquartile range). (c) Cost against the class count; diamonds: median per C (Spearman -0.68). (d) Median accuracy over cells against n for the isotropic, native and matched spectra (interquartile range shaded for the first two).

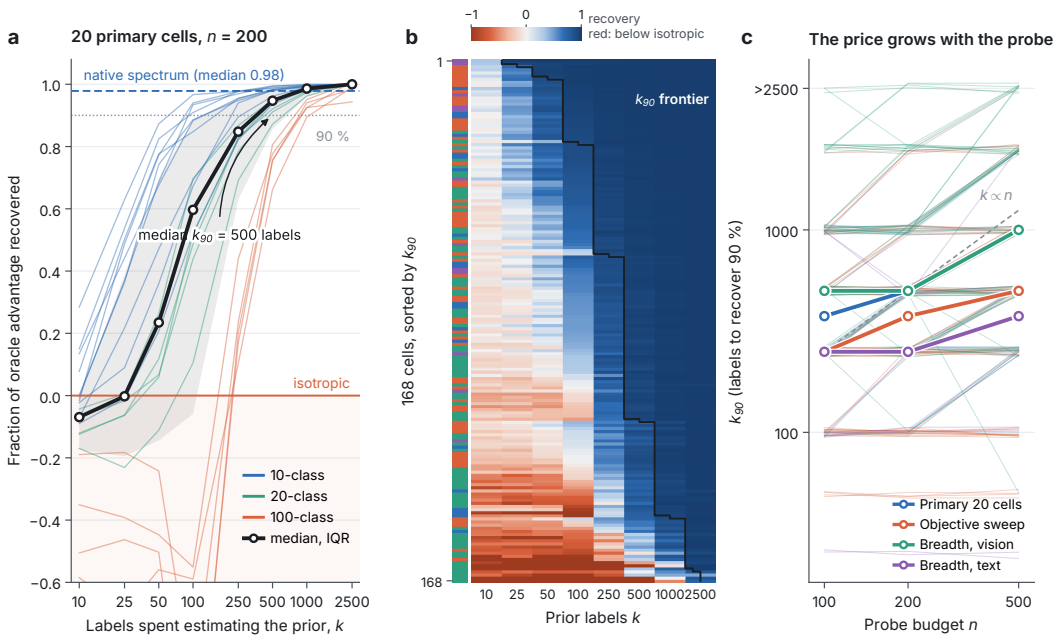


Figure 6: **Recovering the discarded prior.** (a) After isotropising, the fraction of the oracle prior's advantage recovered by a plug-in prior from k labels at $n = 200$: 20 cells coloured by class count, median with interquartile band (black), native spectrum dashed (median 0.98); median $k_{90} = 500$. (b) The same for the 168 cells of all sets whose oracle prior beats isotropy by more than 0.5 points, sorted by k_{90}; red: worse than isotropy; black step: k_{90}; left strip: cell set. (c) k_{90} against the probe budget n per cell (thin) and per-set medians; dashed: $k \propto n$; crossings beyond the grid drawn at > 2500.

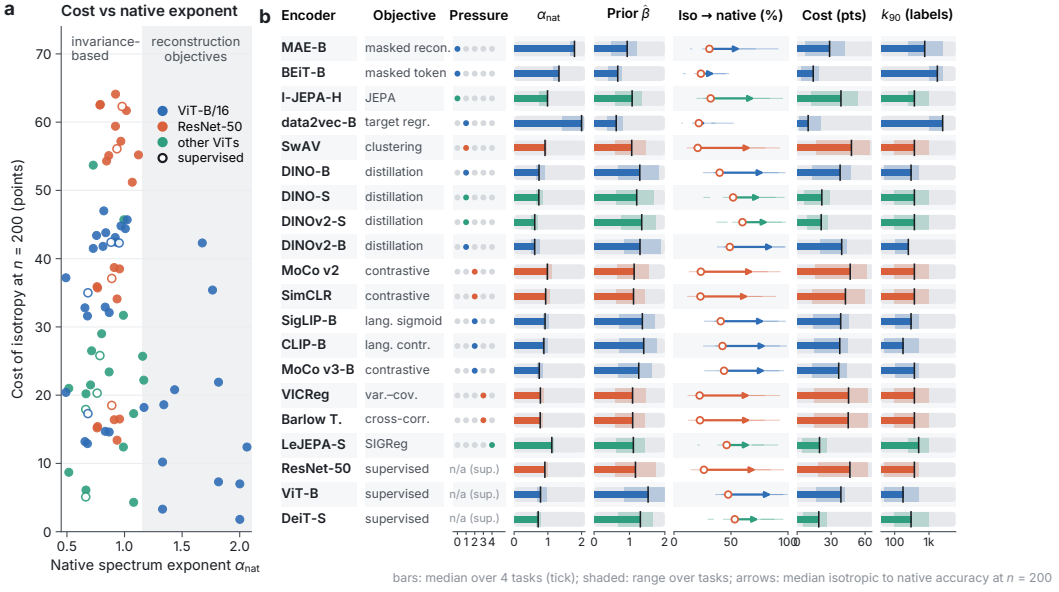


Figure 7: **The objective sweep by encoder.** (a) Native exponent α_{nat} against the cost of isotropy at $n = 200$, one dot per cell; colour = architecture, hollow = supervised; shaded: $\alpha_{\text{nat}} > 1.15$. (b) Encoders ordered by the pressure rank (0 reconstruction to 4 SIGReg): α_{nat}, $\hat{\beta}$, cost and k_{90} as the median over the four tasks (bar and tick) with the range over tasks shaded, and the median isotropic-to-native accuracy (arrow). Reconstruction objectives are steep and every invariance-based objective is flat; the price does not order with the pressure rank.

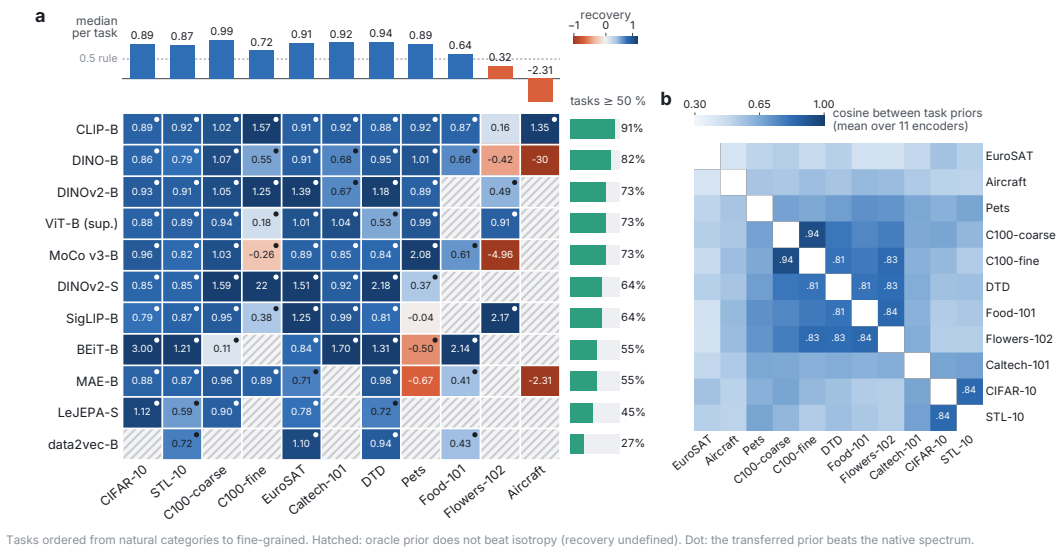


Figure 8: **The family prior transfers label-free.** (a) Fraction of (oracle − isotropic) recovered at $n = 200$ by the population prior averaged over the *other* tasks, per encoder and held-out task; tasks from natural-category to fine-grained, encoders by the fraction of tasks recovering at least 50 % (right bars); hatched: the oracle prior does not beat isotropy (28 of 121 cells, recovery undefined); dot: beats the native spectrum; top: median per task. (b) Cosine similarity between task priors in the shared basis, mean over the 11 encoders, ordered by average-linkage clustering.

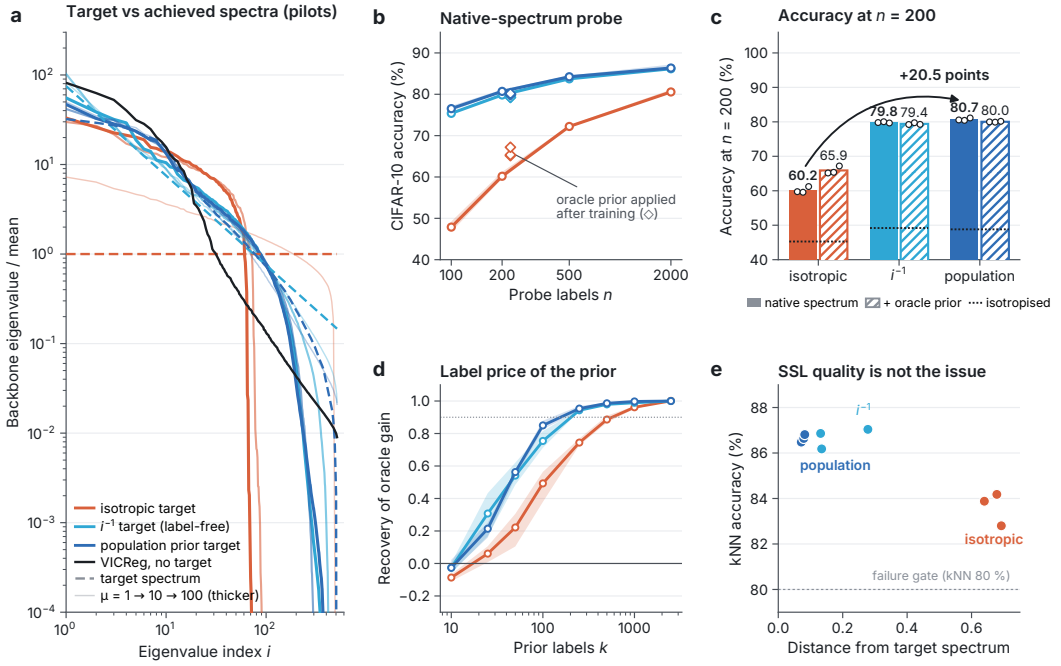


Figure 9: **Training with spectral targets.** (a) Target spectra (dashed) and the backbone spectra reached in the 5-epoch pilots at $\mu = 1, 10, 100$ (thin to thick) and with no target (black); the final runs' distances from their targets are in (e). (b) Native-spectrum probe accuracy against n per arm (mean over three seeds, range shaded); diamonds: oracle prior applied afterwards, $n = 200$. (c) Accuracy at $n = 200$ with the native spectrum (solid) and the oracle prior applied afterwards (hatched), seeds as dots, isotropised spectrum dotted; population minus isotropic: $+20.5$ points. (d) Label-price recovery curve per arm. (e) kNN accuracy against the distance from the target spectrum; every run clears the 0.80 failure gate.

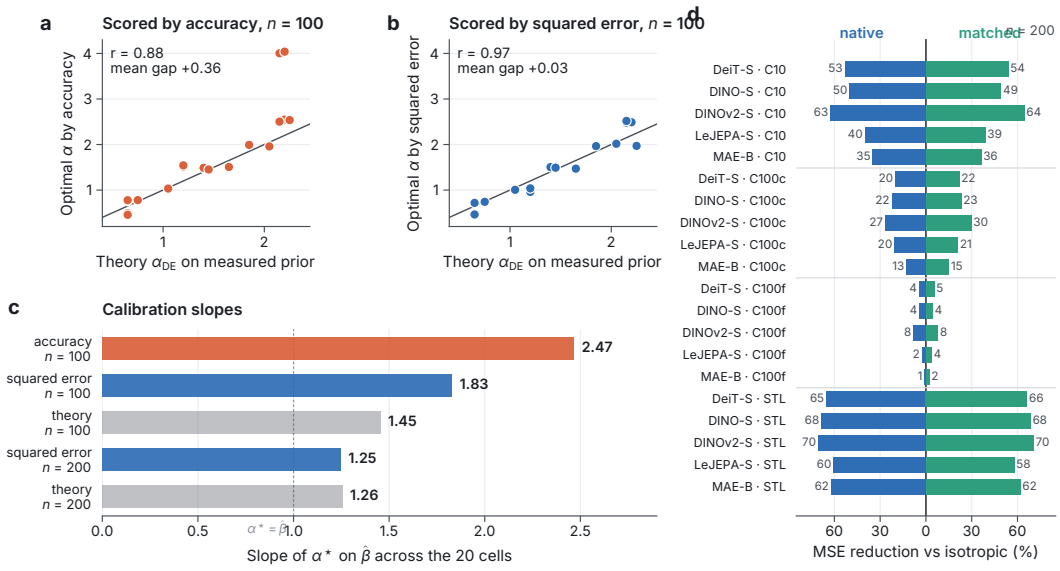


Figure 10: **Accuracy versus squared-error scoring.** (a, b) Accuracy-optimal and squared-error-optimal exponents against the theory's prediction on each cell's measured prior at $n = 100$ ($r = 0.88$ and 0.97; mean gap $+0.36$ and $+0.03$). (c) Slopes of $\alpha^\star$ on $\hat{\beta}$ across the 20 cells. (d) One-hot squared-error reduction relative to the isotropic spectrum at $n = 200$, native (left) and matched (right), per cell.

G REAL-IMAGE ILLUSTRATION

To show on images what isotropising discards, we re-extracted features: a supervised ImageNet ResNet-50 (timm `resnet50` release weights; Wightman, 2019) on 5000 training and 2000 test CIFAR-10 images from a public per-class JPEG copy, resized to 224 px, and put them through the protocol of §4 unchanged. At $n = 200$ the native spectrum reaches 68.1 %, the matched spectrum 63.1 % and the isotropised spectrum 24.3 % (five label draws), a cost of 43.8 points (the objective sweep's supervised `rn50-sup`, a different checkpoint and copy of the data, pays 56). Because the backbone average-pools its last feature map, a linear probe's logit is an exact sum over the 7×7 locations, so Figure 11 shows where each probe's evidence lies rather than a saliency estimate. Figure 12 shows the whitened directions ranked by the measured prior, and Figure 13 what a random linear view of the representation keeps with and without the spectrum.

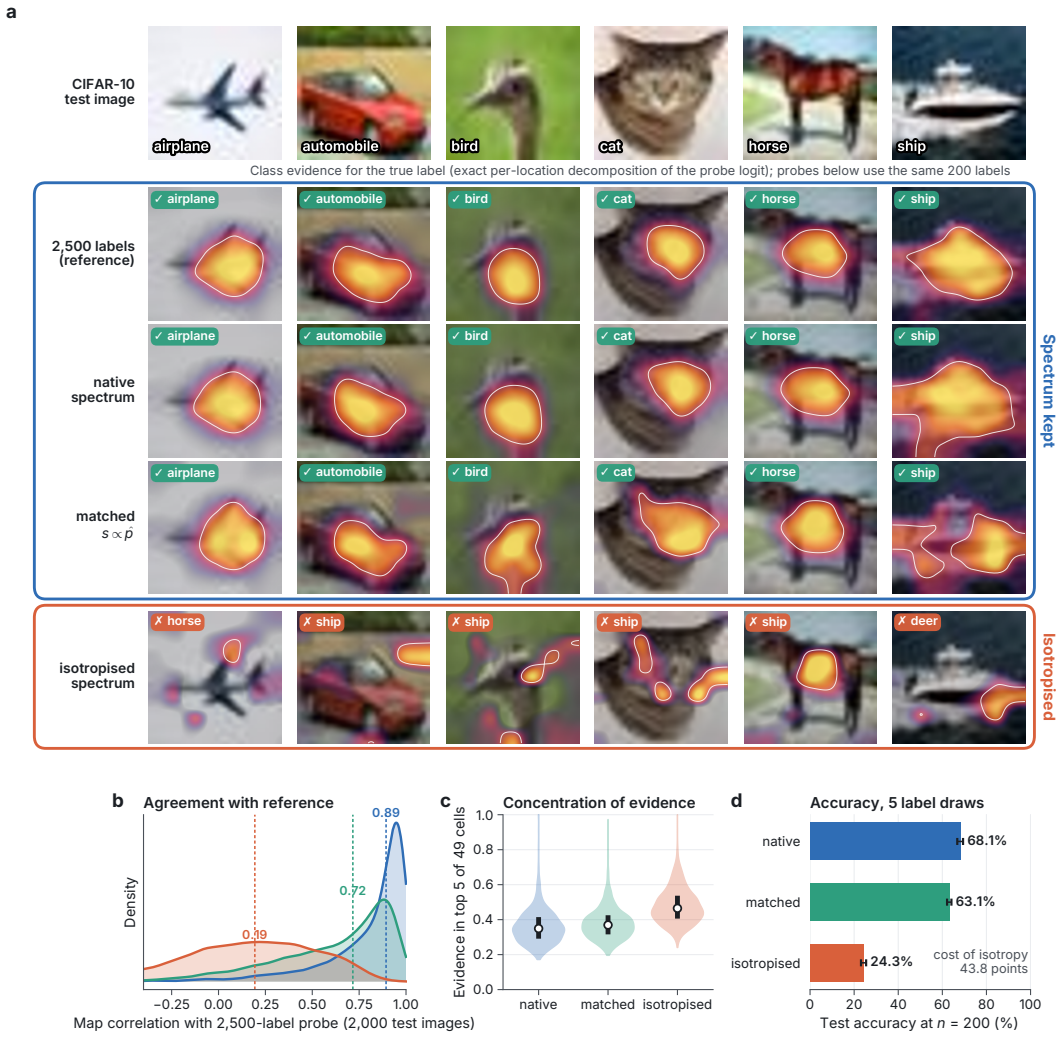


Figure 11: **Where a 200-label probe finds its evidence.** (a) CIFAR-10 test images and the per-location decomposition of the true-class logit of ridge probes on a supervised ResNet-50, each map scaled to its own 99th percentile (white contour at 60 %); the three $n = 200$ probes share the same labels and differ only in the spectrum; chips: each probe's prediction. Examples: per class, the test image with the largest reference margin among those the native probe classifies correctly and the isotropised probe does not. (b) Correlation of each probe's true-class map with the 2500-label probe's map over the 2000 test images (medians 0.89, 0.72, 0.19). (c) Fraction of the positive evidence on the top 5 of 49 locations. (d) Accuracy over five label draws.



Figure 12: **What the spectrum ranks.** (a) Measured prior $\hat{p}_i - c_0$ over the 2048 whitened directions of the ResNet-50, ranked; shaded: the 86 directions above the permutation null's q_{99} (dashed). (b) For eight ranks, the seven training images with the most negative and the most positive score, and the modal class among the 100 most extreme images at each end (chance 10 %). High-prior directions separate classes; directions with no label signal, which an isotropic spectrum weights equally, do not.

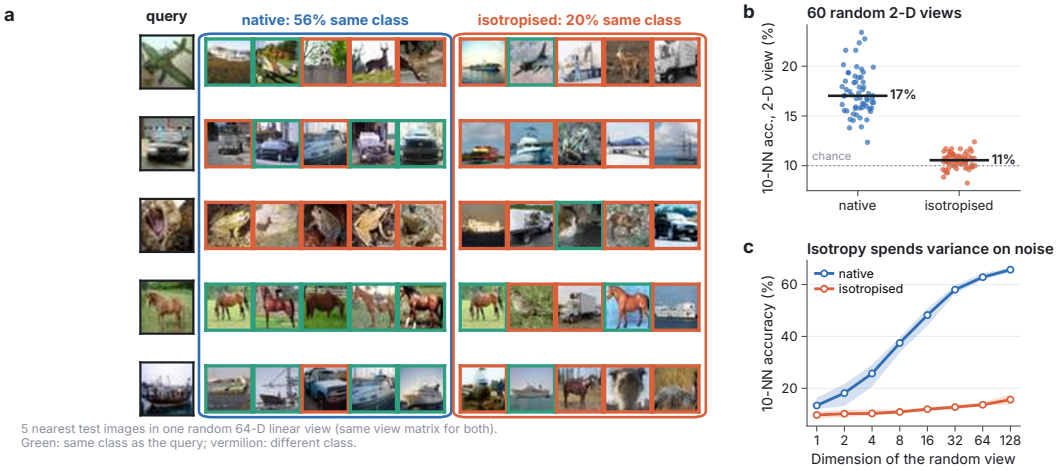


Figure 13: **What a random linear view keeps.** (a) The five nearest test images to five queries in one random 64-d projection of the whitened features scaled by the native spectrum (left) or isotropised (right); green: same class as the query. (b) 10-NN accuracy (five-fold) of 60 random 2-d views; chance 10 %. (c) 10-NN accuracy against the dimension of the random view (eight views per point, range shaded).