SIGNAL-TO-NOISE RATIO AND SAMPLE SIZE GOVERN REPRESENTATIONAL ALIGNMENT IN NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural networks are known to develop latent representations that are *aligned*, namely structurally similar across networks trained with different architectures, training protocols, or training datasets. We study this phenomenon in a controlled and simplified setting, where we train an ensemble of networks on regression and classification tasks using training sets perturbed by independent realizations of a noise process. We show that the signal-to-noise ratio (SNR) and the training sample size influence the alignment in qualitatively similar ways in networks trained on real-world datasets and in an extremely simple *linear* network with a single hidden layer, for which the alignment can be estimated analytically. Across linear and nonlinear networks, regression and classification tasks, and both synthetic and real-world data, we consistently observe that alignment varies monotonically with SNR and typically non-monotonically with training sample size. In particular, the alignment is minimal near the interpolation threshold. In the linear theory, alignment collapses at interpolation in the ridgeless limit, whereas explicit regularization prevents this collapse and yields a nonzero floor, as observed in nonlinear networks; moreover, in the regularized linear model, the regularization strength that minimizes the generalization error is exactly the one that maximizes alignment. Finally, we show that task performance is an insufficient proxy for representational informativeness: a network with lower test error can retain less of the local geometry of its input. These findings reveal a non-trivial dependence of representational alignment on the quality and quantity of the training data and on the regularization of the learning problem.

1 INTRODUCTION

Deep learning has achieved remarkable empirical success across a wide range of domains, including computer vision (LeCun et al., 2015; Krizhevsky et al., 2012), natural language processing (Cho et al., 2014; Vaswani et al., 2017), and speech recognition (Hinton et al., 2012). A common hypothesis attributes this success to the ability of neural networks to learn meaningful internal representations of data (Bengio et al., 2013; LeCun et al., 2015; Goodfellow et al., 2016). This perspective has motivated a growing body of work aimed at quantifying, understanding, and predicting *representational alignment*, that is, the extent to which different models learn similar internal representations (Sucholutsky et al., 2025). A particularly striking finding is that independently trained networks often converge to similar internal representations, even when they differ in initialization, architecture, training objective, or training datasets (Kriegeskorte et al., 2008; Raghu et al., 2017; Kornblith et al., 2019; Bansal et al., 2021; Huh et al., 2024; Acevedo et al., 2025; Gröger et al., 2026).

Despite significant empirical advances, a theoretical and mechanistic understanding of this phenomenon remains elusive. Most existing studies focus on quantifying alignment using tools such as representational similarity analysis (RSA) (Kriegeskorte et al., 2008), singular vector canonical correlation analysis (SVCCA) (Raghu et al., 2017), centered kernel alignment (CKA) (Kornblith et al., 2019), generalized shape metrics (Williams et al., 2021), and nearest-neighbor metrics (Klabunde et al., 2025; Huh et al., 2024). While these studies have made it possible to detect and compare alignment across networks, they offer limited insight into why alignment arises. Several fundamental questions therefore remain open: *(i)* how alignment depends on the quality and quantity of training

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.

data; *(ii)* whether the mechanisms underlying alignment differ between the underparameterized and overparameterized regimes; and *(iii)* whether representational alignment is predictive of model performance. Addressing these questions is important for developing a principled theory of representation learning and for characterizing the conditions under which independently trained networks encode the same information.

We first address these questions in a two-layer linear network, simple enough for closed-form analysis yet exhibiting nontrivial alignment. Although deep linear networks implement a linear map, their layered structure gives rise to hidden representations (Saxe et al., 2013), and their non-convex loss landscape has a manifold of global minima shaped by the data statistics and the architecture (Baldi & Hornik, 1989; Saxe et al., 2013). This makes them a useful model of representation learning (Saxe et al., 2019; Braun et al., 2025; Dominé et al., 2025). We then extend the analysis numerically to nonlinear networks, on the same synthetic tasks and on real-world datasets.

The representation manifolds arising in nonlinear networks can be curved and topologically complex (Ansuini et al., 2019), making measures such as CKA (Kornblith et al., 2019) and SVCCA (Raghu et al., 2017), which are primarily designed to capture correspondences expressible through linear transformations, less suited to such geometries. In this work, we aim to compare representations generated by linear and non-linear networks. We therefore use an alignment measure that is robust to curvature and topology. We define two representations as aligned when knowledge of the neighbors in one representation allows the neighbors in the other to be predicted (Huh et al., 2024; Acevedo et al., 2025; Klabunde et al., 2025). We quantify this predictability using *Conditional Copula Entropy* (CCE), an information-theoretic measure related to Information Imbalance (II) (Glielmo et al., 2022): the negative entropy of the distribution of nearest-neighbor ranks in the target representation, conditioned on the corresponding rank in the source being small (Del Tatto et al., 2024). Whereas II is an average rank, CCE is a Kullback–Leibler divergence and measures alignment as information gain in nats (Appendix B). In linear networks, CCE has a closed form, which we validate against simulations, and is an explicit monotone function of linear CKA (Sec. 4.1). Unlike CKA, CCE is generally asymmetric, so it can reveal whether one representation is more informative about another than vice versa. Our main contributions are summarized as follows:

- We develop a theoretical framework for analyzing representational alignment based on the conditional distribution of copula variables, and derive closed-form expressions for alignment in two-layer linear networks.
- We show analytically in linear networks and numerically in nonlinear ones that alignment depends on signal-to-noise ratio (SNR), and training sample size across regression and classification tasks, using both synthetic and real-world data. Across these settings, alignment increases monotonically with SNR and typically varies non-monotonically with training sample size, attaining a minimum near the interpolation threshold. In ridgeless linear networks, this minimum becomes a complete alignment collapse.
- We show that explicit regularization prevents this alignment collapse: any nonzero effective ridge regularization strength yields a finite alignment floor at interpolation, as observed in nonlinear networks.
- We show that task performance is an insufficient proxy for representational informativeness; one network can outperform another on a specific task, and yet learn less informative representations than a network with lower performance.

2 Representational Alignment from the Conditional Entropy of Rank Distribution

The statistic we use to quantify the alignment between representations is a generalization of the *Information Imbalance* (II) (Glielmo et al., 2022), which quantifies the relative information content of one representation space with respect to another by comparing how the distance ranks change between the two representations. For N data points, the II from space A to space B is

$$\Delta(A \to B) = \frac{2}{N(N-1)} \sum_{i,j|r^A_{i,j}=1} r^B_{i,j}, \tag{1}$$

where $r^A_{i,j} \in \{1, \dots, N-1\}$ denotes the rank of the distance between points i and j in space A, ordered from nearest to farthest from i; $r^B_{i,j}$ is defined analogously. Eq. 1 thus captures how well proximity in A predicts proximity in B. Values near zero indicate strong alignment, while $\Delta(A \to B) \approx 1$ indicates that A carries no information about B. The measure is generally asymmetric: $\Delta(A \to B) < \Delta(B \to A)$ implies that A is more informative than B.

From II to Conditional Copula Entropy. Let d_A, d_B be the pairwise distance variables in representation spaces A, B with cumulative distribution functions F_A, F_B and copula variables $c_A = F_A(d_A)$, $c_B = F_B(d_B)$. Del Tatto et al. (2024) show that II provides an upper bound on a function of the conditional entropy of the copula variables c_A and c_B:

$$\Delta(A \to B) \gtrsim 2 \exp\Big(\lim_{\varepsilon \to 0} H(c_B \mid c_A = \varepsilon) - 1 \Big), \tag{2}$$

where H is a differential entropy of conditional density $p(c_B|c_A)$. The negative of $\lim_{\varepsilon \to 0} H(c_B \mid c_A = \varepsilon)$ measures how much knowing that two points are infinitesimally close in A reduces uncertainty about their distance in B – a geometric characterization of local representational alignment via an information-theoretic quantity. Motivated by this interpretation, we measure representational alignment by the entropy entering the relation in Eq. 2, which, for brevity, we call *Conditional Copula Entropy* (CCE):

$$\mathrm{CCE}(A \to B) = -\lim_{\varepsilon \to 0} H(c_B \mid c_A = \varepsilon) = \lim_{\varepsilon \to 0} \mathrm{D_{KL}}\big(p(d_B \mid d_A = F_A^{-1}(\varepsilon)) \,\big\|\, p(d_B)\big). \tag{3}$$

If $\mathrm{CCE}(A \to B) = \mathrm{CCE}(B \to A) = 0$, the local proximity information in either space does not predict the distance structure in the other and thus the two representations are locally independent. Positive values indicate that the neighborhood structure of one space is informative about the other, with larger values reflecting stronger alignment. Like the II, the CCE can be asymmetric: $\mathrm{CCE}(A \to B) > \mathrm{CCE}(B \to A)$ implies that A is more informative about B than vice versa.

Estimating the CCE. In practice, the conditional density $p(c_B \mid c_A)$ is unknown and must be estimated from data. Since the marginals of c_A and c_B are uniform on $[0, 1]$ (Casella & Berger, 2024), the copula variable c_A can be estimated empirically as $r^A_{i,j}/N$ (and analogously for c_B). Consequently, distance ranks and copula variables are statistically identical; this relation allows us to estimate the CCE directly from the conditional rank samples. Therefore, given a set of N conditional ranks

$$R_{B|A} = \big\{ r^B_{i,j} \in [1, N-1] : r^A_{i,j} = 1 \big\},$$

we partition the rank support $[1, N-1]$ into $M \approx \sqrt{N}$ equal bins and estimate the entropy via the histogram-based estimator of ranks in $R_{B|A}$ (Beirlant et al., 1997; Kraskov et al., 2004). Let p_i denote the empirical probability of bin i, CCE (Eq. 3) is approximated as

$$\mathrm{CCE}(A \to B) \approx \log M + \sum_{i=1}^{M} p_i \log p_i. \tag{4}$$

We find that Eq. 4 provides a good estimate of the true entropy for sufficiently large N, and we use it to compute CCE numerically (Beirlant et al., 1997).

CCE for one-dimensional embedding spaces. Umar et al. (2026) derive a closed-form expression for the CCE when A and B are one-dimensional projections of a Gaussian vector. Let $x \sim \mathcal{N}(0, I_d)$, and set $A = \beta_1^\top x$ and $B = \beta_2^\top x$, where $\beta_1, \beta_2 \in \mathbb{R}^d$. Then A and B are jointly Gaussian, and

$$\mathrm{CCE}(A \to B) = \mathrm{CCE}(B \to A) = -\frac{1}{2} \log(1 - \rho^2) - \frac{1}{2} \rho^2, \tag{5}$$

where $\rho = \mathbb{E}\big[\beta_1^\top \beta_2\big] / \sqrt{\mathbb{E}\big[\|\beta_1\|^2\big]\, \mathbb{E}\big[\|\beta_2\|^2\big]}$ is the correlation coefficient between A and B. We use this result to obtain the CCE between the representations of independently trained linear networks in closed form.

3 LINEAR TEACHER-STUDENT SETTING

We consider a supervised learning problem based on data generated via a noisy linear teacher of the form:

$$y_i = w^* \cdot x_i + \epsilon_i, \tag{6}$$

where $x_i \in \mathbb{R}^d$ are input features, $w^* \in \mathbb{R}^d$ is the teacher weight vector and ϵ_i is scalar additive Gaussian noise. The inputs are drawn independently and identically distributed (i.i.d.) from $\mathcal{N}(0, I_d/d)$. The entries of w^* and ϵ_i are i.i.d. drawn from $\mathcal{N}(0, \sigma_w^2)$ and $\mathcal{N}(0, \sigma_\epsilon^2)$, respectively. We define the SNR $= \sigma_w^2/\sigma_\epsilon^2$. Using the setup in (Advani et al., 2020), the student is a two-layer linear network with input-output mapping

$$\hat{y}_i = W_2 W_1 x_i, \tag{7}$$

where $W_1 \in \mathbb{R}^{k\times d}$ and $W_2 \in \mathbb{R}^{1\times k}$. The network (Eq.7) is trained on n samples $\mathcal{D} = \{(x_i, y_i)\}_{i\le n}$ generated from Eq. 6 using full-batch gradient descent to minimize the mean squared error (MSE)

$$\mathcal{L}(W_1, W_2) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 = \frac{1}{n}\|W_2 W_1 X^\top - y\|^2, \tag{8}$$

where $X \in \mathbb{R}^{n\times d}$ is the data matrix and $y \in \mathbb{R}^{1\times n}$ the vector of the corresponding responses. We study the representational alignment of independently trained students at the global minimum of the loss (Eq. 8) as a function of SNR and sample size n, leaving training dynamics for future work.

Which global minimum does gradient descent select? Under the MSE loss equation 8, many combinations of input and output weights attain the global optimum. The set of minimizers is

$$\mathcal{M}_* = \{W_1^*, W_2^* : W_2^* W_1^* X^\top X = yX\}.$$

The global minimum is therefore not unique, which raises the question of which point of $\mathcal{M}_*$ gradient descent selects. In the small-initialization (rich-learning) regime (Woodworth et al., 2020), the low-rank parameterization of Advani et al. (2020) provides a convenient way to characterize this selection. Specifically, we set $W_1 = r_2\, z V^\top,\ W_2 = u\, r_2^\top$, where V contains the right singular vectors of X, $z \in \mathbb{R}^d$ encodes alignment with the input principal components, $r_2 \in \mathbb{R}^k$ is a unit-norm vector capturing representational degrees of freedom, and $u \in \mathbb{R}$ is a scalar. Let $v = uz$. Under this procedure, gradient descent converges to the stable fixed point given by

$$v^* = \big(w^* V \Lambda + \tilde{\epsilon}\, \Lambda^{1/2}\big)\Lambda^\dagger, \tag{9}$$

where Λ is the diagonal matrix of eigenvalues of $X^\top X$, $\tilde{\epsilon} \sim \mathcal{N}(0, \sigma_\epsilon^2 I_d)$, and $(\cdot)^\dagger$ denotes the pseudo-inverse.

4 RESULTS

We first derive a closed-form expression for the CCE between representations learned by linear networks from independent realizations of the inputs and noise, and validate it against simulations. We then show that the same qualitative behavior extends to nonlinear networks and real-world classification tasks.

4.1 REPRESENTATIONAL ALIGNMENT IN DEEP LINEAR NETWORK

We estimate the asymptotic alignment using the following theorem.

Theorem 1 (Asymptotic Representational Alignment from Independent Training)**.** *Consider two deep linear networks*

$$\hat{y}_i^a = W_2^a W_1^a x_i, \qquad \hat{y}_i^b = W_2^b W_1^b x_i,$$

trained independently by gradient descent from small random initializations on independent datasets $\mathcal{D}^a = \{X^a, y^a\}$ and $\mathcal{D}^b = \{X^b, y^b\}$, drawn independently from the data distribution Eq. 6 with shared w^*, σ_w^2 and σ_ϵ^2. Write the eigendecompositions of the empirical input covariances as $X^{a\top}X^a = V^a\Lambda^a V^{a\top}$ and $X^{b\top}X^b = V^b\Lambda^b V^{b\top}$, and assume that the learned first-layer weights admit the low-rank parameterization

$$W_1^a = r^a z^a V^{a\top} = r^a \tilde{z}^a, \qquad W_1^b = r^b z^b V^{b\top} = r^b \tilde{z}^b,$$

with r^a, r^b unit-norm. Then, as training time $t \to \infty$ and in the proportional high-dimensional limit $n, d \to \infty$ at fixed $\alpha = n/d$, gradient descent converges to a global minimum at which the correlation between $\tilde{z}^a$ and $\tilde{z}^b$ is

$$\rho^* = \begin{cases} \dfrac{SNR}{SNR + \frac{1}{\alpha - 1}}, & \alpha > 1, \\ \dfrac{\alpha \cdot SNR}{SNR + \frac{1}{1-\alpha}}, & \alpha < 1, \end{cases} \tag{10}$$

where $\rho^* = \lim_{t\to\infty} \rho(t)$, $\mathrm{SNR} = \sigma_w^2/\sigma_\epsilon^2$, and $\alpha = n/d$ denotes the ratio of training samples to input dimension.

This result relies on one key assumption: that the learned first-layer weights are low-rank. This assumption is justified by prior work, which has established that deep linear networks extract low-rank structures when initialized with small weights (Arora et al., 2019; Saxe et al., 2013), a form of implicit bias of the training algorithm (Kalimeris et al., 2019; Refinetti et al., 2023). As we will see, our theoretical predictions agree well with the numerical simulations. We give the full derivation in Appendix E and sketch it here. Eq. 10 follows from the global minimum solution in Eq. 9: at convergence, $\tilde{z}^{*q}$ takes the form

$$\tilde{z}^{*q} = \frac{1}{u^{*q}} \big(w^* V^q \Lambda^q + \tilde{\epsilon}\, (\Lambda^q)^{1/2} \big) (\Lambda^q)^\dagger V^{q\top}, \qquad q \in \{a, b\}, \tag{11}$$

where Λ^q is the diagonal matrix of eigenvalues $\{\lambda_i^q\}$ of $X^{q\top} X^q$, $\tilde{\epsilon}^q \sim \mathcal{N}(0, \sigma_\epsilon^2 I_d)$, and u^{*q} is the magnitude of the learned weight. The correlation between $\tilde{z}^{*a}$ and $\tilde{z}^{*b}$ is given by

$$\rho^* = \frac{\mathbb{E}\big[\tilde{z}^{*a} \tilde{z}^{*b\top}\big]}{\sqrt{\mathbb{E}\big[\|\tilde{z}^{*a}\|^2\big]\, \mathbb{E}\big[\|\tilde{z}^{*b}\|^2\big]}} = \frac{\sigma_w^2 \mathbb{E}\big[\mathrm{Tr}\big(\big(V^a \mathrm{diag}(\mathbf{1}_{\lambda_i^a > 0}) V^{a\top}\big)\big(V^b \mathrm{diag}(\mathbf{1}_{\lambda_i^b > 0}) V^{b\top}\big)\big)\big]}{\sqrt{\big[\sum_{i=1}^{R_a} \big(\sigma_w^2 + \frac{\sigma_\epsilon^2}{\lambda_i^a}\big)\big]\big[\sum_{i=1}^{R_b} \big(\sigma_w^2 + \frac{\sigma_\epsilon^2}{\lambda_i^b}\big)\big]}}, \tag{12}$$

where R_a and R_b are the ranks of X^a and X^b, respectively. In the proportional high-dimensional limit $n, d \to \infty$ at fixed ratio $\alpha = n/d$, the empirical spectral distribution of $X^{\top q} X^q$ converges to the Marchenko–Pastur (MP) law (Marčenko & Pastur, 1967; Bai et al., 2010),

$$f_\alpha(\lambda) = \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi\lambda} + (1 - \alpha)^+ \delta(\lambda), \tag{13}$$

with $\lambda_\pm = (1 \pm \sqrt{\alpha})^2$ and $(\cdot)^+ = \max(\cdot, 0)$. The first term is the continuous bulk supported on $[\lambda_-, \lambda_+]$, the second the point mass at zero that appears when $\alpha < 1$. Replacing the empirical average over eigenvalues with an integral against f_α yields the closed form of Eq. 10.

Eq. 10 has a simple interpretation: $\rho^* = \cos^2(\tilde{z}^{*q}, w^*)$, $q \in \{a, b\}$, i.e., the alignment between two students equals each student's squared cosine with the teacher, since their components orthogonal to w^* are uncorrelated across datasets (Appendix G). This links our analysis to work on the fluctuations and agreement of independently trained models (Geiger et al., 2020; Adlam & Pennington, 2020; Loureiro et al., 2022; d'Ascoli et al., 2020; Jiang et al., 2021; Nakkiran & Bansal, 2020).

Theorem 1 allows us to compute, via the CCE, the degree of alignment between representations of independently trained deep linear networks. Let $A = W_1^a X_{\mathrm{test}}^\top$ and $B = W_1^b X_{\mathrm{test}}^\top$ denote the hidden representations of a test set. For any two test inputs $x, x' \in X_{\mathrm{test}}$, the distance induced by W_1^a satisfies

$$\|r^a \tilde{z}^a x - r^a \tilde{z}^a x'\|^2 = \|r^a\|^2 \, \|\tilde{z}^a (x - x')\|^2 = \|\tilde{z}^a (x - x')\|^2, \tag{14}$$

since $\|r^a\| = 1$, and analogously for W_1^b.The neighborhood structure of A and B is thus entirely determined by the one-dimensional projections $\tilde{A} = \tilde{z}^a X_{\mathrm{test}}^\top$ and $\tilde{B} = \tilde{z}^b X_{\mathrm{test}}^\top$. Applying Eq. 5, we obtain the closed-form expression

$$\mathrm{CCE}(A \to B) = \mathrm{CCE}(\tilde{A} \to \tilde{B}) = -\frac{1}{2} \log\big(1 - \rho^{*2}\big) - \frac{1}{2} \rho^{*2}, \tag{15}$$

with ρ^* given by Eq. 10.

Panel (a) of Fig. 1 shows the average generalization error of a two-layer linear network as a function of $\alpha = n/d$, the ratio of training samples to input dimension, for SNR$\in \{0.5, 1, 2, 5, 10\}$. Solid

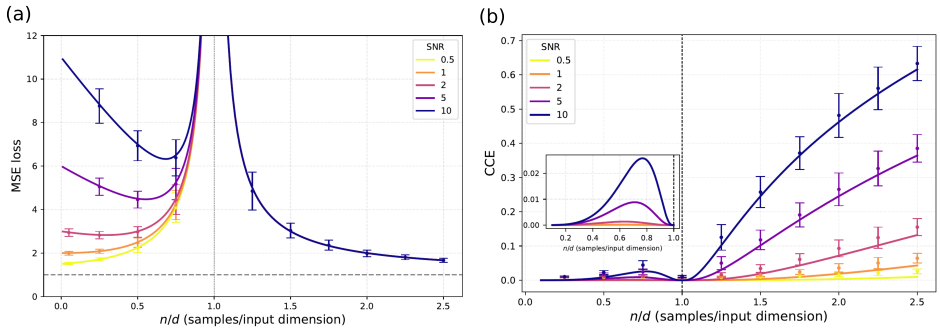

Figure 1: **Linear-network theory and simulations.** (a) Generalization error and (b) CCE between networks trained on independent datasets as a function of n/d, for several SNRs. Solid lines: theory; points: simulations (mean over 50 runs). Parameters: $d = 200$ (input dimension), $k = 100$ (hidden units), and $n_{\text{test}} = 10^4$ (test samples).

curves are the theoretical prediction (see Appendix D for its explicit form); dots with error bars are numerical simulations, averaged over 50 independent runs. The networks are evaluated after convergence, with input dimension $d = 200$ and $k = 100$ hidden units. Theory and simulations agree at all SNRs. In particular, we recover the generalization behavior reported for shallow linear models (Krogh & Hertz, 1992; Hastie et al., 2022; Advani et al., 2020), including the double descent. The interpolation threshold occurs at $n = d$, reflecting that the network is functionally equivalent to a linear model with d effective parameters.

We now turn to our main focus: the relationship between internal representations learned by identical networks trained on independent datasets. Eq. 15 characterizes the directional alignment measures $\text{CCE}(A \to B)$ and $\text{CCE}(B \to A)$, which are symmetric in this case, and thus we only report one. Panel (b) of Fig. 1 shows $\text{CCE}(A \to B)$ as a function of α for the values of the SNR as in panel (a). We observe a strong dependence on the SNR in both the underparameterized ($\alpha > 1$) and overparameterized ($\alpha < 1$) regimes, with alignment increasing monotonically (i.e., higher $\text{CCE}(A \to B)$) with SNR. For any finite SNR > 0, $\text{CCE}(A \to B)$ exhibits a non-monotonic dependence on α, mirroring the double descent of the generalization error: the CCE vanishes at the interpolation threshold $\alpha = 1$ and grows as $\alpha \to \infty$, with an SNR-dependent rate. At the interpolation threshold, $\text{CCE}(A \to B) = 0$, regardless of the SNR, meaning that the two representations carry essentially no information about each other's neighborhood structure–consistent with the divergence of the generalization error at interpolation.

Under the assumptions of Theorem 1, the linear CKA between A and B reduces to the squared Pearson correlation between the corresponding one-dimensional projections. Therefore, in the population limit,

$$\text{CKA}(A, B) = \text{CKA}(\tilde{A}, \tilde{B}) = \rho^{*2}, \tag{16}$$

where ρ^* is the representation correlation defined in Eq. 10. The proof of Eq. 16 is presented in Appendix E.2. Using Eq. 15, the CCE can consequently be written directly in terms of linear CKA as

$$\text{CCE}(A \to B) = -\frac{1}{2}\log\left(1 - \text{CKA}(A, B)\right) - \frac{1}{2}\text{CKA}(A, B). \tag{17}$$

Hence, CCE and linear CKA are monotonic functions of the same scalar order parameter, and therefore exhibit the same ordering and locations of extrema as SNR and sample size are varied. This equivalence is specific to the effectively one-dimensional linear regime. Beyond this setting, CCE retains a directional interpretation in terms of local-neighborhood predictability, whereas CKA is a symmetric global similarity measure.

4.2 Alignment in Non-Linear Networks

We now extend the analysis to nonlinear networks, adding ReLU activation while keeping the architecture of the linear case. Specifically, the two networks are of the form:

$$\hat{y}^a(x) = W_2^a \,\text{ReLU}(W_1^a x), \qquad \hat{y}^b(x) = W_2^b \,\text{ReLU}(W_1^b x). \tag{18}$$

Both networks are trained by full-batch gradient descent on independent datasets from the teacher model (Eq. 6). We compare their hidden representations $A = \text{ReLU}(W_1^a X_{\text{test}}^\top), B =$

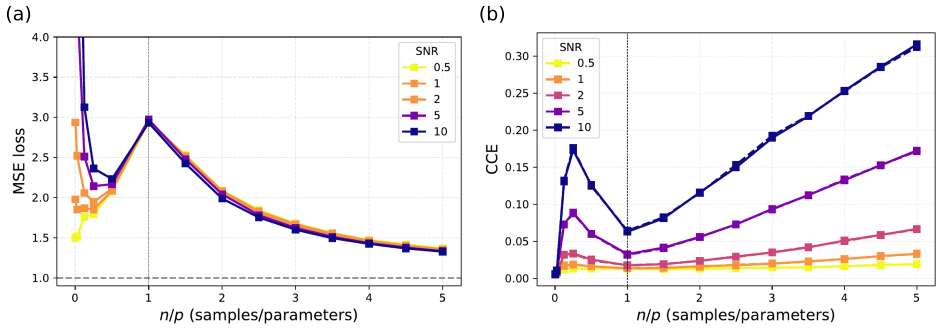

Figure 2: **Nonlinear-network results.** (a) Generalization error and (b) CCE between ReLU networks trained on independent datasets as a function of $n/dk = n/p$, for several SNRs. Curves: simulations (mean over 50 runs). Parameters: $d = 200$ (input dimension), $k = 20$ (number of hidden units), and $n_{\text{test}} = 10^4$ (test samples).

$\text{ReLU}(W_1^b X_{\text{test}}^\top)$ using CCE, as in Sec. 4.1. The nonlinearity makes an analytical treatment of the CCE substantially difficult; so here we proceed numerically.

Fig. 2 summarizes the results. Panel (a) shows the generalization error at several SNRs, panel (b) shows the CCE between representations learned from independent datasets at the same SNR. Both are plotted against the ratio of the number of training samples n to the number of network parameters, $p = dk$. Each curve is averaged over 50 independent experiments, with $d = 200$ and $k = 20$.

Let $\gamma = n/p$. As in the linear model, the generalization error exhibits a double descent around $\gamma \approx 1$, while the CCE has a coincident minimum and increases with the SNR on both sides of the interpolation. However, unlike the linear case, the error does not diverge at the interpolation threshold, and the CCE does not vanish there. Instead, the CCE settles at an SNR-dependent floor, suggesting that the learned representations share some information even at interpolation. This raises the question of whether the analytic framework can be extended to account for the nonzero floor. We therefore return to the linear model and ask whether an explicit regularization can reproduce the same effect, in which the networks minimize the MSE loss together with a Layer-wise weight decay penalty,

$$\mathcal{L}_{\text{wd}}(W_1, W_2) = \frac{1}{n}\left[\|W_2 W_1 X^\top - y\|^2 + \omega\left(\|W_1\|_F^2 + \|W_2\|_2^2\right)\right], \tag{19}$$

where ω denotes the weight decay strength.

We computed the correlation $\rho_\delta^*(\alpha)$ at the global minimum, where δ is an effective ridge regularization parameter determined self-consistently by ω. See Appendix F for the derivation and more discussion. We summarize the main takeaway in Fig. 3. Panel (a) shows the generalization error at several δ for SNR$= 5$; panel (b) shows the CCE between representations learned from independent datasets with the same SNR$= 5$. In the ridgeless limit $\delta \to 0$ we recover the theory of Sec. 4.1, whereas for every $\delta > 0$, the correlation stays strictly positive at the interpolation threshold, $\rho_{\delta>0}^*(1) > 0$, explicit regularization removes the alignment collapse. Remarkably, we found that the effective regularization strength δ_{opt} that minimizes the generalization error also maximizes representational alignment (see Appendix F.3 for the proof). We validate this theory, and it is consistent with numerical simulation. We provide the result in Fig 7. For weak regularization, $\delta \ll (2\,\text{SNR})^{-2}$, the correlation at the threshold scales as $\rho_\delta^*(1) \simeq 2\,\text{SNR}\sqrt{\delta}$, so that $\text{CCE} \simeq 4\,\text{SNR}^4\delta^2 > 0$ (Appendix F.4).

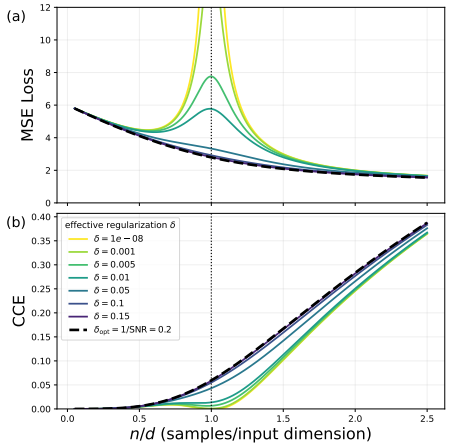

Figure 3: **Effect of regularization.** (a) Generalization error and (b) CCE as a function of n/d for effective ridge regularization $\delta \in [10^{-8}, 0.2]$ at SNR $= 5$. Dashed black: optimal $\delta_{\text{opt}} = 1/\text{SNR} = 0.2$.

Regularization therefore restores an SNR-dependent, nonzero alignment at the threshold. This is consistent with the ReLU network exhibiting a regularization-like effect that suppresses the small-eigenvalue instability responsible for the ridgeless alignment collapse. We do not establish this mechanism analytically for nonlinear networks, but the regularized linear theory predicts the nonzero floor at the interpolation threshold.

Next, we compare the representations learned by the two networks considered so far — one linear, and one with ReLU activation — trained on the same dataset ($d = 200$, $k = 20$, and SNR$= 5$). Panel (a) of Fig. 4 shows the generalization error as a function of γ. The two behave markedly differently: the linear network (orange) achieves a lower error than the ReLU network (blue) everywhere except at its own interpolation threshold, $\gamma = 0.05$, where the error diverges (Hastie et al., 2022; Advani et al., 2020). This follows directly from the structure of the learning problem: the optimal student here is linear by construction, so the linear network operates closer to the Bayes-optimal learner (Bishop & Nasrabadi, 2006; Cui et al., 2023; 2025). Panel (b) shows the opposite behavior. The ReLU representation is more predictive of the linear representation's local neighborhood geometry under the directional CCE, despite the linear network performing better. At $\gamma = 0.05$, the CCE collapses to zero: the linear representation carries no structure, so the two are mutually unpredictable. The CCE has a second, shallower minimum at the interpolation threshold of the ReLU network, $\gamma = 1$, where the network fits the noisy training data (Belkin et al., 2019; Zhang et al., 2016; Mei & Montanari, 2022; d'Ascoli et al., 2020) – yet even there its representations remain more predictive than the linear ones. The asymmetry reflects what the two representations retain: the linear representation better captures the teacher signal, while the ReLU representation preserves more local input geometry, allowing it to better predict the linear representation's neighbourhood structure. Beyond the linear interpolation threshold, we find that the ReLU representation is more informative than the linear representation in predicting the input data (see Fig. 9 in Appendix). This is consistent with the greater expressivity of nonlinear models (Barbier et al., 2025), which allows them to build a richer representational structure.

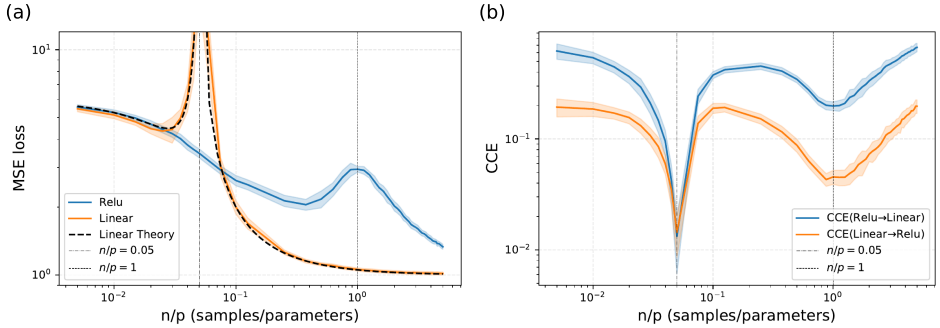

Figure 4: **Linear versus ReLU representations.** (a) Generalization error and (b) directional CCE as a function of n/p at SNR $= 5$. Orange: linear; blue: ReLU. Mean over 20 runs; $d = 200$, $k = 20$.

4.3 Beyond synthetic settings: classification on real-world data

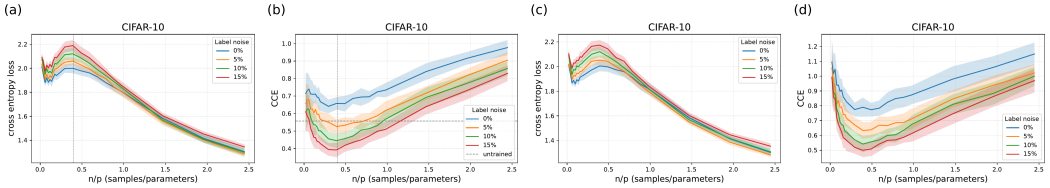

Figure 5: **CIFAR-10 classification.** (a), (c) Generalization error and (b), (d) CCE between penultimate-layer representations as a function of n/p, for different label-noise levels. Networks are trained on non-overlapping datasets from independent initialization in (a), (b), and from a shared initialization in (c), (d).

Finally, we ask whether our findings extend to real-world classification. We consider CIFAR-10 (Krizhevsky et al., 2009) with a standard convolutional neural network (CNN) with $p \approx 10^4$ trainable parameters and 40 feature maps in the last convolutional layer. We train two networks with the same architecture and training protocol on non-overlapping subsets of size n, ranging from 10^2 to half the training set (Appendix H.1), from either independent initializations (Fig. 5 (a),(b)) or a shared initialization (Fig. 5 (c),(d)). We introduce label noise by replacing each training label with a

random one with probability P, independently across examples; the corruption is applied once and held fixed during training, and test labels remain clean. Here P plays the role of inverse SNR.

Panels (a) and (c) show the generalization error as a function of $\gamma = n/p$; the two protocols give nearly identical curves. As on the synthetic data, we recover double descent, with a peak at $\gamma \approx 0.4$ that grows with label noise (Nakkiran et al., 2021). A peak is present even at 0% noise, which may reflect label errors in the original annotations (Northcutt et al., 2021).

Panels (b) and (d) show the CCE between penultimate-layer representations. Under both protocols, representations learned at low label noise align better than those learned at high noise across the whole range of γ, and alignment varies non-monotonically with γ, reaching a minimum near the empirical test loss peak ($\gamma = 0.4$). A shared initialization raises CCE at every γ, most strongly at small n. In panel (b), the dashed lines show the CCE between two untrained, independently initialized networks, i.e., the alignment due to the architecture alone. Trained networks stay close to this level at small γ, fall well below it near the loss peak at 10% and 15% noise, and rise above it at large γ. As in the linear model, alignment increases with SNR and is minimal near the empirical generalization error peak; unlike in the linear model, it does not vanish there, and at small γ it is set largely by the architecture. On MNIST with a fully connected network (shared initialization, Fig. 8), label noise again lowers alignment, but the CCE minimum lies at larger γ than the generalization error peak, and at 0% label noise there is no minimum; this might come from the absence of intrinsic noise in the original labels.

5 Conclusion and Discussion

We have characterized the alignment between the hidden representations of independently trained neural networks using the Conditional Copula Entropy (CCE), a statistic that can be computed analytically in simple but informative reference cases. The picture that emerges is consistent across linear and nonlinear networks: the CCE increases monotonically with the signal-to-noise ratio (SNR) and typically varies non-monotonically with the training sample size, reaching a minimum near the interpolation threshold (Figs. 1, 2, and 5).

We further find that a network can outperform another on the task while learning less informative representations (Figs. 4,-9). Task performance is therefore not a proxy for how much local input-neighborhood structure a representation retains, as quantified by CCE. Our result is also consistent with recent work on nonlinear autoencoders, which shows that the network can incur a higher reconstruction loss than linear methods while learning more complete representations (Mendes et al., 2026). Directional differences in representational predictability cannot be captured by symmetric measures such as CKA, which assign a single similarity value to the pair. This highlights the complementary value of asymmetric alignment measures for characterizing directional relationships between representations (Bansal et al., 2021; Sucholutsky et al., 2025; Umar et al., 2026; Acevedo et al., 2025).

Limitations and Future Directions. We have focused on the single-modality setting, whereas representational convergence has also been reported across networks trained on different modalities, such as vision and language (Huh et al., 2024; Acevedo et al., 2025; Gröger et al., 2026); extending the framework in that direction is a natural next step. A second direction is to push the theory beyond the linear case, which would test whether the universality we observe empirically has a principled explanation and would expose the roles of depth and hidden-layer width — both invisible in our linear results, where the network is functionally equivalent to a single linear map. Finally, we have characterized alignment only at convergence; how it develops over the course of training remains open. Despite these limitations, our results clarify the structure of learned representations across learning regimes and provide a starting point for understanding why independently trained networks converge to similar ones.

AI use statement

In this work, we used generative AI tools to rephrase and edit text, to check mathematical computations for consistency, and to obtain critical feedback on drafts, including on methodology, experiments, the interpretation of results and related literature. We have not used generative AI tools to develop

theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, generate code, assist in writing proofs. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Santiago Acevedo, Andrea Mascaretti, Riccardo Rende, Matéo Mahaut, Marco Baroni, and Alessandro Laio. A quantitative analysis of semantic information in deep representations of text and images. *arXiv preprint arXiv:2505.17101*, 2025.

Ben Adlam and Jeffrey Pennington. Understanding double descent requires a fine-grained bias-variance decomposition. *Advances in neural information processing systems*, 33:11022–11032, 2020.

Madhu Advani and Surya Ganguli. An equivalence between high dimensional bayes optimal inference and m-estimation. *Advances in Neural Information Processing Systems*, 29, 2016a.

Madhu Advani and Surya Ganguli. Statistical mechanics of optimal convex inference in high dimensions. *Physical Review X*, 6(3):031034, 2016b.

Madhu S Advani, Andrew M Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in neural information processing systems*, 32, 2019.

Zhidong Bai, Jack W Silverstein, et al. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. *Advances in neural information processing systems*, 34:225–236, 2021.

Jean Barbier, Francesco Camilli, Minh-Toan Nguyen, Mauro Pastore, and Rudy Skerk. Statistical physics of deep learning: Optimal learning of a multi-layer perceptron near interpolation. *arXiv preprint arXiv:2510.24616*, 2025.

Jan Beirlant, Edward J Dudewicz, László Györfi, Edward C Van der Meulen, et al. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6(1):17–39, 1997.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Lukas Braun, Erin Grant, and Andrew M Saxe. Not all solutions are created equal: An analytical dissociation of functional and representational similarity in deep linear neural networks. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=YucuAuXMpT.

George Casella and Roger Berger. *Statistical inference*. Chapman and Hall/CRC, 2024.

Kyunghyun Cho, Bart Van Merriënboer, Çağlar Gulçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1724–1734, 2014.

Hugo Cui, Florent Krzakala, and Lenka Zdeborova. Bayes-optimal learning of deep random networks of extensive-width. In *International Conference on Machine Learning*, pp. 6468–6521. PMLR, 2023.

Hugo Cui, Florent Krzakala, and Lenka Zdeborová. Bayes-optimal learning of deep random networks of extensive-width. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(1):014001, 2025.

Vittorio Del Tatto, Gianfranco Fortunato, Domenica Bueti, and Alessandro Laio. Robust inference of causality in high-dimensional dynamical processes from the information imbalance of distance ranks. *Proceedings of the National Academy of Sciences*, 121(19):e2317256121, 2024.

Clémentine Carla Juliette Dominé, Nicolas Anguita, Alexandra Maria Proca, Lukas Braun, Daniel Kunin, Pedro A. M. Mediano, and Andrew M Saxe. From lazy to rich: Exact learning dynamics in deep linear networks. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ZXaocmXc6d.

Clémentine CJ Dominé, Nicolas Anguita, Alexandra M Proca, Lukas Braun, Daniel Kunin, Pedro AM Mediano, and Andrew M Saxe. From lazy to rich: Exact learning dynamics in deep linear networks. *arXiv preprint arXiv:2409.14623*, 2024.

Stéphane d'Ascoli, Maria Refinetti, Giulio Biroli, and Florent Krzakala. Double trouble in double descent: Bias and variance (s) in the lazy regime. In *International Conference on Machine Learning*, pp. 2280–2290. PMLR, 2020.

Mario Geiger, Arthur Jacot, Stefano Spigler, Franck Gabriel, Levent Sagun, Stéphane d'Ascoli, Giulio Biroli, Clément Hongler, and Matthieu Wyart. Scaling description of generalization with number of parameters in deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(2):023401, 2020.

Aldo Glielmo, Claudio Zeni, Bingqing Cheng, Gábor Csányi, and Alessandro Laio. Ranking the information content of distance measures. *PNAS nexus*, 1(2):pgac039, 2022.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the platonic representation hypothesis: An aristotelian view. *arXiv preprint arXiv:2602.14486*, 2026.

Yufei Gu, Xiaoqing Zheng, and Tomaso Aste. Unraveling the enigma of double descent: An in-depth analysis through the lens of learned feature space. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=CEkIyshNbC.

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

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Devon Jarvis, Sebastian Lee, Clémentine Carla Juliette Dominé, Andrew M Saxe, and Stefano Sarao Mannelli. A theory of initialisation's impact on specialisation. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(11):114001, 2025.

Yiding Jiang, Vaishnavh Nagarajan, Christina Baek, and J Zico Kolter. Assessing generalization of sgd via disagreement. *arXiv preprint arXiv:2106.13799*, 2021.

Dimitris Kalimeris, Gal Kaplun, Preetum Nakkiran, Benjamin Edelman, Tristan Yang, Boaz Barak, and Haofeng Zhang. Sgd on neural networks learns functions of increasing complexity. *Advances in neural information processing systems*, 32, 2019.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9):1–52, 2025.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/kornblith19a.html`.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 69(6):066138, 2004.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Anders Krogh and John A Hertz. Generalization in a linear perceptron in the presence of noise. *Journal of Physics A: Mathematical and General*, 25(5):1135–1147, 1992.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent learning: Do different neural networks learn the same representations? *arXiv preprint arXiv:1511.07543*, 2015.

Bruno Loureiro, Cedric Gerbelot, Maria Refinetti, Gabriele Sicuro, and Florent Krzakala. Fluctuations, bias, variance & ensemble of learners: Exact asymptotics for convex losses in high-dimension. In *International conference on machine learning*, pp. 14283–14314. PMLR, 2022.

Vladimir A Marčenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75 (4):667–766, 2022.

Vicente Conde Mendes, Lorenzo Bardone, Cédric Koller, Jorge Medina Moreira, Vittorio Erba, Emanuele Troiani, and Lenka Zdeborová. A solvable high-dimensional model where nonlinear autoencoders learn structure invisible to pca while test loss misaligns with generalization. *arXiv preprint arXiv:2602.10680*, 2026.

Preetum Nakkiran and Yamini Bansal. Distributional generalization: A new kind of generalization. *arXiv preprint arXiv:2009.08092*, 2020.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

Roger B Nelsen. *An introduction to copulas*. Springer, 2006.

Curtis G Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*, 2021.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.

Maria Refinetti, Alessandro Ingrosso, and Sebastian Goldt. Neural networks trained with sgd learn distributions of increasing complexity. In *International Conference on Machine Learning*, pp. 28843–28863. PMLR, 2023.

Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

Andrew M Saxe, James L McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23): 11537–11546, 2019.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Christopher J Cueva, Erin Grant, Iris Groen, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine Hermann, Kerem Oktar, Klaus Greff, Martin N Hebart, Nathan Cloos, Nikolaus Kriegeskorte, Nori Jacoby, Qiuyi Zhang, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas O'Connell, Thomas Unterthiner, Andrew Kyle Lampinen, Klaus Robert Muller, Mariya Toneva, and Thomas L. Griffiths. Getting aligned on representational alignment. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=Hiq7lUh4Yn. Survey Certification, Expert Certification.

Ali Hussaini Umar, Franky Kevin Nando Tezoh, Jean Barbier, Santiago Acevedo, and Alessandro Laio. The effect of label noise on the information content of neural representations. *Frontiers in Physics*, 14:1717253, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Alex H Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. Generalized shape metrics on neural representations. *Advances in neural information processing systems*, 34:4738–4750, 2021.

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pp. 3635–3673. PMLR, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

A Related Work

Quantifying when two information processing systems represent information in compatible ways is a long-standing question, originating in cognitive neuroscience with Representational Similarity Analysis (Kriegeskorte et al., 2008), which compares systems through their pairwise dissimilarity matrices rather than raw activations. Modern measures developed for deep networks — Singular Vector Canonical Correlation Analysis (SVCCA) (Raghu et al., 2017), Centered Kernel Alignment (CKA) (Kornblith et al., 2019), and shape metrics based on Procrustes distances (Williams et al., 2021) — extend this idea while imposing different invariance assumptions (orthogonal, affine, or invertible linear). A growing body of empirical work shows that independently trained networks often arrive at structurally similar representations (Li et al., 2015; Raghu et al., 2017; Kornblith et al., 2019; Bansal et al., 2021), a phenomenon recently described as the Platonic Representation Hypothesis (Huh et al., 2024), which conjectures that independently trained deep networks—despite being trained

with different objectives and on different data—tend to learn aligned/similar internal representations that converge to a universal statistical model of reality. A subsequent re-evaluation (Gröger et al., 2026) argues that, after appropriate calibration, this convergence is most robustly visible in local neighborhood structure rather than in global geometry, precisely the regime our CCE statistic is designed to probe.

B About CCE

B.1 Conditional Copula Entropy

The information content between the pairwise distance variables (d_x and d_y) of the representation spaces x and y is measured by the mutual information $\mathrm{I}(d_x, d_y)$. However, mutual information will not be useful in quantifying how well a certain representation space predicts a geometric property of another representation space due to its symmetric nature. Del Tatto et al. (2024) proposed an asymmetric information-theoretic quantity called restricted mutual information, which is defined as

$$\mathrm{I}^{\varepsilon}(d_x \to d_y) = \int_0^{\eta} \int_0^{\infty} dd_x dd_y p(d_x, d_y) \log\left(\frac{p(d_x, d_y)}{p(d_x)p(d_y)}\right)$$

where $\eta = F_x^{-1}(\varepsilon)$ and $F_x(r) = \int_0^r p(d_x) dd_x$ denote the cumulative distribution function of d_x. This quantity is proposed based on the intuition that a good distance space should be locally semantically meaningful. i.e, a space where a data point is enclosed by points that are "similar". Therefore, in the limit of small ε, the restricted mutual information quantifies the shared information by the random variables d_x and d_y when conditioned on the events $d_x = \eta$. With the assumption that large distances are not informative, the authors (Del Tatto et al. (2024)) interpret this measure as the actual information in d_y that is contained in d_x. By expanding the restricted mutual information in the limit $\varepsilon \to 0$, the first nonzero term is its first derivative around zero:

$$\begin{aligned}
\lim_{\varepsilon\to 0} \frac{\mathrm{I}^{\varepsilon}(d_x \to d_y)}{\varepsilon} &= \frac{\partial}{\partial \varepsilon} \mathrm{I}^{\varepsilon}(d_x; d_y)|_{\varepsilon} \\
&= \frac{1}{p(d_x = \eta)} \frac{\partial}{\partial d_x} \mathrm{I}^{\varepsilon}(d_x; d_y)|_{\eta} \\
&= \frac{1}{p(d_x = \eta)} \frac{\partial}{\partial d_x} \left[\int_0^{c} \int_0^{\infty} dd_x dd_y p(d_x, d_y) \log\left(\frac{p(d_x, d_y)}{p(d_x)p(d_y)}\right) \right] \Bigg|_{\eta} \\
&= \frac{1}{p(d_x = \eta)} \frac{\partial}{\partial d_x} \int_0^{c} dd_x p(d_x) \left[\int_0^{\infty} dd_y p(d_y|d_x) \log\left(\frac{p(d_y|d_x)}{p(d_y)}\right) \right] \Bigg|_{\eta} \\
&= \lim_{\eta\to 0} \int_0^{\infty} dd_y p(d_y|d_x = \eta) \log\left(\frac{p(d_y|d_x = \eta)}{p(d_y)}\right)
\end{aligned}$$

Therefore, in the limit of $\varepsilon \to 0$, the restricted mutual information is given by

$$\lim_{\varepsilon\to 0} \frac{\mathrm{I}^{\varepsilon}(d_x \to d_y)}{\varepsilon} = \lim_{\eta\to 0} D_{\mathrm{KL}}[p(d_y|d_x = \eta) \parallel p(d_y)], \tag{20}$$

where D_{KL} denotes the Kullback-Leibler divergence (a pseudo-distance). This divergence represents the amount of information (in nats) gained about d_y when conditioning on $d_x = \eta$, compared to the marginal distribution of d_y. It quantifies how much the distribution of d_y changes when conditioned on small distances in space x—the greater the change, the more informative space x is with respect to space y. Let

$$c = (c_x, c_y) = (F_x(d_x), F_y(d_y)),$$

where $F_x(d_x)$ and $F_y(d_y)$ are the cumulative densities of random variables d_x and d_y. According to Sklar's theorem (Nelsen, 2006), the distribution of $d = (d_x, d_y)$ can be decomposed as the product of the marginals and the joint copula distribution:

$$P(d_x, d_y) = p_c(c_x, c_y) p(d_x) p(d_y),$$

where $p_c(c_x, c_y)$ is the joint copula distribution. This relationship between the distribution of c and its respective copula distribution allows us to rewrite the restricted mutual information as

$$\lim_{\varepsilon\to 0}\frac{\mathrm{I}^{\varepsilon}(d_x \to d_y)}{\varepsilon} = \lim_{\eta\to 0}\int_0^{\infty} dd_y p(d_y \mid d_x = \eta)\log\left(\frac{p(d_y \mid d_x = \eta)}{p(d_y)}\right) \tag{21}$$

$$= \lim_{\varepsilon\to 0}\int_0^{1} dc_y p_c(c_y \mid c_x = \varepsilon)\log p_c(c_y \mid c_x = \varepsilon) \tag{22}$$

$$= -\lim_{\varepsilon\to 0} H(c_y|c_x = \varepsilon). \tag{23}$$

B.2 Properties of CCE

Importantly, CCE (similarly, II) satisfies the standard invariance properties expected from the representation alignment measures (Kornblith et al., 2019). In particular, it is invariant to isotropic scaling and orthogonal transformations, as it depends only on the distribution of pairwise distances. Formally, for any $c_1, c_2 \in \mathbb{R}$ such that $c_1, c_2 \neq 0$ and orthogonal matrices O_1, O_2,

$$\text{CCE}(c_1 A \to c_2 B) = \text{CCE}(A \to B), \quad \text{CCE}(AO_1 \to BO_2) = \text{CCE}(A \to B).$$

C Deep Linear Network Gradient Flow Equations

For a deep linear network with D hidden layers, continuous-time gradient descent on the mean squared training error yields the following dynamics for each weight matrix W_l:

$$\tau\frac{dW_l}{dt} = \left(\prod_{i=l+1}^{D} W_i\right)^T \left[yX - \left(\prod_{i=1}^{D} W_i\right)X^T X\right]\left(\prod_{i=1}^{l-1} W_i\right)^T, \tag{24}$$

where $\tau = n/2$ is a time constant inversely proportional to the learning rate. These dynamics capture the behavior of gradient descent in the limit of a small learning rate.

Advani et al. (2020) provides a detailed description of these dynamics starting from small random initializations, a setting commonly referred to as the rich learning regime (Woodworth et al., 2020). Based on the observation that gradient descent dynamics in deep linear networks initialized with small random weights extract low rank structure (Saxe et al., 2013), they begin by using a singular value decomposition (SVD) based change of variables. Specifically, they define

$$W_1 = r_2 z(t) V^T,$$
$$W_l = c(t) r_{l+1} r_l^T, \quad \text{for } l = 2, \ldots, D,$$

where V is the matrix of right singular vectors of X, $z(t) \in \mathbb{R}^{1\times d}$ is a vector encoding the time-varying overlap with each principal axis in the input, $r_l \in \mathbb{R}^{d_l \times 1}$ are arbitrary unit-norm vectors that capture the freedom in the network's internal representations, and $c(t)$ is a scalar which encodes the change in representation over time. With these definitions, the total weight matrix $W_{\text{tot}} = \prod_{i=1}^{D} W_i$ simplifies to

$$W_{\text{tot}} = \left(\prod_{i=2}^{D} c(t) r_{i+1} r_i^T\right) r_2 z(t) V^T = u(t) z(t) V^T,$$

where $u(t) = c(t)^{D-1}$ and we set $r_{D+1} = 1$.

Further decomposing the input correlation matrix $X^T X$ and the input–output correlation matrix yX as

$$X^T X = V \Lambda V^T, \qquad yX = \tilde{s} V^T,$$

with $\tilde{s} = z^* \Lambda + \tilde{\epsilon}\Lambda^{1/2}$ and $z^* = w^* V$, the network dynamics reduce to the following system of equations:

$$\tau\frac{du(t)}{dt} = u(t)^{\frac{2D-4}{D-1}}\left(\tilde{s} z(t)^T - u(t) z(t) \Lambda z(t)^T\right), \tag{25}$$

$$\tau\frac{dz(t)}{dt} = u(t)\left(\tilde{s} - u(t) z(t) \Lambda\right). \tag{26}$$

Or in component-wise notation

$$\tau \frac{du(t)}{dt} = u(t)^{\frac{2D-4}{D-1}} \sum_{i=1}^{d} \left(\lambda_i z_i^* z_i(t) - u(t) \lambda_i z_i^2(t) + \lambda^{1/2} \tilde{\epsilon}_i z_i(t) \right), \tag{27}$$

$$\tau \frac{dz_i(t)}{dt} = u(t) \left(\lambda_i z_i^* - u(t) \lambda_i z_i(t) + \lambda^{1/2} \tilde{\epsilon}_i \right) \quad \text{for } i = 1, ..., d. \tag{28}$$

These equations represent a substantial reduction in complexity. While full gradient descent in a D-layer network with d hidden units per layer involves $O(d^2 D)$ parameters, the reduced system requires only $d + 1$ parameters, irrespective of depth and hidden layer width. Despite this dramatic compression, Advani et al. (2020) demonstrates that the reduced dynamics provide an accurate description of the full training trajectory, provided the network is initialized with sufficiently small random weights.

C.1 Gradient Flow fixed points

For a two-layer linear neural network, i.e., $D = 2$, the gradient flow training dynamic on the mean squared training error yields the following dynamics for W_1 and W_2:

$$\tau \frac{dW_1}{dt} = W_2^T \left(yX - W_2 W_1 X^T X \right), \tag{29}$$

$$\tau \frac{dW_2}{dt} = \left(yX - W_2 W_1 X^T X \right) W_1^T, \tag{30}$$

These dynamics of the weight contain two manifolds of fixed points: An unstable fixed point at zero $\mathcal{U}$ and a stable fixed point at the global minimum $\mathcal{S}$, which are given by

$$\mathcal{U} = \{ W_2 = 0, W_1 = 0 \}$$
$$\mathcal{S} = \{ W_2, W_1 \mid W_2 W_1 X^T X = YX \}.$$

Using the reduced network dynamics in equations 26, 25 of z and u

$$\tau \frac{du(t)}{dt} = \left(\tilde{s} z(t)^T - u(t) z(t) \Lambda z(t)^T \right), \tag{31}$$

$$\tau \frac{dz(t)}{dt} = u(t) \left(\tilde{s} - u(t) z(t) \Lambda \right). \tag{32}$$

Let $v(t) = u(t) z(t)$. Then

$$\begin{aligned}
\tau \frac{dv(t)}{dt} &= \tau \left[u(t) \frac{dz(t)}{dt} + z(t) \frac{du(t)}{dt} \right] \\
&= \tau \left[u(t)^2 \tilde{s} - u(t)^3 \Lambda z(t) + \tilde{s} z(t) z(t)^\top z(t) - u(t) \Lambda z(t) z^\top z(t) \right] \\
&= \tau \left[\tilde{s} \left(u(t)^2 I + z(t) z(t)^\top \right) - u(t) \Lambda \left(u(t)^2 I + z(t) z(t)^\top \right) z(t) \right] \\
&= \tau \big(\tilde{s} - v(t) \Lambda \big) \big(u(t)^2 I + z(t)^\top z(t) \big).
\end{aligned}$$

Because the matrix $u(t)^2 I + z(t)^\top z(t)$, which depends on $z(t)$ and $u(t)$ in a nonlinear way, the dynamics remain coupled. As a result, a closed-form solution for the finite time is not feasible. However, at the global minimum / stable fixed point (i.e., when $t \to \infty$), we must have

$$\big(\tilde{s} - v^* \Lambda \big) \big(u^{*2} I + z^{*\top} z^* \big) = 0$$

$$\big(w^* V \Lambda + \tilde{\epsilon} \Lambda^{1/2} - v^* \Lambda \big) \big(u^{*2} I + z^{*\top} z^* \big) = 0,$$

where Λ is the diagonal matrix of eigenvalues of $X^\top X$ and $\tilde{\epsilon} \sim \mathcal{N}(0, \sigma_\epsilon^2 I_d)$. Since $u^{*2} I + z^{*\top} z^*$ is always positive definite, this implies that at the stable fixed point

$$v^* = \big(w^* V \Lambda + \tilde{\epsilon} \Lambda^{1/2} \big) \Lambda^\dagger. \tag{33}$$

where $(\cdot)^\dagger$ denotes the pseudoinverse.

D Asymptotic Generalization Error of Two-layer Linear Neural Network

As shown by Advani et al. (2020), the generalization error of the student network at the global minimum is

$$\mathcal{E}_g := \mathbb{E}_{w^*, x_{\text{test}}, \epsilon}[(y - \hat{y})^2] = \frac{1}{d}\mathbb{E}_{w^*, v}[(w^* V^\top - v)^2] + \sigma_\epsilon^2, \tag{34}$$

where the expectation is over all potential teacher weight w^*, input x_{test} and noise ϵ. Substituting the global minimum in Eq. 34 yields

$$\mathcal{E}_g = \sigma_w^2 + \sigma_\epsilon^2 - \frac{1}{d}\sum_{i=1}^{\text{rank}(X^\top X)} \left(\sigma_w^2 - \frac{\sigma_\epsilon^2}{\lambda_i}\right), \tag{35}$$

where $\{\lambda_i\}$ are the non-zero eigenvalues of $X^\top X$, and σ_w^2, σ_ϵ^2 are the variance of the teacher weight and the noise, respectively. Eq. 35 reveals the critical role of the eigenvalue spectrum of the input covariance matrix. Small eigenvalues degrade generalization, while large eigenvalues improve it. Directions with $\lambda_i = 0$ are never learned: the corresponding components v_i remain at their initialized values throughout training (Advani et al., 2020), which is why the sum in Eq. 35 runs only over directions with non-zero eigenvalues. Consequently, initialization directly affects generalization: if weights are initialized to small values, the contribution of these frozen directions to the output remains negligible.

In the proportional high-dimensional limit $n, d \to \infty$ with $\alpha = n/d$ fixed, the empirical spectral distribution of $X^\top X$ converges to the Marchenko–Pastur law (Marčenko & Pastur, 1967),

$$f_\alpha(\lambda) = \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi\lambda} + (1 - \alpha)^+ \delta(\lambda), \tag{36}$$

where $\lambda_\pm = (1 \pm \sqrt{\alpha})^2$ and $(\cdot)^+ = \max(\cdot, 0)$. The first term is the continuous bulk on $[\lambda_-, \lambda_+]$; the second captures the point mass at zero when $\alpha < 1$. Replacing the empirical average with an integral over f, the generalization error admits the asymptotic form

$$\mathcal{E}_g = \sigma_w^2 + \sigma_\epsilon^2 - \int \left(\sigma_w^2 - \frac{\sigma_\epsilon^2}{\lambda}\right) f_\alpha(\lambda)\, d\lambda. \tag{37}$$

E Asymptotic Alignment

Consider two independently trained linear networks $\hat{y}_i^a = W_2^a W_1^a x_i$ and $\hat{y}_i^b = W_2^b W_1^b x_i$ by gradient descent from small random initializations on independent datasets $\mathcal{D}^a = \{X^a, y^a\}$ and $\mathcal{D}^b = \{X^b, y^b\}$, drawn independently from the same data distribution. Suppose the correlation matrices are diagonalizable with orthogonal singular-vector matrices V^a and V^b, respectively. We adopt the SVD parameterization that is commonly used in theoretical analysis of deep linear networks (Saxe et al., 2013; Advani et al., 2020; Dominé et al., 2024; Jarvis et al., 2025) and express the first-layer weights as

$$W_1^a = r^a z^a V^{a\top} = r^a \tilde{z}^a, \qquad W_1^b = r^b z^b V^{b\top} = r^b \tilde{z}^b.$$

This parameterization expresses that the first-layer weights have a low-rank structure and is inspired by the well-known result from prior work that deep linear networks tend to extract low-rank structures when initialized with small weights (Arora et al., 2019; Saxe et al., 2013), which is a form of implicit inductive bias by the training algorithm (Kalimeris et al., 2019; Refinetti et al., 2023). We notice that for any test inputs x, x', the distance induced by W_1^a satisfies

$$\left\| r^a \tilde{z}^a x - r^a \tilde{z}^a x' \right\|^2 = ||r^a||^2 \left\| \tilde{z}^a (x - x') \right\|^2 = \left\| \tilde{z}^a (x - x') \right\|^2, \tag{38}$$

and analogously for W_1^b. Hence, letting $A = W_1^a X_{\text{test}}^\top$, $B = W_1^b X_{\text{test}}^\top$, the pairwise distance structure of A and B are equivalent to that of the one-dimensional projections $\tilde{A} = \tilde{z}^a X_{\text{test}}^\top$ and $\tilde{B} = \tilde{z}^b X_{\text{test}}^\top$. As a consequence, $\text{CCE}(A \to B) = \text{CCE}(\tilde{A} \to \tilde{B})$. By the closed-form result of Umar et al. (2026) for CCE between Gaussian random variables $U = u^\top x$ and $V = v^\top x$, where $x \sim \mathcal{N}(0, I_d)$ and $u, v \in \mathbb{R}^d$ are fixed vectors. The CCE admits the following closed-form expression

$$\text{CCE}(U \to V) = \text{CCE}(V \to U) = -\frac{1}{2}\log(1 - \rho^2) - \frac{1}{2}\rho^2, \tag{39}$$

where $\rho = \mathbb{E}[u^\top v]/\sqrt{\mathbb{E}[\|u\|^2]\ \mathbb{E}[\|v\|^2]}$ denotes the correlation coefficient between U and V. Therefore, the CCE of A and B depends only on the correlation ρ between $\tilde{A}$ and $\tilde{B}$.

$$\rho = \frac{\mathbb{E}[\tilde{z}^a \cdot \tilde{z}^{b\top}]}{\sqrt{\mathbb{E}[\|\tilde{z}^a\|^2]\ \mathbb{E}[\|\tilde{z}^b\|^2]}}. \tag{40}$$

E.1 Correlation between $\tilde{A}$ and $\tilde{B}$

At the global minimum, $\tilde{z}^{*a}$ takes the form

$$\begin{aligned}\tilde{z}^{*a} &= \frac{1}{u^{*a}}\left(w^* V^a \Lambda^a + \tilde{\epsilon}^a (\Lambda^a)^{1/2}\right)(\Lambda^a)^\dagger V^{a\top} \\ &= \frac{1}{u^{*a}}\left(w^* V^a \Lambda^a (\Lambda^a)^\dagger V^{a\top} + \tilde{\epsilon}^a (\Lambda^a)^{1/2}(\Lambda^a)^\dagger V^{a\top}\right),\end{aligned}$$

where Λ^a is the diagonal matrix of eigenvalues of $X^{a\top}X^a$, $\tilde{\epsilon}^a_i \sim \mathcal{N}(0, \sigma^2_\epsilon)$, and u^{*a} is network a's learned second-layer magnitude (an analogous expression, with u^{*b}, holds for z^{*b}). Let $P_a := V^a \Lambda^a (\Lambda^a)^\dagger V^{a\top}$. If $\alpha = n/d > 1$, $(\Lambda^a)^\dagger = (\Lambda^a)^{-1}$ and $P_a = I_d$. If $\alpha < 1$, $X^{a\top}X^a$ is rank deficient: from the dynamics of the weight only $R_a = \text{rank}(X^{a\top}X^a) = n$ are learned; the remaining $d - R_a$ remains at (small) initialization. and $P_a \neq I_d$ is instead the rank-R_a orthogonal projector onto row(X^a).

Case $\alpha > 1$.

$$\mathbb{E}[\tilde{z}^{*a} \cdot \tilde{z}^{*b\top}] = \frac{1}{u^{*a}u^{*b}}\mathbb{E}\left[\left(w^* + \tilde{\epsilon}^a(\Lambda^a)^{-1/2}V^{a\top}\right)\cdot\left(w^* + \tilde{\epsilon}^b(\Lambda^b)^{-1/2}V^{b\top}\right)^\top\right]. \tag{41}$$

Expanding $\tilde{z}^{*a} \cdot \tilde{z}^{*b\top}$, all the cross terms involving $w^*, \tilde{\epsilon}^a, \tilde{\epsilon}^b$ vanish in expectation by mutual independence. For *any* V^a, V^b,

$$\mathbb{E}[\tilde{z}^{*a} \cdot \tilde{z}^{*b\top}] = \frac{1}{u^{*a}u^{*b}}\mathbb{E}[||w^*||^2] = \frac{d\sigma^2_w}{u^{*a}u^{*b}}. \tag{42}$$

Similarly, using the independence of w^* and $\tilde{\epsilon}^a$ and that right-multiplication by the orthogonal matrix $V^a\top$ preserves the norm

$$\mathbb{E}[||\tilde{z}^{*a}||^2] = \frac{1}{(u^{*a})^2}\mathbb{E}\left[\left(||w^*||^2 + ||\tilde{\epsilon}^a(\Lambda^a)^{-1/2}V^{a\top}||^2\right)\right] \tag{43}$$

$$= \frac{d\sigma^2_w}{(u^{*a})^2} + \frac{\sigma^2_\epsilon}{(u^{*a})^2}\sum_{i=1}^{d}\frac{1}{\lambda_i} \tag{44}$$

where $\{\lambda^a_i\} = \Lambda_{ii}$. i.e the non-zero eigenvalues of $X^{a\top}X^a$. Replacing the empirical average over eigenvalues with an integral against f (Eq. 36) yields

$$\mathbb{E}[||\tilde{z}^{*a}||^2] = \frac{d}{(u^{*a})^2}\left(\sigma^2_w + \frac{\sigma^2_\epsilon}{\alpha - 1}\right), \tag{45}$$

Similarly for $\mathbb{E}[||\tilde{z}^{*b}||^2]$. Therefore,

$$\rho^* = \frac{\mathbb{E}[\tilde{z}^{*a} \cdot \tilde{z}^{*b\top}]}{\sqrt{\mathbb{E}[\|\tilde{z}^{*a}\|^2]\ \mathbb{E}[\|\tilde{z}^{*b}\|^2]}} = \frac{\sigma^2_w}{\sigma^2_w + \frac{\sigma^2_\epsilon}{\alpha-1}} = \frac{\text{SNR}}{\text{SNR} + \frac{1}{\alpha-1}}, \tag{46}$$

where $\text{SNR} = \sigma^2_w/\sigma^2_\epsilon$.

Case $\alpha < 1$.

$$\mathbb{E}[\tilde{z}^{*a} \cdot \tilde{z}^{*b\top}] = \frac{1}{u^{*a}u^{*b}}\mathbb{E}\left[\left(w^* P_a + \tilde{\epsilon}^a(\Lambda^a)^{1/2}(\Lambda^a)^\dagger V^{a\top}\right)\cdot\left(w^* P_b + \tilde{\epsilon}^b(\Lambda^b)^{1/2}(\Lambda^b)^\dagger V^{b\top}\right)^\top\right]. \tag{47}$$

Expanding $\tilde{z}^{*a} \cdot \tilde{z}^{*b\top}$, as before every cross term involving $w^*, \tilde{\epsilon}^a, \tilde{\epsilon}^b$ vanishes in expectation by mutual independence, leaving only

$$\mathbb{E}\big[\tilde{z}^{*a} \cdot \tilde{z}^{*b\top}\big] = \frac{1}{u^{*a}u^{*b}}\mathbb{E}\big[w^* P_a P_b w^{*\top}\big] = \frac{\sigma_w^2}{u^{*a}u^{*b}}\mathbb{E}\big[\mathrm{Tr}(P_a P_b)\big]. \tag{48}$$

Because $X^a \perp X^b$ are independent matrices with i.i.d. Gaussian entries, P_a, P_b are independent Haar-random rank-R projectors and the $\mathbb{E}\big[\mathrm{Tr}(P_a P_b)\big] = R_a R_b/d = n^2/d$, so

$$\mathbb{E}\big[\tilde{z}^{*a} \cdot \tilde{z}^{*b\top}\big] = \frac{n^2\sigma_w^2}{du^{*a}u^{*b}}. \tag{49}$$

Similarly for the expected norm

$$\mathbb{E}\big[||\tilde{z}^a||^2\big] = \frac{1}{(u^{*a})^2}\mathbb{E}\Big[\Big(||w^* P_a||^2 + ||\tilde{\epsilon}^a(\Lambda^a)^{1/2}(\Lambda^a)^\dagger V^{a\top}||^2\Big)\Big] \tag{50}$$

$$= \frac{R_a\sigma_w^2}{(u^{*a})^2} + \frac{\sigma_\epsilon^2}{(u^{*a})^2}\sum_{i=1}^{R_a}\frac{1}{\lambda_i} \tag{51}$$

Replacing the empirical average over eigenvalues with an integral against f (Eq. 36) yields

$$\mathbb{E}\big[||\tilde{z}^{*a}||^2\big] = \frac{R_a\sigma_w^2}{u^{*2}} + \frac{d\sigma_\epsilon^2}{u^{*^2}}\left(\frac{\alpha}{1-\alpha}\right) \tag{52}$$

Similarly for $\mathbb{E}\big[||\tilde{z}^{*b}||^2\big]$ with u^{*b}. Therefore,

$$\rho^* = \frac{\frac{n^2\sigma_w^2}{du^{*a}u^{*b}}}{\frac{n\sigma_w^2}{u^{*a}u^{*b}} + \frac{d\sigma_\epsilon^2}{u^{*a}u^{*b}}\left(\frac{\alpha}{1-\alpha}\right)} = \frac{\alpha^2\sigma_w^2\left(\frac{1}{u^{*a}u^{*b}}\right)}{\left(\alpha\sigma_w^2 + \frac{\alpha\sigma_\epsilon^2}{1-\alpha}\right)\left(\frac{1}{u^{*a}u^{*b}}\right)} = \frac{\alpha\mathrm{SNR}}{\mathrm{SNR} + \frac{1}{1-\alpha}}. \tag{53}$$

E.2 Equivalence with linear CKA

The rank-one structure of the learned representations makes the linear CKA between A and B an explicit function of the same correlation ρ^*. Write $A = W_1^a X_{\mathrm{test}}^\top = r^a\tilde{a}$ with $\tilde{a} = \tilde{z}^a X_{\mathrm{test}}^\top \in \mathbb{R}^{1\times N}$, $N = n_{\mathrm{test}}$, and analogously $B = r^b\tilde{b}$. Since $\|r^a\| = \|r^b\| = 1$, the sample Gram matrices are rank one,

$$K_A = A^\top A = \tilde{a}^\top\tilde{a}, \qquad K_B = B^\top B = \tilde{b}^\top\tilde{b}.$$

Let $H = I_N - \frac{1}{N}\mathbf{1}\mathbf{1}^\top$ be the centering matrix and $\bar{a} = \tilde{a}H$, $\bar{b} = \tilde{b}H$ the centered projections, so that $\widetilde{K}_A = HK_AH = \bar{a}^\top\bar{a}$ and $\widetilde{K}_B = \bar{b}^\top\bar{b}$ remain rank one. For rank-one Gram matrices,

$$\langle\widetilde{K}_A, \widetilde{K}_B\rangle_F = \mathrm{Tr}\big(\bar{a}^\top\bar{a}\,\bar{b}^\top\bar{b}\big) = \big(\bar{a}\cdot\bar{b}\big)^2, \qquad \|\widetilde{K}_A\|_F = \|\bar{a}\|^2,$$

and therefore

$$\mathrm{CKA}(A,B) = \frac{\langle\widetilde{K}_A, \widetilde{K}_B\rangle_F}{\|\widetilde{K}_A\|_F\,\|\widetilde{K}_B\|_F} = \frac{(\bar{a}\cdot\bar{b})^2}{\|\bar{a}\|^2\|\bar{b}\|^2} = \hat{\rho}^{\,2}, \tag{54}$$

the squared empirical Pearson correlation between the two one-dimensional projections. Note that the argument uses only $\|r^a\| = \|r^b\| = 1$ and the rank-one form of W_1^a, W_1^b; it is exactly the low-rank assumption of Theorem 1 that makes the two measures functions of the same scalar. Once the representation spans more than one direction, K_A has rank > 1, the Frobenius inner product no longer collapses to a single squared correlation, and the correspondence fails.

F Layer-wise weight decay Objective

F.1 Global minimum of the loss

For each network, we minimize the layer-wise regularized objective

$$\mathcal{L}_{\mathrm{wd}}(W_1, W_2) = \frac{1}{n}\Big[\|W_2W_1X^\top - y\|^2 + \omega\big(\|W_1\|_F^2 + \|W_2\|_2^2\big)\Big], \tag{55}$$

where $\omega \geq 0$ is the Layer-wise weight decay strength. The continuous-time gradient descent dynamics of the weights are as follows:

$$\tau \frac{dW_1}{dt} = W_2^\top \big(yX - W_2 W_1 X^\top X\big) - \omega W_1, \tag{56}$$

$$\tau \frac{dW_2}{dt} = \big(yX - W_2 W_1 X^\top X\big) W_1^\top - \omega W_2, \tag{57}$$

where $\tau = n/2$ is a time constant inversely proportional to the learning rate. Unlike the case when $\omega = 0$, a non-zero stationary point does not satisfy $\big(yX - W_2 W_1 X^\top X\big) = 0$. Instead, the residual correlation is balanced against the layerwise decay.

Reduction of the dynamics. Diagonalize the empirical covariance and express the input-output correlation in the same basis,

$$X^T X = V \Lambda V^T \quad \text{and} \qquad yX = \tilde{s} V^T, \qquad \tilde{s} = w^* V \Lambda + \tilde{\epsilon}\, \Lambda^{1/2}, \quad \tilde{\epsilon} \sim \mathcal{N}(0, \sigma_\epsilon^2 I_d),$$

and apply the low-rank parameterization

$$W_1 = r z V^T \quad \text{and} \quad W_2 = u r^T, \quad \|r\| = 1,$$

where $z \in \mathbb{R}^{1\times d}, u \in \mathbb{R}$. The end-to-end predictor in the eigenbasis is $v = uz$. Substituting the low-rank ansatz in the dynamics gives

$$\tau \frac{du}{dt} = (\tilde{s} - uz\Lambda) z^T - \omega u, \tag{58}$$

$$\tau \frac{dz}{dt} = u(\tilde{s} - uz\Lambda) - \omega z. \tag{59}$$

Balancing of the two layers: Let $\mathcal{B}(t) = u(t)^2 - \|z(t)\|^2$ be the layer imbalance. From Eqs.58-59 the residual terms cancel exactly and

$$\tau \frac{d\mathcal{B}}{dt} = 2u\tau \frac{du}{dt} - 2z\left(\tau \frac{dz}{dt}\right)^\top = -2\omega(u(t)^2 - \|z(t)\|^2). \tag{60}$$

So that

$$\mathcal{B}(t) = \mathcal{B}(0) \exp\left(-\frac{2\omega t}{\tau}\right). \tag{61}$$

For $\omega > 0$ the imbalance decays exponentially from *any* initialization; in particular every stationary point is exactly balanced,

$$u^{*2} = \|z^*\|^2, \qquad \text{hence} \qquad \|v^*\| = u^* \|z^*\| = u^{*2}. \tag{62}$$

Stationary Solution and effective ridge: Setting Eq. 59 to zero,

$$u^* \tilde{s} - u^{*2} z^* \Lambda - \omega z^* = 0 \quad \Longrightarrow \quad z^* = \frac{u^* \tilde{s}}{u^* 2\Lambda + \omega I}, \tag{63}$$

and the end-to-end predictor $v^* = u^* z^*$ is a ridge estimator,

$$v^* = \frac{\tilde{s}}{\Lambda + \frac{\omega}{u^{*2}} I}. \tag{64}$$

The relation $u^{*2} = ||z^*||^2$ generates an effective ridge-like parameter given by

$$\delta = \frac{\omega}{||v^*||}. \tag{65}$$

Here, δ is determined self-consistently by the norm of the solution through

$$\omega^2 = \delta^2 \sum_i^d \frac{\tilde{s}_i^2}{(\lambda_i + \delta)^2}. \tag{66}$$

F.2 CCE BETWEEN THE REPRESENTATIONS:

Consider two networks $y^a(x)$ and $y^b(x)$ trained independently on data generated by the same teacher but with independent input and noise realizations. Let $\tilde{z}^{*q} = z^{*q}V^{q\top}$, $q \in \{a, b\}$, denote the learned first-layer direction expressed on the input basis. From Eq. 64,

$$\tilde{z}^{*q} = \frac{1}{u^{*}{}^{q}}\left(\frac{\tilde{s}^q}{\Lambda^q + \delta I}\right)V^{q\top} = \frac{1}{u^{*}{}^{q}}\left(\frac{w^* V^q \Lambda^q V^{q\top}}{\Lambda^q + \delta I} + \frac{\epsilon^q V^q (\Lambda^q)^{1/2} V^{q\top}}{\Lambda^q + \delta I}\right). \tag{67}$$

Introducing the signal filter and the noise-driven component,

$$\Pi^q_\delta = V^q\left(\frac{\Lambda^q}{\Lambda^q + \delta I}\right)V^{q\top} \quad \text{and} \quad \xi^q = \frac{\epsilon^q V^q (\Lambda^q)^{1/2} V^{q\top}}{\Lambda^q + \delta I}, \tag{68}$$

we have $\tilde{z}^{*q} = (w^*\Pi^q_\delta + \xi^q)/u^{*q}$ and

$$\mathbb{E}\big[\tilde{z}^{*a} \cdot \tilde{z}^{*b\top}\big] = \frac{1}{u^{*a}u^{*b}}\mathbb{E}\Big[w^*\Pi^a_\delta\Pi^b_\delta w^{*\top} + w^*\Pi^a_\delta\xi^{b\top} + \xi^a\Pi^b_\delta w^{*\top} + \xi^a\xi^{b\top}\Big]. \tag{69}$$

The two noise realizations are independent of each other and of w^*, and have zero mean, so only the signal term survives:

$$\mathbb{E}\big[\tilde{z}^{*a} \cdot \tilde{z}^{*b\top}\big] = \frac{1}{u^{*a}u^{*b}}\mathbb{E}\Big[w^*\Pi^a_\delta\Pi^b_\delta w^{*\top}\Big] = \frac{\sigma_w^2}{u^{*a}u^{*b}}\mathbb{E}\big[\mathrm{Tr}(\Pi^a_\delta\Pi^b_\delta)\big]. \tag{70}$$

Writing $O = V^{a\top}V^b$, the trace can be simply expressed as follows

$$\mathrm{Tr}(\Pi^a_\delta\Pi^b_\delta) = \mathrm{Tr}\left(\left(\frac{\Lambda^a}{\Lambda^a + \delta I}\right)O^{ab}\left(\frac{\Lambda^b}{\Lambda^b + \delta I}\right)O^{ab\top}\right) \tag{71}$$

$$= \sum_{i,j}\left(\frac{\lambda^a_i}{\lambda^a_i + \delta}\right)\left(\frac{\lambda^b_j}{\lambda^b_j + \delta}\right)|O^{ab}_{ij}|^2 \tag{72}$$

Because the inputs are isotropic Gaussian, each $X^{q\top}X^q$ belongs to a rotationally invariant ensemble: V^q is Haar distributed on $O(d)$ and independent of Λ^q. The two networks being independent, $O^{ab} = V^{a\top}V^b$ is again Haar orthogonal, and since its rows are unit vectors with exchangeable entries, $\sum_j \mathbb{E}|O^{ab}_{ij}|^2 = 1$ forces

$$\mathbb{E}[|O^{ab}_{ij}|^2] = \frac{1}{d}. \tag{73}$$

The basis overlap is therefore unbiased and diffuse: no eigendirection of one network is preferentially aligned with any eigendirection of the other, and the double sum factorizes,

$$\mathbb{E}\big[\tilde{z}^{*a} \cdot \tilde{z}^{*b\top}\big] = \frac{\sigma_w^2}{u^{*a}u^{*b}d}\sum_{i,j}\left(\frac{\lambda^a_i}{\lambda^a_i + \delta}\right)\left(\frac{\lambda^b_j}{\lambda^b_j + \delta}\right) \tag{74}$$

$$= \frac{\sigma_w^2 d}{u^{*a}u^{*b}}\left[\frac{1}{d}\sum_i\left(\frac{\lambda^a_i}{\lambda^a_i + \delta}\right)\right]\left[\frac{1}{d}\sum_j\left(\frac{\lambda^b_j}{\lambda^b_j + \delta}\right)\right]. \tag{75}$$

In the proportional high-dimensional limit

$$\mathbb{E}\big[\tilde{z}^{*a} \cdot \tilde{z}^{*b\top}\big] = \frac{\sigma_w^2 d}{u^{*a}u^{*b}}\left[\int\left(\frac{\lambda}{\lambda + \delta}\right)f_\alpha(\lambda)\,d\lambda\right]^2. \tag{76}$$

Similarly, for $q \in \{a, b\}$. The norm of a learned direction is

$$\mathbb{E}\big[||\tilde{z}^{*q}||^2\big] = \frac{1}{(u^{*q})^2}\mathbb{E}\Big[\left(||w^*\Pi^q_\delta||^2 + ||\xi^q||^2\right)\Big] \tag{77}$$

$$= \frac{1}{(u^{*q})^2}\left[\sigma_w^2\sum_i\left(\frac{\lambda^q_i}{\lambda^q_i + \delta}\right)^2 + \sigma_\epsilon^2\sum_i\frac{\lambda^q_i}{(\lambda^q_i + \delta)^2}\right] \tag{78}$$

$$= \frac{d}{(u^{*q})^2}\left[\sigma_w^2\int\left(\frac{\lambda}{\lambda + \delta}\right)^2 f_\alpha(\lambda)d\lambda + \sigma_\epsilon^2\int\frac{\lambda}{(\lambda + \delta)^2}f_\alpha(\lambda)d\lambda\right] \tag{79}$$

Let

$$\mathcal{M}_1(\alpha,\delta) = \int \frac{\lambda}{\lambda+\delta} f_\alpha(\lambda)\, d\lambda,$$
$$\mathcal{M}_2(\alpha,\delta) = \int \left(\frac{\lambda}{\lambda+\delta}\right)^2 f_\alpha(\lambda)\, d\lambda,$$
$$\mathcal{N}(\alpha,\delta) = \int \frac{\lambda}{(\lambda+\delta)^2} f_\alpha(\lambda)\, d\lambda$$

Using the preceding results, the asymptotic cosine correlation is

$$\rho^*_{\text{wd}} = \frac{\mathbb{E}\big[\tilde{z}^{*a}\cdot\tilde{z}^{*b\top}\big]}{\sqrt{\mathbb{E}\big[\|\tilde{z}^{*a}\|^2\big]\ \mathbb{E}\big[\|\tilde{z}^{*b}\|^2\big]}} = \frac{\sigma_w^2\mathcal{M}_1(\alpha,\delta)^2}{\sigma_w^2\mathcal{M}_2(\alpha,\delta)+\sigma_\epsilon^2\mathcal{N}(\alpha,\delta)}. \tag{80}$$

Dividing both the numerator and the denominator by σ_ϵ^2 gives

$$\rho^*_{\text{wd}} = \frac{\text{SNR}\mathcal{M}_1(\alpha,\delta)^2}{\text{SNR}\mathcal{M}_2(\alpha,\delta)+\mathcal{N}(\alpha,\delta)}. \tag{81}$$

The correlation (thus CCE) depends on the regularization only through the effective ridge δ, and hence on ω only through the self-consistent Eq. 66. As $\omega \to 0$, so $\delta \to 0$, the ρ^*_{wd} in Eq. 81 tends to ρ^* in Eq. 10. i.e one recovers the unregularized formulas. Since $A = W_1^a X_{\text{test}}^\top$ and $B = W_1^b X_{\text{test}}^\top$ reduce, in terms of pairwise distances, to one-dimensional Gaussian projections, the CCE between A and B is given by

$$\text{CCE}_{\text{wd}}(\alpha,\delta,\text{SNR}) = -\tfrac{1}{2}\log\big(1-\rho^{*2}_{\text{wd}}\big) - \tfrac{1}{2}\rho^{*2}_{\text{wd}}. \tag{82}$$

F.3 Optimal Regularization

For a test input x_{test}, the expected MSE is

$$\mathcal{E}_g(\alpha,\delta) := \mathbb{E}\left[\big(y(x_{\text{test}}) - vV^\top x_{\text{test}}\big)^2\right] = \sigma_\epsilon^2 + \frac{1}{d}\mathbb{E}||w^* - Vv^\top||^2. \tag{83}$$

Using Eq. 64, the asymptotic result is

$$\mathcal{E}_g(\alpha,\delta) = \sigma_\epsilon^2 + \sigma_w^2\,(1 - 2\mathcal{M}_1(\alpha,\delta) + \mathcal{M}_2(\alpha,\delta)) + \sigma_\epsilon^2\mathcal{N}(\alpha,\delta). \tag{84}$$

Differentiate the generalization error Eq. 84 with respect to the effective parameter δ. One obtains

$$\frac{\partial E_g}{\partial\delta} = 2\left(\sigma_w^2\delta - \sigma_\epsilon^2\right)\int \frac{\lambda}{(\lambda+\delta)^3} f_\alpha(\lambda)\, d\lambda. \tag{85}$$

The integral is non-negative, so optimal generalization under layerwise weight decay occurs when the effective regularization strength is

$$\delta = \frac{\sigma_\epsilon^2}{\sigma_w^2} = \frac{1}{\text{SNR}}. \tag{86}$$

Moreover, it is well known that, under Gaussian noise and parameter distributions, no algorithm can outperform optimally tuned effective layer-wise weight decay in terms of generalization error (Advani & Ganguli, 2016a;b).

The derivative of Eq. 81 with respect to δ can be written as

$$\frac{d\rho^*_{\text{wd}}}{d\delta} = \frac{2\text{SNR}\mathcal{M}_1(1-\text{SNR}\delta)\left(\mathcal{M}_1\mathcal{S}-\mathcal{N}^2\right)}{(\text{SNR}\mathcal{M}_2+\mathcal{N})^2} \tag{87}$$

where

$$\mathcal{S}(\delta) = \int \frac{\lambda}{(\lambda+\delta)^3} f_\alpha(\lambda)\, d\lambda. \tag{88}$$

By Cauchy–Schwarz,

$$\mathcal{M}_1\mathcal{S} - \mathcal{N}^2 \geq 0. \tag{89}$$

Therefore, the representation correlation is also maximized at

$$\delta = \frac{1}{\mathrm{SNR}}.$$

At this value, one has

$$\mathrm{SNR}\mathcal{M}_2 + \mathcal{N} = SNR\mathcal{M}_1. \tag{90}$$

Therefore, the correlation simplifies to

$$\rho^*_{\mathrm{opt}} = \mathcal{M}_1\left(\alpha, \frac{1}{\mathrm{SNR}}\right). \tag{91}$$

Hence the regularization strength that optimizes generalization is exactly the one that maximizes the correlation between independently learned representations: under this model, generalization error and representational alignment are optimized by the same amount of regularization strength.

F.4 Weak-δ behavior of the CCE at the interpolation threshold

At $\alpha = 1$ the MP edges are $\lambda_- = 0$ and $\lambda_+ = 4$, and the limiting density is

$$f_1(\lambda) = \frac{1}{2\pi}\sqrt{\frac{4-\lambda}{\lambda}}, \qquad \lambda \in (0, 4]. \tag{92}$$

The lower edge has reached the origin and the density *diverges* there, $f_1(\lambda) \simeq 1/(\pi\sqrt{\lambda})$ as $\lambda \to 0^+$. This pile-up of arbitrarily small eigenvalues is what makes the unregularized variance blow up at the interpolation threshold and drives the sharp $\mathrm{CCE} = 0$ result of the unregularized theory; a non-zero effective ridge cuts it off at the scale $\lambda \sim \delta$, and the question is how quickly the alignment is restored as δ is switched on at the interpolation threshold.

At $\alpha = 1$ the three spectral integrals are available in closed form. Writing $D(\delta) = \sqrt{\delta(\delta+4)}$,

$$\mathcal{M}_1(1,\delta) = 1 + \frac{\delta}{2} - \frac{D}{2}, \qquad \mathcal{N}(1,\delta) = \frac{\delta+2}{2D} - \frac{1}{2}, \qquad \mathcal{M}_2(1,\delta) = \mathcal{M}_1 - \delta\mathcal{N}. \tag{93}$$

The first follows from the Stieltjes transform of Eq. 92 evaluated at $-\delta$,

$$\mathcal{M}_1 = 1 - \delta\int \frac{f_1(\lambda)}{\lambda+\delta}\, d\lambda = 1 - \delta\,\frac{\sqrt{\delta+4} - \sqrt{\delta}}{2\sqrt{\delta}}, \tag{94}$$

and the other two from $N = -\partial_\delta M_1$ and the following identity

$$\mathcal{M}_1(\alpha,\delta) \;=\; M_2(\alpha,\delta) + \delta\,\mathcal{N}(\alpha,\delta), \tag{95}$$

obtained by writing $\lambda/(\lambda+\delta) = \lambda^2/(\lambda+\delta)^2 + \delta\lambda/(\lambda+\delta)^2$. Expanding Eq. 93 for $\delta \ll 1$,

$$\mathcal{M}_1 = 1 - \sqrt{\delta} + O(\delta), \qquad \mathcal{M}_2 = 1 - \tfrac{3}{2}\sqrt{\delta} + O(\delta), \qquad \mathcal{N} = \frac{1}{2\sqrt{\delta}} - \frac{1}{2} + O(\sqrt{\delta}). \tag{96}$$

The terms $\mathcal{M}_1$ and $\mathcal{M}_2$ stay $O(1)$ while the term $\mathcal{N}$ *diverges* as $\delta^{-1/2}$: the square-root singularity of Eq. 92 integrated against $\lambda(\lambda+\delta)^{-2}$ produces $\pi^{-1}\int_0^\infty \sqrt{\lambda}(\lambda+\delta)^{-2}d\lambda = 1/(2\sqrt{\delta})$. Substituting Eq. 96 into Eq. 81, the denominator is dominated by $\mathcal{N}$ as long as $\mathcal{N} \gg \mathrm{SNR}\,\mathcal{M}_2$, i.e. $\sqrt{\delta} \ll (2\,\mathrm{SNR})^{-1}$, and in that regime

$$\rho^*_{\mathrm{wd}}(1,\delta) = \frac{\mathrm{SNR}\,\mathcal{M}_1^2}{\mathrm{SNR}\,\mathcal{M}_2 + \mathcal{N}} = 2\,\mathrm{SNR}\sqrt{\delta}\left[1 - (1+2\,\mathrm{SNR})\sqrt{\delta} + O(\delta)\right] \;\simeq\; 2\,\mathrm{SNR}\sqrt{\delta}. \tag{97}$$

For small correlation the CCE admits the expansion

$$\mathrm{CCE}(\rho) = -\tfrac{1}{2}\log(1-\rho^2) - \tfrac{1}{2}\rho^2 = \frac{\rho^4}{4} + \frac{\rho^6}{6} + O(\rho^8), \tag{98}$$

obtained from $-\frac{1}{2}\log(1-x) = \frac{x}{2} + \frac{x^2}{4} + \frac{x^3}{6} + \cdots$ with $x = \rho^2$: the leading $x/2 = \rho^2/2$ cancels exactly against the subtracted $-\rho^2/2$, leaving a quartic leading order. Combining Eq. 98 with Eq. 97,

$$\mathrm{CCE}\big(\rho^*_{\mathrm{wd}}(1,\delta)\big) \;\simeq\; \frac{\big(2\,\mathrm{SNR}\sqrt{\delta}\big)^4}{4} \;=\; 4\,\mathrm{SNR}^4\,\delta^2. \tag{99}$$

The alignment at the threshold is therefore strictly positive for any $\delta > 0$, and it approaches zero *quadratically* in δ with a *fourth-power* dependence on SNR. This weak-δ window, $\delta \ll (2\,\mathrm{SNR})^{-2}$, lies well below the optimum $\delta_{\mathrm{opt}} = 1/\mathrm{SNR}$, where the alignment is instead $\rho^*_{\mathrm{opt}} = M_1(1, 1/\mathrm{SNR}) = O(1)$.

F.5 RESULT AS FUNCTION OF ω

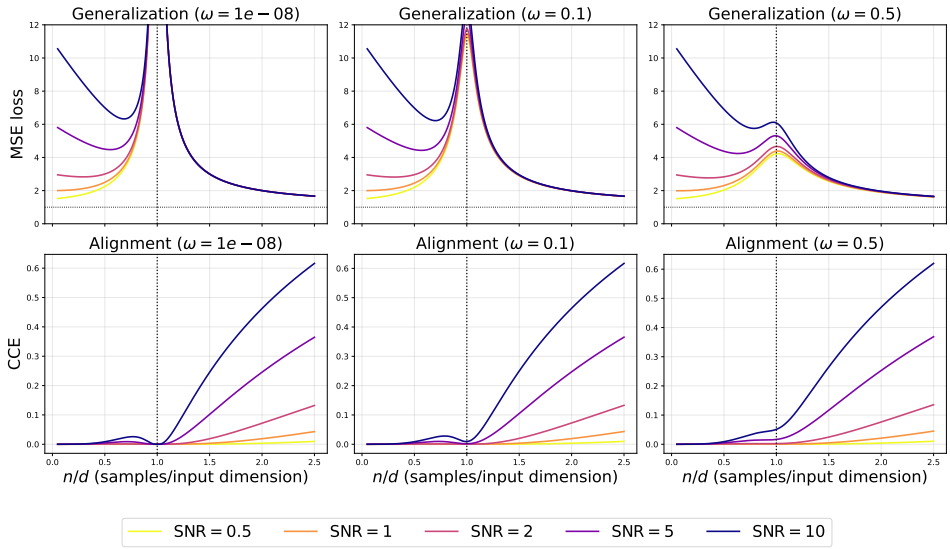

Figure 6: Effect of weight decay on generalization error (top) and CCE (bottom) as a function of α, for $\omega \in \{10^{-8}, 0.1, 0.5\}$ (each in a separate column) and SNR$\in \{0.5, 1, 2, 5, 10\}$.

Figure 6 shows generalization error and CCE for $\omega \in \{10^{-8}, 0.1, 0.5\}$ and SNR$\in \{0.5, 1, 2, 5, 10\}$. The generalization error peak at $\alpha = 1$ drops rapidly with ω, and the CCE curve is lifted off in a neighborhood of $\alpha = 1$.

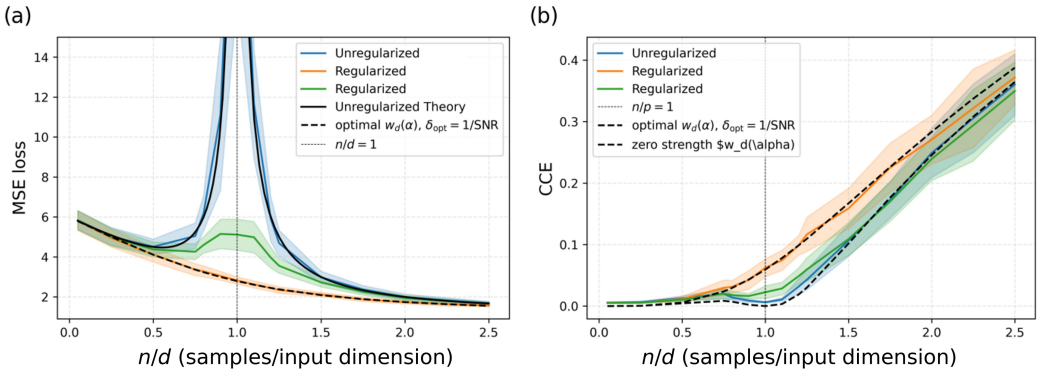

Figure 7: **Layer-wise theory versus Simulations.** (a) Generalization error and (b) CCE as a function of n/d at SNR $= 5$. Orange: optimal effective regularization $\delta_{\text{opt}} = 1/\text{SNR}$ with $\omega_{\text{opt}}(\delta_{\text{opt}}, \alpha)$ computed apriori from Eq 66; green: $\omega = 0.5$, blue: zero regularization strength. Mean over 20 runs; $d = 200$, $k = 20$

G ALIGNMENT AS STUDENT–TEACHER OVERLAP

Since $\tilde{z}^{*q}$ is the end-to-end predictor divided by the scalar u^{*q}, the scalar cancels in all cosines.

For $\alpha > 1$,

$\mathbb{E}[\tilde{z}^{*q} \cdot w^{*\top}] = d\sigma_w^2/u^{*q}$, and Eq. 45 gives

$$\cos^2(\tilde{z}^{*q}, w^*) = \frac{(d\sigma_w^2)^2}{d\sigma_w^2\, d\left(\sigma_w^2 + \frac{\sigma_\epsilon^2}{\alpha - 1}\right)} = \frac{\text{SNR}}{\text{SNR} + \frac{1}{\alpha - 1}} = \rho^*.$$

For $\alpha < 1$, $\mathbb{E}[\tilde{z}^{*q} \cdot w^{*\top}] = \sigma_w^2 \, \mathbb{E}\,\mathrm{Tr}(P^q)/u^{*q} = n\sigma_w^2/u^{*q}$, and Eq. 52 gives

$$\cos^2(\tilde{z}^{*q}, w^*) = \frac{\alpha \mathrm{SNR}}{(\mathrm{SNR} + \frac{1}{1-\alpha})} = \rho^*.$$

For the ridge solutions of Appendix F, $\mathbb{E}[\tilde{z}^{*q} \cdot w^{*\top}] = d\sigma_w^2 \mathcal{M}_1/u^{*q}$, and Eq. 79 gives

$$\cos^2(\tilde{z}^{*q}, w^*) = \mathrm{SNR}\,\mathcal{M}_1^2/(\mathrm{SNR}\,\mathcal{M}_2 + \mathcal{N}) = \rho^*_{\mathrm{wd}}.$$

In every case the components of $\tilde{z}^{*a}$ and $\tilde{z}^{*b}$ orthogonal to w^* are uncorrelated, so

$$\rho^* = \cos(\tilde{z}^{*a}, w^*)\cos(\tilde{z}^{*b}, w^*).$$

H Classification Task

H.1 Experimental Setup

We performed classification experiments using two baseline neural-network architectures: a fully connected neural network (FCNN) and a standard convolutional neural network (CNN). The two networks have the following structure:

1. A single-layer FCNN with k hidden units, with input and output dimensions fixed to match the number of pixels in the input image and the number of classes in the task, respectively.
2. A family of CNNs consisting of 4 convolutional stages of width $[k, 2k, 4k, 8k]$, where k is the width parameter, followed by a fully connected layer as a classifier. The MaxPool is $[2, 2, 2, 4]$. For the entire convolution layer, the kernel size $= 3$, stride $= 1$, and padding $= 1$. This architecture is identical to the one considered in (Gu et al., 2024)

The networks are trained using the mini-batch SGD algorithm to minimize the cross entropy loss. A batch size of 128 samples is used. For FCNNs, we set the learning rate to start with $l_0 = 0.1$, and after every successive 50 epochs it changes to $l_r = l_0/(\sqrt{1 + epoch/50})$. While for CNNs, the learning rate starts with $l_0 = 0.05$, and updates as $l_r = l_0/(\sqrt{1 + epoch \times 50})$ after every 50 epoch. This learning rate scheme was used in a recent paper by (Gu et al., 2024). The FCNN is trained for image classification tasks with the MNIST dataset, and the CNN is trained for image classification tasks with the CIFAR-10 dataset. For both datasets, we partition all the training data into two equal parts. i.e., we have two sets of training data, for MNIST each set has 3×10^4 samples. While for CIFAR-10, each set has 25×10^3 samples.

H.2 Results on Mnist

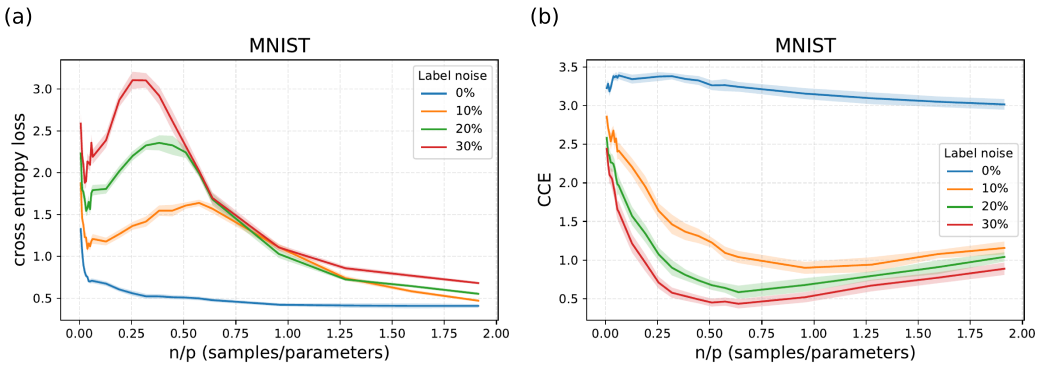

Figure 8: **MNIST classification.** (a) Generalization error and (b) CCE between penultimate-layer representations as a function of n/p, for different label-noise levels. Networks are trained on non-overlapping datasets from the same initialization.

I Supportive Figures

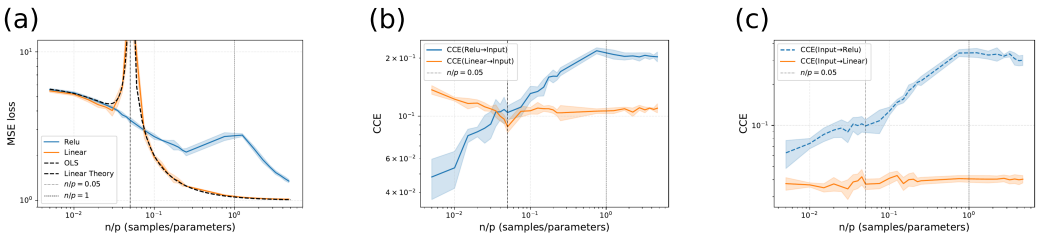

Figure 9: **Input versus Linear/ReLU representations.** (a) Generalization error and (b), (c) directional CCE as a function of n/p at SNR $= 5$. Orange: linear; blue: ReLU. Mean over 5 runs; $d = 200$, $k = 20$.