Redundancy as a Structural Information Principle for Learning and Generalization

Anonymous authors
Paper under double-blind review

Abstract

We present a theoretical paradigm that extends classical information theory to finite and structured systems by redefining *redundancy* as a fundamental quantity of information organization rather than inefficiency. Within an f-divergence framework, redundancy is formalized as $\mathcal{R}_f(X) = D_f(P_X \| \Pi_X) = \mathbb{E}_{\Pi_X}\big[f\big(\frac{p(x)}{\prod_i p_i(x_i)}\big)\big]$, where $p(x)$ is the joint density of $(X_1, \ldots, X_n)$, $p_i(x_i)$ their marginals, and f a convex kernel defining the geometry of informational dependence. Different choices of f recover mutual information, χ^2 redundancy, and spectral redundancy as special cases, unifying diverse notions under a single mathematical principle. This reveals that classical measures are not isolated heuristics but projections of a single redundancy geometry. The framework shows that redundancy is bounded both above and below, yielding a natural equilibrium R^* between over-compression (loss of structure) and over-coupling (collapse). In contrast to the asymptotic regime where minimizing redundancy optimizes transmission efficiency, finite, structured systems—where real-world learning operates—achieve maximal stability and generalization near this equilibrium. Thus, redundancy emerges as a *structural information principle*: a self-organizing property that governs how information is coherently structured rather than transmitted. Experiments with masked autoencoders (MAE) serve to *verify and visualize* the theory rather than pursue performance benchmarks. They confirm the predicted equilibrium R^*, where latent redundancy stabilizes and generalization peaks. Together, these results establish redundancy as a measurable and tunable quantity bridging the asymptotic world of communication and the finite world of learning.

1 Introduction

Redundancy is one of the oldest yet least unified concepts in the quantitative sciences. From Shannon's original formulation of information theory in the 1940s, redundancy has been viewed primarily as an inefficiency—a measure of wasted bits relative to the entropy limit of a code Shannon (1948). In neuroscience and sensory coding, however, the term acquired an almost opposite meaning: redundant neural firing or overlapping receptive fields were interpreted as signatures of robustness and fault tolerance in biological systems Barlow (2001); Narayanan et al. (2005); Tononi et al. (1994) with methods often designed to leverage redundancy explicitly Kazemivash et al. (2025). In statistics and machine learning, redundancy appears again under diverse guises—multicollinearity in regression Gunst & Webster (1975), overparameterization in neural networks Zhang et al. (2016); Belkin et al. (2019), and correlated features in representation learning—each carrying context-dependent interpretations of inefficiency, degeneracy, or regularization Li et al. (2023); Berchenko (2024); Allen-Zhu et al. (2019); Chan et al. (2022). In physics and complex systems, redundant degrees of freedom are sometimes regarded as the very basis of stability and emergence, echoing the notion that structure arises from constrained repetitions rather than minimal codes May (1972); Lloyd (2007); Prigogine & Hiebert (1982). Despite its ubiquity, these perspectives remain largely fragmented, and no single mathematical framework consistently links redundancy across these domains. Historically, classical information theory and signal processing treated redundancy as something to be eliminated. In the Shannon framework, an optimal code is one that minimizes redundancy, achieving the highest information rate per transmitted symbol Shannon (1948). Source coding theorems and compression algorithms were

built precisely on this ideal—removing repeated or correlated patterns to approach the entropy limit of a signal Huffman (2007); Ziv & Lempel (2003). In statistics and data representation, the same logic underlies many foundational techniques: principal component analysis (PCA) seeks orthogonal directions of maximal variance to eliminate correlated features; independent component analysis (ICA) explicitly minimizes statistical dependence between latent sources Hyvärinen & Oja (2000); Bell & Sejnowski (1997); Tishby et al. (2000); and sparse or efficient coding models in neuroscience posit that sensory neurons should represent the world with minimal overlap or redundancy among firing patterns Olshausen & Field (1996). Across these traditions, redundancy has been equated with inefficiency, overfitting, or wasted capacity—a byproduct to be compressed away rather than a structural property to be understood.

Yet in finite, noisy, and structured regimes, the classical view of redundancy as inefficiency breaks down. We propose a paradigm shift that extends Shannon's information theory from asymptotic coding efficiency to finite-sample informational organization. In this new framework, redundancy is not the waste of bits, but the *geometry of informational dependence*—a structural degree of freedom through which data organize meaning, stability, and generalization.

Formally, we define redundancy as an f-divergence from statistical independence, $\mathcal{R}_f(X) = D_f(P_X \| \Pi_X) = \mathbb{E}_{\Pi_X}\big[f\big(\frac{p(x)}{\prod_i p_i(x_i)}\big)\big]$, where f specifies the geometric kernel of informational coupling. This single functional family $\{\mathcal{R}_f\}$ unifies diverse notions of redundancy across disciplines as its different projections: in *information theory*, $\mathcal{R}_{\mathrm{KL}}$ corresponds to mutual information and total correlation (entropy-based projection); in *statistics*, $\mathcal{R}_{\chi^2}$ captures covariance redundancy (second-order projection); in *neuroscience and complex systems*, spectral redundancy $\mathcal{R}_{\mathrm{spec}}$ describes degeneracy and shared variance (eigen-structural projection). All are manifestations of the same underlying quantity that measures the departure of data from independence—the structure that sustains meaning in finite systems Zollikofer et al. (2024).

Within this unified geometry, redundancy is shown to be bounded both above and below, giving rise to an intrinsic equilibrium R^* between over-compression (loss of structure) and over-coupling (collapse). This equilibrium complements Shannon's asymptotic ideal: while minimizing redundancy maximizes channel efficiency, maintaining an optimal level of redundancy enhances stability and generalization in structured, finite regimes.

More broadly, the framework provides a theoretical bridge between generative and discriminative paradigms of self-supervised learning Nanda et al. (2023). In generative pretraining, such as masked autoencoders (MAE) or diffusion-based foundation models Ho et al. (2020); Rombach et al. (2022); He et al. (2024); Croitoru et al. (2023), understanding emerges from the organization of data redundancy; generalization improves when redundancy is balanced rather than minimized. In contrast, discriminative or contrastive objectives, such as SimCLR or MoCo Chen et al. (2020); He et al. (2020), still tend to minimize redundancy, emphasizing separability over structure. This dichotomy highlights redundancy as a controllable variable that governs how information is structured, transferred, and understood across learning paradigms.

To illustrate this principle, we examine the generative regime through masked autoencoders (MAE). The experiments show that model understanding and generalization both peak when latent redundancy fluctuates around the theoretical equilibrium R^*. These experiments are not intended for performance benchmarking but to *visualize and verify* the theoretical prediction that generative self-supervised learning achieves its most coherent internal organization at an optimal redundancy balance. We thus establish redundancy as a *structural information principle*—a quantitative law unifying its projections across fields and revealing its equilibrium as the organizing center of modern representation learning.

The remainder of this paper is organized as follows. Section 3 introduces the unified redundancy framework, defining $\mathcal{R}_f$ as a general f-divergence from independence and establishing its fundamental properties, including nonnegativity, the data-processing inequality, and Gaussian or spectral reductions. Section 3.4 develops the redundancy–loss equilibrium theory, proving the existence of an interior optimum R^* and analyzing its stability under stochastic learning dynamics. Section 3.7 presents the information–theoretic "sandwich" argument, demonstrating that both insufficient and excessive redundancy impair performance, whereas a natural balance emerges at R^*. Section 3.5 provides empirical validation in masked autoencoders (MAE), confirming that model performance peaks near the theoretical optimum, and Appendix A contains com-

plete mathematical proofs. Together, these sections establish a unified theory of redundancy that bridges information theory, statistical learning, and complex systems—elevating redundancy from a byproduct of compression to a governing principle of organization.

Table 1: A cross-disciplinary summary of how redundancy has been defined, interpreted, and utilized across different scientific domains. Our framework unifies these views by treating redundancy as a measurable and optimizable property of data organization.

Field / Domain	Classical view of redundancy	Recent perspective	Representative works
Information theory	Redundancy reduces channel capacity and should be minimized for efficient coding.	Redundancy ensures reliability, error correction, and structured transmission in noisy environments.	Shannon (1948); Huffman (1952); Ziv & Lempel (1977).
Neuroscience / Sensory coding	Redundant firing viewed as inefficiency in neural encoding.	Overlapping receptive fields increase robustness, fault tolerance, and predictive stability.	Barlow (2001); Narayanan (2005); Kazemivash et al. (2025).
Statistics / Machine learning	Multicollinearity and overparameterization seen as overfitting or inefficiency.	Redundant features improve generalization and stability; redundancy becomes a tunable regularization target.	Zhang et al. (2016); Belkin et al. (2019); Li et al. (2023).
Physics / Complex systems	Redundancy implies degeneracy and non-essential degrees of freedom.	Redundancy enables stability, emergence, and robustness in high-dimensional systems.	May (1972); Prigogine (1982); Lloyd (2007).
Deep representation learning	Redundancy reduction as compression objective.	Networks self-organize redundant subspaces that enhance generalization and robustness.	Doimo et al. (2022); Nanda et al. (2023); Wollstadt et al. (2023).

2 Related Works

Redundancy has re-emerged as a central theme in recent studies of representation learning and overparameterized networks. Empirical analyses have shown that deep models often develop *structured repetition* rather than mere inefficiency. Doimo *et al.* Doimo et al. (2022) demonstrated that wide neural networks naturally form groups of nearly identical neurons whose activations differ only by small perturbations, offering a mechanistic account of "benign overfitting." Similarly, Nanda *et al.* Nanda et al. (2023) reported *diffuse redundancy* in pretrained representations: randomly subsampling neurons from intermediate layers hardly degrades downstream accuracy, indicating that information is distributed across many interchangeable units. Parallel work in information-theoretic feature analysis, such as Wollstadt *et al.* Wollstadt et al. (2023), reached similar conclusions through the lens of partial information decomposition (PID), formally partitioning predictive information into unique, synergistic, and redundant components. Together, these studies suggest that redundancy is not a statistical artifact to be removed, but an intrinsic property of high-dimensional representations that contributes to robustness and generalization.

Building on these empirical insights, subsequent research has sought to *measure*, *control*, and *organize* redundancy in a principled way. The **RedTest** framework Lu et al. (2024) provides a large-scale empirical methodology for quantifying neuron- and layer-level redundancy, revealing its correlation with robustness and energy efficiency. Yavuz and Yanikoglu Yavuz & Yanikoglu (2025) proposed a coupling-matrix-based redundancy index $\rho(C)$ that captures latent-space correlations as a proxy for internal representational geometry, reinforcing the view that redundancy is a continuous structural variable rather than a defect. Wang *et al.* Wang et al. (2020) interpreted redundancy reduction as a regularization objective in unsupervised domain adaptation, showing that selectively removing domain-specific redundancies improves transferability. More dynamically, Dorovatas *et al.* Dorovatas et al. (2025) introduced **Auto-Compressing Networks** that self-organize to prune redundant pathways during training while maintaining representational quality, implicitly converging toward an internal redundancy equilibrium. Beyond deep learning, Li *et al.* Li et al. (2023) demonstrated that exploiting intrinsic redundancy in large materials-science datasets enhances retrieval efficiency without compromising performance, underscoring the universality of redundancy as a reusable structural resource.

From a theoretical neuroscience perspective, Sajid *et al.* Sajid et al. (2020) provided a formal treatment of *degeneracy* and *redundancy* within the free-energy and active-inference framework. They defined degeneracy as the entropy of posterior beliefs—the flexibility afforded by multiple internal explanations for sensory outcomes—and redundancy as the complexity cost of forming those beliefs. Their analysis showed that adaptive agents minimize redundancy while maintaining sufficient degeneracy, achieving robust yet efficient inference. This formulation extends the role of redundancy beyond coding efficiency to a quantitative component of self-organization in both biological and synthetic intelligence, resonating with our view that redundancy regulates informational stability and adaptability rather than constituting noise.

Collectively, these developments trace a conceptual transition: from viewing redundancy as inefficiency to recognizing it as a *principle of organization*. Our work builds on this transition by offering a unified mathematical framework that formalizes redundancy across domains and demonstrates its theoretical and empirical role in stabilizing learning and enhancing generalization.

3 A Unified Redundancy Framework

Notation. Let $(\Omega, \mathcal{F})$ be a measurable space. A random vector $X = (X_1, \ldots, X_n)$ takes values in the product space $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$. For each coordinate, let μ_i be a σ-finite base measure on $\mathcal{X}_i$ and define $\mu = \bigotimes_{i=1}^n \mu_i$. Assume that $P_X \ll \mu$ with joint density $p(x) = \frac{dP_X}{d\mu}(x)$, and that each marginal $P_{X_i} \ll \mu_i$ has density $p_i(x_i) = \frac{dP_{X_i}}{d\mu_i}(x_i)$. The corresponding independent product measure is

$$\Pi_X = \bigotimes_{i=1}^n P_{X_i}, \qquad \frac{d\Pi_X}{d\mu}(x) = \prod_{i=1}^n p_i(x_i).$$

For a positive semidefinite matrix M, denote by $\lambda(M)$ its eigenvalues, tr its trace, and $\|\cdot\|_F$ the Frobenius norm. All logarithms are natural.

Master Definition: Redundancy as Divergence from Independence

Definition 1 (Unified redundancy functional)**.** *Let $f : \mathbb{R}_{>0} \to \mathbb{R}$ be a convex function satisfying $f(1) = 0$. The* redundancy *of a random vector $X = (X_1, \ldots, X_n)$ is defined as the f-divergence between its joint law and the independence manifold:*

$$\mathcal{R}_f(X) := D_f\big(P_X \,\|\, \Pi_X\big) = \int_{\mathcal{X}} f\left(\frac{p(x)}{\prod_{i=1}^n p_i(x_i)}\right) \prod_{i=1}^n p_i(x_i) \, d\mu(x) = \mathbb{E}_{\Pi_X}\left[f\left(\frac{p(x)}{\prod_{i=1}^n p_i(x_i)}\right)\right]. \quad (1)$$

This quantity is finite whenever $P_X \ll \Pi_X$; otherwise we set $\mathcal{R}_f(X) = +\infty$. We refer to $\mathcal{R}_f$ as the redundancy functional *and to f as its* redundancy kernel*, specifying the geometry of the divergence.*

Remark 1 (Specializations and the role of the kernel f)**.** ***Role of the kernel.*** *In an f-divergence, the convex kernel f determines how deviations from independence are measured and weighted.*

(i) *It defines the* local geometry *near independence through $f''(1)$, which governs the quadratic (Gaussian) approximation of the divergence.*

(ii) *It controls the* tail sensitivity *via the growth of $f(t)$ as $t \to 0$ or $t \to \infty$, specifying how strongly rare but large dependencies are penalized.*

(iii) *Its derivative $f'(t)$ acts as an* influence function *(contrast function), determining which ranges of the likelihood ratio contribute most to the redundancy.*

(iv) *Different choices of f induce distinct yet data-processing–consistent geometries on the independence manifold, all obeying the same monotonicity under coarse-graining.*

- **Kullback–Leibler kernel:** For $f(t) = t \log t$,

$$\mathcal{R}_f(X) = D_{\mathrm{KL}}(P_X \| \Pi_X),$$

  the total correlation (multi-information). This kernel grows superlinearly and is highly sensitive to rare but strong dependencies.

- **Pearson χ^2 kernel:** For $f(t) = \frac{1}{2}(t-1)^2$,

$$\mathcal{R}_f(X) = \frac{1}{2} \int \frac{(p - \prod_i p_i)^2}{\prod_i p_i} \, dx,$$

  a quadratic-energy formulation that is less sensitive to tails and well suited to weak or near-Gaussian dependence.

- **Rényi-α family:** Power-type kernels yield Rényi-style redundancies, tuning tail emphasis continuously with α: larger α magnifies the influence of strong dependencies, while $\alpha \downarrow 1$ recovers the KL case.

- **Other classical choices:** Hellinger, Jensen–Shannon, and related kernels provide alternative redundancy geometries, each with its own balance between robustness and sensitivity.

Summary. Information-theoretic, statistical, and geometric notions of redundancy are thus unified under the single principle

$$\mathcal{R}_f(X) = D_f(P_X \| \Pi_X),$$

where the kernel f acts as the lens that shapes the sensitivity, robustness, and intrinsic geometry of the measure.

3.1 Fundamental Properties of Redundancy

Proposition 1 (Nonnegativity and vanishing iff independence)**.** *For any convex function $f : \mathbb{R}_{>0} \to \mathbb{R}$ satisfying $f(1) = 0$, the redundancy functional satisfies*

$$\mathcal{R}_f(X) \geq 0, \qquad and \qquad \mathcal{R}_f(X) = 0 \iff P_X = \Pi_X \quad (i.e.,\ X_1, \ldots, X_n\ are\ mutually\ independent).$$

Sketch. This follows from standard properties of f-divergences: for any probability measures P, Q with $P \ll Q$, we have $D_f(P \| Q) \geq 0$, with equality if and only if $P = Q$ almost everywhere. $\qquad \square$

Proposition 2 (Data processing inequality (DPI))**.** *Let $Y_i = g_i(X_i)$ be measurable mappings or coordinate-wise random channels, and let $Y = (Y_1, \ldots, Y_n)$. Then*

$$\mathcal{R}_f(Y) = D_f(P_Y \| \Pi_Y) \leq D_f(P_X \| \Pi_X) = \mathcal{R}_f(X).$$

Sketch. f-divergences are contractive under Markov kernels. In this setting, the overall channel acting on X is the product of the coordinate-wise channels (g_i), which maps Π_X to Π_Y. Hence the contraction inequality implies $\mathcal{R}_f(Y) \leq \mathcal{R}_f(X)$. $\qquad \square$

Proposition 3 (Bounds). *Assume $P_X \ll \Pi_X$ and that $f(L(x))$ is finite on a Π_X-full set, where $L(x) = p(x)/\prod_i p_i(x_i)$. Then:*

1. *(Lower bound)* $\mathcal{R}_f(X) \geq 0$ *(Prop. 1).*

2. *(Upper bounds) If $0 < m \leq L(x) \leq M < \infty$ Π_X-a.e., then $\mathcal{R}_f(X) \leq f(M)$; in particular for $f(t) = t \log t$, $\mathcal{R}_f(X) \leq M \log M$.*

3.2 Gaussian and Geometric Reductions

Proposition 4 (Gaussian total correlation). *Let $X \sim \mathcal{N}(0, \Sigma)$ with correlation matrix $C = \mathrm{corr}(X)$. Then the redundancy under the Kullback–Leibler kernel equals the Gaussian total correlation:*

$$\mathcal{R}_{\mathrm{KL}}(X) \;=\; D_{\mathrm{KL}}(P_X \| \Pi_X) \;=\; -\tfrac{1}{2} \log \det C.$$

Sketch. For a centered Gaussian X, the joint density is $p(x) \propto \exp\left(-\tfrac{1}{2} x^\top \Sigma^{-1} x\right)$, and the independent-product density corresponds to replacing Σ by its diagonal $\mathrm{diag}(\Sigma)$. Direct integration yields

$$D_{\mathrm{KL}}(P_X \| \Pi_X) = \tfrac{1}{2}(\mathrm{tr}(C) - \log \det C - n),$$

which reduces to $-\tfrac{1}{2} \log \det C$ when C is a correlation matrix ($\mathrm{tr}(C) = n$). □

Proposition 5 (Quadratic approximation and covariance form). *Let $C = \mathrm{corr}(X)$ and assume weak dependence (all off-diagonal entries of C are small). For the quadratic kernel $f(t) = \tfrac{1}{2}(t-1)^2$,*

$$\mathcal{R}_f(X) \;\approx\; \tfrac{1}{4} \|C - I\|_F^2 \;=\; \tfrac{1}{4} \sum_{i \neq j} C_{ij}^2.$$

Moreover, for $\|C - I\|_2$ sufficiently small,

$$-\log \det C \;\approx\; \mathrm{tr}(I - C) \;+\; \tfrac{1}{2} \|C - I\|_F^2,$$

showing that the Gaussian total correlation $\mathcal{R}_{\mathrm{KL}}$ and the quadratic redundancy $\mathcal{R}_f$ coincide up to second order.

Sketch. Second-order Taylor expansion of f and $\log \det$ around independence $C = I$. □

3.3 Spectral and Geometric Redundancy

Definition 2 (Spectral (geometric) redundancy). *Let $Z \in \mathbb{R}^D$ be a random vector with covariance matrix Σ_Z. Denote its eigenvalues by $\lambda_1, \ldots, \lambda_D > 0$, and define the normalized spectrum*

$$\tilde{\lambda}_i \;=\; \frac{\lambda_i}{\sum_{j=1}^D \lambda_j}, \qquad \sum_{i=1}^D \tilde{\lambda}_i = 1.$$

The spectral entropy *and the corresponding* effective rank *are*

$$H_\lambda(Z) \;=\; -\sum_{i=1}^D \tilde{\lambda}_i \log \tilde{\lambda}_i, \qquad r_{\mathrm{eff}}(Z) \;=\; \exp\left(H_\lambda(Z)\right).$$

We define the spectral redundancy *as*

$$\mathcal{R}_{\mathrm{spec}}(Z) \;=\; 1 - \frac{r_{\mathrm{eff}}(Z)}{D}, \qquad \mathcal{R}_{\mathrm{spec}}(Z) \in \left[0, \, 1 - \tfrac{1}{D}\right].$$

Proposition 6 (Bounds and extremal cases). *The spectral redundancy $\mathcal{R}_{\mathrm{spec}}(Z)$ satisfies:*

$$\mathcal{R}_{\mathrm{spec}}(Z) = 0 \quad \Longleftrightarrow \quad \tilde{\lambda}_1 = \cdots = \tilde{\lambda}_D = \tfrac{1}{D},$$

i.e., when the spectrum is uniform (energy evenly distributed across dimensions); and

$$\mathcal{R}_{\mathrm{spec}}(Z) = 1 - \tfrac{1}{D} \quad \Longleftrightarrow \quad \tilde{\lambda}_1 = 1, \ \tilde{\lambda}_{i>1} = 0,$$

i.e., when Σ_Z is rank-one (complete collapse onto a single direction).

Sketch. The result follows from the properties of Shannon entropy. H_λ is maximized for the uniform distribution, yielding $r_{\mathrm{eff}} = D$, and minimized when one component carries all the mass, giving $r_{\mathrm{eff}} = 1$. Substituting these values into $\mathcal{R}_{\mathrm{spec}} = 1 - r_{\mathrm{eff}}/D$ establishes the bounds. □

3.4 Redundancy Balance and the Existence of an Interior Optimum

Let $\mathcal{R}(Z)$ denote any redundancy measure (e.g., $\mathcal{R}_{\mathrm{KL}}$, $\mathcal{R}_{\chi^2}$, or $\mathcal{R}_{\mathrm{spec}}$) evaluated on latent representations $Z = f_\theta(X)$.

Assumption 1 (Loss–redundancy trade-off). *Assume the training objective decomposes as*

$$\mathcal{L}(\theta) \;=\; \mathbb{E}[\ell_{\mathrm{rec}}(X; \theta)] \;+\; \lambda\, \mathcal{R}\big(f_\theta(X)\big), \qquad \lambda > 0.$$

Let $D(R)$ denote the minimal achievable reconstruction risk at a fixed redundancy level $R \in [\underline{R}, \overline{R}]$. Assume that $D(R)$ is continuous, strictly decreasing on $[\underline{R}, R_1)$, and strictly increasing on $(R_2, \overline{R}]$ for some $\underline{R} < R_1 \leq R_2 < \overline{R}$.

Theorem 1 (Existence of a redundancy equilibrium). *Under Assumption 1, the scalar objective*

$$\Phi(R) \;=\; D(R) + \lambda R$$

admits at least one minimizer $R^* \in (\underline{R}, \overline{R})$. If D is twice differentiable and $D''(R^*) > 0$, then R^* is a locally stable equilibrium point corresponding to the optimal redundancy.

Proof. By continuity of $D(R)$ and the assumed monotonicity near the boundaries, the derivative $\Phi'(R) = D'(R) + \lambda$ is negative for R close to $\underline{R}$ and positive for R close to $\overline{R}$. Hence, by the intermediate value theorem, there exists $R^* \in (\underline{R}, \overline{R})$ satisfying $\Phi'(R^*) = 0$. If, in addition, $D''(R^*) > 0$, then $\Phi''(R^*) > 0$, and R^* is a local minimizer—interpreted as a stable redundancy equilibrium. □

Corollary 1 (Banded convergence signature). *If the training dynamics θ_t monotonically decrease $\mathcal{L}(\theta_t)$ and both $\mathcal{L}(\theta_t)$ and $\mathcal{R}(f_{\theta_t}(X))$ are Lipschitz-continuous in t with additive stochastic noise, then $\mathcal{R}(f_{\theta_t}(X))$ converges to a narrow band around R^* and exhibits small fluctuations whose amplitude is proportional to the noise scale.*

3.5 Recovering Classical Notions and Practical Proxies

Total correlation (multi-information). For $f(t) = t \log t$,

$$\mathcal{R}_{\mathrm{KL}}(X) \;=\; \sum_{i=1}^n H(X_i) - H(X),$$

which exactly measures the amount of *shared information* among the coordinates.

Quadratic / covariance proxy. For near-Gaussian representations,

$$\mathcal{R}_{\chi^2}(X) \;\approx\; \tfrac{1}{4}\, \|C - I\|_F^2,$$

providing a simple, differentiable redundancy regularizer that penalizes excess pairwise correlation.

Spectral proxy. The spectral redundancy $\mathcal{R}_{\mathrm{spec}}(Z)$ is norm- and scale-invariant, computationally inexpensive, and captures the trade-off between *collapse* (energy concentration) and *fragmentation* (over-dispersion) in the representation spectrum.

Attention-head redundancy. Let $A_h \in \mathbb{R}^{N \times N}$ denote the attention map of head h. A simple head-redundancy proxy is

$$\mathcal{R}_{\mathrm{head}} \;=\; \frac{1}{H(H-1)} \sum_{h \neq h'} \frac{\langle \mathrm{vec}(A_h),\, \mathrm{vec}(A_{h'}) \rangle^2}{\|A_h\|_F^2 \, \|A_{h'}\|_F^2},$$

which lies in $[0, 1]$ and contracts under head-wise post-processing, consistent with the data processing inequality for linear maps.

3.6 Practical Lemmas for Boundedness During Training

Lemma 1 (Bound via spectral norm control)**.** *Let* $Z = WX$ *where* $\|W\|_2 \leq M$ *and* $\mathbb{E}\|X\|^2 \leq \sigma^2$. *Then the covariance of* Z *satisfies* $\mathrm{tr}(\Sigma_Z) \leq M^2 \sigma^2$, *and consequently the effective rank* $r_{\mathrm{eff}}(Z)$ *lies in* $[1, D]$. *Hence,*

$$0 \;\leq\; \mathcal{R}_{\mathrm{spec}}(Z) \;\leq\; 1 - \tfrac{1}{D}.$$

Sketch. Since $\Sigma_Z = W \Sigma_X W^\top$ and $\|W\|_2 \leq M$, we have $\mathrm{tr}(\Sigma_Z) = \mathrm{tr}(W \Sigma_X W^\top) \leq \|W\|_2^2 \, \mathrm{tr}(\Sigma_X) \leq M^2 \sigma^2$. The normalization of eigenvalues then ensures $r_{\mathrm{eff}}(Z) \in [1, D]$, yielding the stated bounds. $\qquad\square$

Lemma 2 (KL–Frobenius bound near independence)**.** *Let* C *be a correlation matrix with spectral radius* $\rho(C) < 1$, *and define* $A = C - I$. *Then*

$$-\log \det C \;=\; -\mathrm{tr} \log(I + A) \;\geq\; \tfrac{1}{2}\|A\|_F^2 - \tfrac{1}{3}\|A\|_F^3,$$

implying that for small $\|A\|_F$,

$$\mathcal{R}_{\mathrm{KL}}(X) \;\gtrsim\; \tfrac{1}{2}\|C - I\|_F^2.$$

Sketch. Using the Taylor expansion $\log(I+A) = A - \tfrac{1}{2}A^2 + \tfrac{1}{3}A^3 - \cdots$ and $\mathrm{tr}(A) = 0$ for correlation matrices, we obtain $-\log \det C = -\mathrm{tr} \log(I+A) \geq \tfrac{1}{2}\|A\|_F^2 - \tfrac{1}{3}\|A\|_F^3$. $\qquad\square$

3.7 Information-Theoretic Sandwich Proof: Why Redundancy Is Not "The Less, the Better"

We now demonstrate that the classical Shannon view of redundancy—interpreting it purely as inefficiency to be minimized—fails to hold in finite, noisy, and structured regimes. Instead, the task error $\mathcal{E}(R)$ (or risk as a function of redundancy R) typically follows a *U-shaped* dependence: both insufficient and excessive redundancy degrade performance, while an intermediate value R^* yields optimal generalization and manifests as a stable equilibrium band around which $\mathcal{R}(f_{\theta_t}(X))$ fluctuates during training.

Setting. Let $Z = f_\theta(X)$ be a learned representation of data X that is relevant to a downstream target S (e.g., a latent factor, class label, or generative source). Denote redundancy by $R = \mathcal{R}(Z)$, measured by any f-divergence form (Def. 1) or by its geometric proxies such as $\mathcal{R}_{\chi^2}$ or $\mathcal{R}_{\mathrm{spec}}$. Assume a bounded entropy budget $\sum_i H(Z_i) \leq C_0$ and that the channel $X \to Z$ is affected by finite noise or perturbation power.

Lemma 3 (Capacity-side lower bound)**.** *The mutual information between* Z *and the task variable S satisfies*

$$I(Z; S) \;\leq\; H(Z) \;\leq\; \sum_i H(Z_i) - \mathrm{TC}(Z) \;=\; C_0 - \mathcal{R}_{\mathrm{KL}}(Z),$$

and therefore

$$\mathcal{E}(R) \;\geq\; g_{\mathrm{info}}(C_0 - R),$$

where g_{info} *is a strictly increasing function.*

Sketch. From the entropy decomposition $H(Z) = \sum_i H(Z_i) - \mathrm{TC}(Z)$ and the finite entropy budget $\sum_i H(Z_i) \leq C_0$, we obtain $H(Z) \leq C_0 - \mathrm{TC}(Z)$. Since $I(Z;S) \leq H(Z)$, either Fano's inequality (for classification) or the I–MMSE relationship (for Gaussian regression) implies that the task risk increases monotonically as $I(Z;S)$ decreases. Hence, excessive redundancy (large total correlation) reduces the effective information capacity available for the task. $\square$

Lemma 4 (Robustness-side upper bound). *If the encoding $X \to Z$ and decoding $Z \to \hat{S}$ operate under bounded stochastic noise, dropout, or adversarial perturbations, then*

$$\mathcal{E}(R) \ \leq \ g_{\mathrm{robust}}(R),$$

for some continuous function g_{robust} that is strictly decreasing on $[0, R_1)$.

Sketch. When redundancy is close to zero (i.e., the coordinates of Z are nearly independent), the representation lacks alternative pathways for recovering corrupted or missing information. Information-theoretic rate–distortion bounds and classical error-correction arguments show that replication or correlated encoding across coordinates (nonzero redundancy) reduces the expected reconstruction or classification error. Consequently, increasing small amounts of redundancy strictly improves robustness to perturbations, yielding a decreasing error function $g_{\mathrm{robust}}(R)$ in the low-redundancy regime. $\square$

3.8 U-shape and Interior Optimum

Theorem 2 (Interior optimal redundancy and stability band). *Assume that $\mathcal{E}(R)$ is continuous and satisfies*

$$g_{\mathrm{robust}}(R) \ \leq \ \mathcal{E}(R) \ \leq \ g_{\mathrm{info}}(C_0 - R),$$

where g_{robust} is strictly decreasing on $[0, R_1)$ and g_{info} is strictly increasing on $(R_2, C_0]$ for some $0 < R_1 \leq R_2 < C_0$. Then there exists an interior minimizer $R^* \in (R_1, R_2)$ such that

$$\mathcal{E}(R^*) \ = \ \min_{R \in [0, C_0]} \mathcal{E}(R).$$

If the training dynamics follow noisy gradient descent that decreases $\mathcal{E}$ and $\mathcal{R}(f_{\theta_t}(X))$ is Lipschitz-continuous in time, then $\mathcal{R}(f_{\theta_t}(X))$ converges to a stability band $[R_{\min}, R_{\max}]$ around R^*.

Proof. By Lemmas 3–4, $\mathcal{E}(R)$ is bounded between two functions with opposing monotonic trends near the boundaries. By continuity, $\mathcal{E}'(R)$ changes sign, and by the intermediate value theorem, there exists R^* such that $\mathcal{E}'(R^*) = 0$. Local convexity of g_{info} and g_{robust} implies that $\mathcal{E}''(R^*) > 0$, yielding a stable minimum. Under small gradient noise, stochastic approximation theory ensures convergence to a narrow fluctuation band $[R_{\min}, R_{\max}]$ around this equilibrium point. $\square$

Corollary 2 (Operational interpretation). *For $R < R_1$, the system is under-redundant—fragile and non-generalizing. For $R > R_2$, it is over-redundant—collapsed and low in effective capacity. Between them lies a redundancy equilibrium band $[R_{\min}, R_{\max}]$ where efficiency, robustness, and generalization coexist in balance.*

Interpretation. Shannon's 1948 paradigm addressed the asymptotic and noiseless limit, in which redundancy merely wastes bits. In contrast, real-world and learning systems operate in finite-sample, noisy, and structured regimes, where redundancy serves as a functional resource rather than a defect. It provides (i) structural stability through multiple consistent cues, (ii) error correction and smooth interpolation, and (iii) invariance under perturbations. Thus, *redundancy is not the opposite of information, but the geometry that sustains it.* An optimal, nonzero level of redundancy forms the basis of structure itself—and constitutes the mathematical condition for understanding.

4 Experiments

Goal. The goal of our experiments is to empirically validate the theoretical predictions of the redundancy–balance framework rather than to introduce any architectural innovation. Specifically, we test Theorem 2, which predicts that the generalization performance of a learning system depends non-monotonically on its internal redundancy R: both insufficient and excessive redundancy impair stability, while an intermediate value R^* yields optimal organization and generalization.

Setup. We adopt the masked autoencoder (MAE) He et al. (2022) as a representative self-supervised architecture and incorporate our redundancy regularization into its pretraining objective. The model reconstructs masked image patches under a fixed mask ratio of 0.75 using a ViT-Base encoder–decoder backbone. To test the redundancy–balance hypothesis, we augment the original reconstruction loss with a spectral-redundancy term, yielding the total objective.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \lambda_{\text{red}} \mathcal{R}_{\text{spec}}(Z), \qquad \text{where} \quad \mathcal{L}_{\text{recon}} = \| \hat{X} - X \|_2^2. \qquad (2)$$

Here, $\mathcal{R}_{\text{spec}}(Z) = 1 - r_{\text{eff}}(Z)/D$ is the spectral-redundancy functional defined in Section 3.3. The coefficient λ_{red} controls the trade-off between pixel-level reconstruction fidelity and structural redundancy in the latent representation Z. By varying $\lambda_{\text{red}} \in \{0, 10^{-3}, 10^{-2}, 5 \times 10^2\}$, we obtain models spanning distinct redundancy regimes. Except for this single coefficient, all experimental settings—including architecture, optimizer, learning-rate schedule, data augmentation, and mask ratio—are kept identical across runs. After pretraining, the encoder is frozen and evaluated through a linear-probe classifier on CIFAR-100, providing a downstream measure of generalization ability.

Metrics. We monitor two redundancy proxies during pretraining: (i) the covariance-based redundancy $\mathcal{R}_{\chi^2}$, and (ii) the spectral redundancy $\mathcal{R}_{\text{spec}}$, computed from the effective rank r_{eff} of the covariance spectrum as $1 - r_{\text{eff}}/D$. While validation losses remain nearly identical across different λ_{red}, the resulting representations exhibit markedly different redundancy profiles, making these metrics essential for quantifying structural organization. Downstream generalization is assessed via the linear-probe Top-1 accuracy on CIFAR-100, which serves as a sensitive indicator of information balance.

Training protocol. All models are trained for the same number of epochs with identical random seeds, optimizer hyperparameters, learning-rate schedules, and data augmentations. **This strict control ensures that any observed difference in generalization or redundancy originates solely from the redundancy-regularization term λ_{red}.** Redundancy statistics are logged at each epoch to analyze both equilibrium levels and dynamic trajectories during training.

Summary. This experimental design isolates the effect of redundancy regulation in a fully controlled setting and directly tests the theoretical predictions of equilibrium and self-organization described in Sections 3.4 and 3.8.

5 Results and Analysis

Empirical validation of the redundancy–balance theorem. Figure 1 shows the empirical relationship between spectral redundancy and downstream generalization performance. The results confirm the key prediction of Theorem 2: generalization follows a clear U-shaped dependence on redundancy. When no redundancy regularization is applied ($\lambda_{\text{red}}=0$), the model exhibits high spectral coupling ($\mathcal{R}_{\text{spec}} \approx 0.96$) and poor generalization (35.6% Top-1 on CIFAR-100). Introducing mild redundancy control ($\lambda_{\text{red}}=10^{-3}$) slightly improves generalization (36.3%) but remains in the over-coupled regime. At an intermediate regularization strength ($\lambda_{\text{red}}=10^{-2}$), spectral redundancy decreases to approximately 0.5 and yields the best generalization (41.4%). When regularization becomes too strong ($\lambda_{\text{red}}=5 \times 10^2$), redundancy is over-suppressed ($\mathcal{R}_{\text{spec}} \approx 0.19$), leading to a decline in performance (39.6%). This non-monotonic trend demonstrates the existence of an interior optimum R^* balancing compression and coupling, consistent with the theoretical redundancy equilibrium.
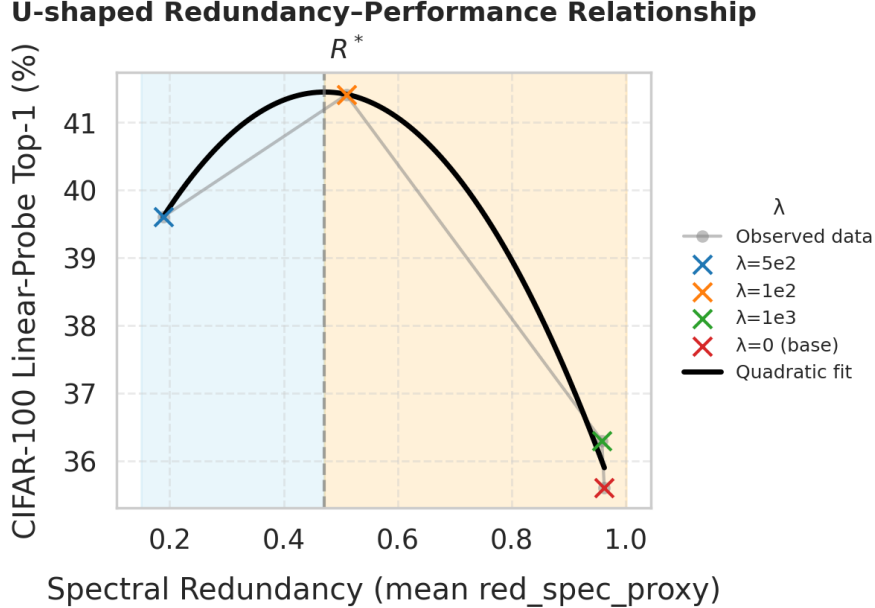
Figure 1: **U-shaped redundancy–performance relationship.** CIFAR-100 linear-probe Top-1 accuracy (vertical axis) varies non-monotonically with the spectral redundancy $\mathcal{R}_{\mathrm{spec}}$ (horizontal axis). Both over-redundant (high $\mathcal{R}_{\mathrm{spec}}$) and under-redundant (low $\mathcal{R}_{\mathrm{spec}}$) models exhibit degraded performance, while an intermediate redundancy ($\lambda_{\mathrm{red}}=10^{-2}$, $\mathcal{R}_{\mathrm{spec}} \approx 0.5$) achieves the best generalization (41.4%). The dashed line marks the fitted optimum R^* predicted by Theorem 2, and shaded regions indicate under- and over-redundant regimes.

Quantitative summary. Table 2 summarizes the relationship between redundancy, validation loss, and downstream generalization across different regularization regimes. While validation losses remain nearly identical across runs, the redundancy levels and probe accuracies vary significantly, indicating that the improvement in generalization arises from structural organization rather than overfitting or optimization bias.

Table 2: Summary of MAE pretraining under varying redundancy regularization strengths. Spectral redundancy $\mathcal{R}_{\mathrm{spec}}$ is computed as the mean value over the final training epochs. Top-1 accuracy is measured by linear probing on CIFAR-100.

Redundancy weight λ_{red}	Spectral redundancy $\mathcal{R}_{\mathrm{spec}}$	Validation loss	CIFAR-100 Top-1 (%)	Regime characterization
0 (baseline)	0.96	0.0293	35.6	Over-coupled (collapse)
10^{-3}	0.96	0.0293	36.3	Slightly over-coupled
10^{-2}	0.51	0.0297	**41.4**	Balanced (near R^*)
5×10^2	0.19	0.0315	39.6	Under-redundant (fragmented)

Interpretation. The observed U-shaped pattern provides direct empirical support for the redundancy–balance theory. It indicates that redundancy is not merely a form of noise to be suppressed, but

a structured degree of freedom that must be maintained at an optimal level to achieve stability and generalization. This equilibrium behavior also implies that learning dynamics naturally converge to a self-organized regime around R^*, rather than requiring fine-tuned regularization. Together, these results establish redundancy as a measurable and predictive quantity governing the geometry of learned representations. This motivates further examination of how redundancy evolves during training and whether it naturally stabilizes around the predicted equilibrium R^*.

Training dynamics and equilibrium formation. Figure 2 illustrates the joint evolution of effective rank (solid lines) and spectral redundancy (dashed lines) across different regularization strengths λ_{red}. The trajectories reveal three key phenomena consistent with the redundancy–balance theory. First, redundancy regularization exerts a *nonlinear* influence on representational structure: small increases in λ_{red} induce moderate reductions in redundancy, whereas large values (e.g., 5×10^2) lead to a rapid collapse of redundancy to near-zero levels. This sensitivity implies that λ_{red} controls a structural, rather than purely statistical, phase transition in the organization of representations.

Second, redundancy and effective rank evolve as strongly coupled variables, showing clear inverse trajectories. As redundancy decreases, the effective dimensionality of latent features expands, indicating that redundancy acts as an internal regulator of usable representational degrees of freedom. However, when redundancy is over-suppressed, the representation becomes overly fragmented—large in rank but poor in coherence—highlighting that high dimensionality alone does not guarantee generalization.

Third, moderate regularization ($\lambda_{\mathrm{red}} = 10^{-2}$) leads to a stable *equilibrium band*, where redundancy and effective rank co-stabilize after early fluctuations. This regime corresponds to the predicted redundancy equilibrium R^*: a self-organizing attractor in learning dynamics where compression and coupling are balanced. The emergence of this equilibrium without fine-tuning supports the theoretical claim that redundancy is a self-organizing property of data manifolds, not an artifact of optimization.
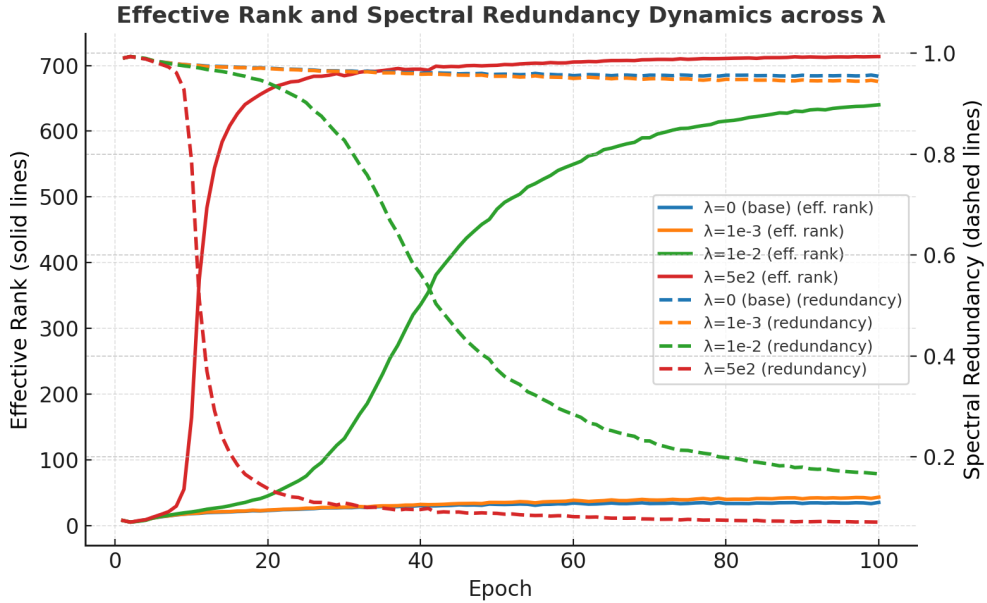

Figure 2: **Training dynamics of effective rank and spectral redundancy across different redundancy regularization strengths** λ_{red}**.** Solid lines denote the evolution of effective rank, and dashed lines denote spectral redundancy. Moderate regularization ($\lambda_{\mathrm{red}} = 10^{-2}$) yields a stable equilibrium where redundancy and rank co-stabilize, while excessive regularization (5×10^2) collapses redundancy. The relationship between λ_{red} and redundancy is markedly non-linear, confirming that redundancy regulation induces a self-organizing equilibrium rather than monotonic suppression.

6 Discussion

Summary and perspective. This work proposes a theoretical paradigm that redefines redundancy as a fundamental quantity of information organization rather than inefficiency. We formulate a unified family of redundancy functionals $\{\mathcal{R}_f\}$ based on f-divergences from statistical independence, showing that classical measures such as mutual information, χ^2 dependence, and spectral redundancy are projections of the same underlying geometry. The theory demonstrates that redundancy is bounded both above and below, giving rise to an intrinsic equilibrium R^* between over-compression and over-coupling. Regularizing redundancy toward this equilibrium enhances stability and generalization in finite, structured regimes. Together, these results establish redundancy as a *structural information principle*—a paradigm shift that extends Shannon's efficiency-centric information theory into the finite, noisy, and organized world of learning systems.

Redundancy and foundation models. Our framework provides conceptual and practical insights for the design of *generative self-supervised and foundation models*. These systems—ranging from masked autoencoders and diffusion models to large language and multimodal models Ho et al. (2020); Rombach et al. (2022); He et al. (2024); Caro et al. (2023); Cox et al. (2024); Wu et al. (2025); Sun et al. (2025)—learn by reconstructing or predicting structured dependencies within data, a process inherently governed by redundancy. The redundancy–balance principle suggests that such generative models achieve their best understanding and transferability when latent redundancy is regulated near its equilibrium R^*, maintaining a stable internal geometry between over-coupling and over-compression. In practice, *explicit redundancy regularization*—for instance via spectral or covariance-based constraints—could provide a self-organizing mechanism that stabilizes large generative pretraining and enhances robustness across modalities. Thus, our theory complements current scale-driven approaches by offering a structural principle for efficient, coherent, and generalizable generative learning.

Limitations and open directions. While the proposed theory provides a unifying view of redundancy, several limitations remain. First, our empirical validation focuses on generative self-supervised learning (e.g., masked autoencoders) rather than on large-scale foundation models such as GPTs Achiam et al. (2023) or multimodal LLMs Xu et al. (2024). Although we hypothesize that the redundancy–balance principle also governs their emergent generalization, confirming this requires extensive large-scale experiments. Second, the equilibrium theorem is not expected to be universally optimal: discriminative or contrastive paradigms may still benefit from minimizing redundancy to maximize separability. The framework thus captures one axis of the generative–discriminative continuum, but not the entirety of learning dynamics. Third, the theory has not yet been tested in cross-disciplinary domains such as neuroscience, where redundancy is hypothesized to underlie degeneracy and robustness in brain activity. Validating the redundancy equilibrium on neural and cognitive datasets will be an essential step toward bridging machine and biological learning. Finally, while our analysis concentrates on vector-valued data, redundancy in other modalities—particularly graphs, multimodal signals, and spatiotemporal structures—remains largely unexplored. Quantifying "graph redundancy" and integrating it with graph neural networks presents especially promising directions for future work Gates et al. (2021); Kumar et al. (2024).

Outlook. Despite these limitations, the redundancy–balance principle offers a new foundation for understanding learning systems as self-organizing processes that regulate—not eliminate—dependence. It reframes generalization as a property of data organization rather than model complexity, suggesting that redundancy forms the structural backbone through which stability and understanding emerge. Future research should explore how data redundancy interacts with other fundamental factors such as model capacity, architectural complexity, and optimization dynamics. Jointly characterizing these variables may yield a unified theory of learning equilibria, where data structure and model structure co-adapt to maintain informational balance. Such an extension could guide the principled design of large-scale foundation models, multimodal architectures, and biological neural systems alike. In this broader view, redundancy may ultimately serve as the missing link connecting statistical learning, information theory, and cognitive computation—a structural law governing how complex systems learn to understand.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International conference on machine learning*, pp. 242–252. PMLR, 2019.

Horace Barlow. Redundancy reductionrevisited. *Network: computation in neural systems*, 12(3):241, 2001.

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.

Anthony J Bell and Terrence J Sejnowski. The "independent components" of natural scenes are edge filters. *Vision research*, 37(23):3327–3338, 1997.

Yakir Berchenko. Simplicity bias in overparameterized machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11052–11060, 2024.

Josue Ortega Caro, Antonio H de O Fonseca, Christopher Averill, Syed A Rizvi, Matteo Rosati, James L Cross, Prateek Mittal, Emanuele Zappala, Daniel Levine, Rahul M Dhodapkar, et al. Brainlm: A foundation model for brain activity recordings. *bioRxiv*, pp. 2023–09, 2023.

Jireh Yi-Le Chan, Steven Mun Hong Leow, Khean Thye Bea, Wai Khuen Cheng, Seuk Wai Phoong, Zeng-Wei Hong, and Yen-Lin Chen. Mitigating the multicollinearity problem and its machine learning approach: a review. *Mathematics*, 10(8):1283, 2022.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Joseph Cox, Peng Liu, Skylar E Stolte, Yunchao Yang, Kang Liu, Kyle B See, Huiwen Ju, and Ruogu Fang. Brainsegfounder: Towards 3d foundation models for neuroimage segmentation. *Medical Image Analysis*, 97:103301, 2024.

Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(9):10850–10869, 2023.

Diego Doimo, Aldo Glielmo, Sebastian Goldt, and Alessandro Laio. Redundant representations help generalization in wide neural networks. *Advances in Neural Information Processing Systems*, 35:19659–19672, 2022.

Vaggelis Dorovatas, Georgios Paraskevopoulos, and Alexandros Potamianos. Auto-compressing networks. *arXiv preprint arXiv:2506.09714*, 2025.

Alexander J Gates, Rion Brattig Correia, Xuan Wang, and Luis M Rocha. The effective graph reveals redundancy, canalization, and control pathways in biochemical regulation and signaling. *Proceedings of the National Academy of Sciences*, 118(12):e2022598118, 2021.

RF Gunst and JT Webster. Regression analysis and problems of multicollinearity. *Communications in Statistics-Theory and Methods*, 4(3):277–292, 1975.

Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Zhang, Bingbing Liu, and Ying-Cong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. *arXiv preprint arXiv:2409.18124*, 2024.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

David A Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101, 2007.

Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

Behnam Kazemivash, Pranav Suresh, Dong Hye Ye, Armin Iraji, Jingyu Liu, Sergey Plis, Peter Kochunov, David C Zhu, and Vince D Calhoun. st-densevit: A weakly supervised spatiotemporal vision transformer for dense prediction of dynamic brain networks. *Human Brain Mapping*, 46(14):e70364, 2025.

Pawan Kumar, Siddique Merajuddin, Shariefuddin Pirzada, and Yilun Shang. On the spectral redundancy of pineapple graphs. *Symmetry*, 16(10):1267, 2024.

Kangming Li, Daniel Persaud, Kamal Choudhary, Brian DeCost, Michael Greenwood, and Jason Hattrick-Simpers. Exploiting redundancy in large materials datasets for efficient machine learning with less data. *Nature Communications*, 14(1):7283, 2023.

Seth Lloyd. *Programming the universe: a quantum computer scientist takes on the cosmos*. Vintage, 2007.

Yao Lu, Peixin Zhang, Jingyi Wang, Lei Ma, Xiaoniu Yang, and Qi Xuan. Redtest: Towards measuring redundancy in deep neural networks effectively. *arXiv preprint arXiv:2411.10507*, 2024.

Robert M May. Will a large complex system be stable? *Nature*, 238(5364):413–414, 1972.

Vedant Nanda, Till Speicher, John Dickerson, Krishna Gummadi, Soheil Feizi, and Adrian Weller. Diffused redundancy in pre-trained representations. *Advances in Neural Information Processing Systems*, 36:4055–4079, 2023.

Nandakumar S Narayanan, Eyal Y Kimchi, and Mark Laubach. Redundancy and synergy of neuronal ensembles in motor cortex. *Journal of Neuroscience*, 25(17):4207–4216, 2005.

Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

Ilya Prigogine and Erwin N Hiebert. From being to becoming: Time and complexity in the physical sciences, 1982.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Noor Sajid, Thomas Parr, Thomas M Hope, Cathy J Price, and Karl J Friston. Degeneracy and redundancy in active inference. *Cerebral Cortex*, 30(11):5750–5766, 2020.

Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3): 379–423, 1948.

Yue Sun, Limei Wang, Gang Li, Weili Lin, and Li Wang. A foundation model for enhancing magnetic resonance images and downstream segmentation, registration and diagnostic tasks. *Nature Biomedical Engineering*, 9(4):521–538, 2025.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Giulio Tononi, Olaf Sporns, and Gerald M Edelman. A measure for brain complexity: relating functional segregation and integration in the nervous system. *Proceedings of the National Academy of Sciences*, 91 (11):5033–5037, 1994.

Mengzhu Wang, Xiang Zhang, Long Lan, Wei Wang, Huibin Tan, and Zhigang Luo. Improving unsupervised domain adaptation by reducing bi-level feature redundancy. *arXiv preprint arXiv:2012.15732*, 2020.

Patricia Wollstadt, Sebastian Schmitt, and Michael Wibral. A rigorous information-theoretic definition of redundancy and relevancy in feature selection based on (partial) information decomposition. *Journal of Machine Learning Research*, 24(131):1–44, 2023.

Jiamin Wu, Zichen Ren, Junyu Wang, Pengyu Zhu, Yonghao Song, Mianxin Liu, Qihao Zheng, Lei Bai, Wanli Ouyang, and Chunfeng Song. Adabrain-bench: Benchmarking brain foundation models for brain-computer interface applications. *arXiv preprint arXiv:2507.09882*, 2025.

Mengwei Xu, Wangsong Yin, Dongqi Cai, Rongjie Yi, Daliang Xu, Qipeng Wang, Bingyang Wu, Yihao Zhao, Chen Yang, Shihe Wang, et al. A survey of resource-efficient llm and multimodal foundation models. *arXiv preprint arXiv:2401.08092*, 2024.

Mehmet Can Yavuz and Berrin Yanikoglu. Evaluating the efficiency of latent spaces via the coupling-matrix. *arXiv preprint arXiv:2509.06314*, 2025.

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

Jacob Ziv and Abraham Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on information theory*, 23(3):337–343, 2003.

David Zollikofer, Béni Egressy, Frederik Benzing, Matthias Otth, and Roger Wattenhofer. Beyond pairwise correlations: Higher-order redundancies in self-supervised representation learning. *arXiv preprint arXiv:2412.01926*, 2024.

A Full Mathematical Proofs

A.1 Preliminaries and Conventions

We recall that the f-divergence between two probability measures $P \ll Q$ on $(\mathcal{X}, \mathcal{F})$ with respect to a common σ-finite base measure μ and corresponding densities $p = \frac{dP}{d\mu}$, $q = \frac{dQ}{d\mu}$ is defined as

$$D_f(P\|Q) \;=\; \int_{\mathcal{X}} q(x)\, f\left(\frac{p(x)}{q(x)}\right)\, d\mu(x),$$

where $f : \mathbb{R}_{>0} \to \mathbb{R}$ is convex and satisfies $f(1) = 0$. We use the following standard facts:

(i) (**Nonnegativity**) $D_f(P\|Q) \geq 0$, with equality if and only if $P = Q$ (Csiszár–Morimoto identity).

(ii) (**Data processing**) D_f is nonincreasing under Markov kernels, i.e., $D_f(KP\|KQ) \leq D_f(P\|Q)$ for any stochastic map K.

(iii) (**Local quadratic expansion**) If f is twice differentiable at 1 with $f''(1) > 0$, then $D_f(P\|Q)$ admits a second-order expansion around $P = Q$.

Throughout, $\Pi_X = \bigotimes_{i=1}^{n} P_{X_i}$ denotes the product of the marginals of P_X. We write p for the joint density of P_X and p_i for that of P_{X_i}. For any symmetric matrix A, we denote by $\|A\|_F$ its Frobenius norm, by $\rho(A)$ its spectral radius, and by $\log \det(A)$ the principal matrix logarithm on the cone Sym^{++} of positive definite matrices.

A.2 Proof of Proposition 1 (Nonnegativity and Identity of Indiscernibles)

Proof. By definition,

$$\mathcal{R}_f(X) \;=\; D_f(P_X\|\Pi_X) \;=\; \int_{\mathcal{X}} f\Big(\frac{p(x)}{\prod_i p_i(x_i)}\Big) \prod_i p_i(x_i)\, dx.$$

Since f is convex and satisfies $f(1) = 0$, the Csiszár–Morimoto theorem ensures that $D_f(P\|Q) \geq 0$ with equality if and only if $P = Q$ almost everywhere. Applying this result to $(P, Q) = (P_X, \Pi_X)$ gives the desired claim:

$$\mathcal{R}_f(X) \geq 0 \quad \text{and} \quad \mathcal{R}_f(X) = 0 \iff P_X = \Pi_X,$$

that is, the components $X_1, \ldots, X_n$ are mutually independent. $\qquad\square$

A.3 Proof of Proposition 2 (Data Processing Inequality)

Proof. Let $K = \bigotimes_{i=1}^{n} K_i$ denote the product Markov kernel mapping X to Y, where each K_i acts on coordinate X_i. Then $P_Y = P_X K$ and, since K factorizes coordinate-wise, $\Pi_Y = \Pi_X K$. By the contraction property of f-divergences under Markov kernels,

$$\mathcal{R}_f(Y) \;=\; D_f(P_Y\|\Pi_Y) \;=\; D_f(P_X K \,\|\, \Pi_X K) \;\leq\; D_f(P_X\|\Pi_X) \;=\; \mathcal{R}_f(X),$$

which establishes the inequality. $\qquad\square$

A.4 Proof of Proposition 3 (Bounds)

Proof. **(1) Lower bound.** This follows directly from Proposition 1, which guarantees $\mathcal{R}_f(X) \geq 0$ for any convex f satisfying $f(1) = 0$.

(2) Upper bound. If $m \leq L(x) := \frac{p(x)}{\prod_i p_i(x_i)} \leq M$ Π_X-a.e., then by the monotonicity of f on $[m, M]$ and Jensen's inequality under Π_X,

$$\mathcal{R}_f(X) \;=\; \mathbb{E}_{\Pi_X}[f(L)] \;\leq\; \max_{t \in [m,M]} f(t) \;\leq\; f(M).$$

For $f(t) = t \log t$, the function $t \log t$ is increasing on $[1, \infty)$ and decreasing on $(0, 1]$. Hence, if $L \leq M$ almost everywhere,

$$\mathbb{E}_{\Pi_X}[L \log L] \;\leq\; M \log M,$$

and a symmetric bound with $m \log m$ holds when $m < 1$ dominates. This establishes the stated boundedness of $\mathcal{R}_f(X)$. $\qquad\square$

A.5 Proof of Proposition 4 (Gaussian Total Correlation)

Proof. Let $X \sim \mathcal{N}(0, \Sigma)$ with correlation matrix $C = \operatorname{diag}(\Sigma)^{-1/2}\, \Sigma\, \operatorname{diag}(\Sigma)^{-1/2}$. Then $P_X = \mathcal{N}(0, \Sigma)$ and $\Pi_X = \bigotimes_i \mathcal{N}(0, \Sigma_{ii})$. The KL divergence between these two Gaussian measures is

$$D_{\mathrm{KL}}\Big(\mathcal{N}(0, \Sigma)\,\Big\|\,\bigotimes_i \mathcal{N}(0, \Sigma_{ii})\Big) \;=\; \frac{1}{2}\Big(\sum_i \log \Sigma_{ii} - \log \det \Sigma\Big).$$

Since $\log \det \Sigma = \sum_i \log \Sigma_{ii} + \log \det C$, this simplifies to

$$D_{\mathrm{KL}}(P_X\|\Pi_X) \;=\; -\tfrac{1}{2} \log \det C,$$

which proves Proposition 4. $\qquad\square$

A.6 Proof of Proposition 5 (Quadratic Approximation and Covariance Form)

Proof. Let $C = I + A$ with $\|A\|_2$ small. For $f(t) = \frac{1}{2}(t-1)^2$, expanding around independence yields the classical χ^2-divergence approximation of $D_f(P_X\|\Pi_X)$, which in the jointly Gaussian or weakly dependent regime reduces to a quadratic form in the off-diagonal covariances. Concretely (e.g., via Edgeworth or Isserlis' formula),

$$\mathcal{R}_f(X) \approx \frac{1}{4}\sum_{i\neq j} C_{ij}^2 = \frac{1}{4}\|C - I\|_F^2.$$

For the Kullback–Leibler case, recall the series expansion

$$\log\det(I+A) = \operatorname{tr}\log(I+A) = \sum_{k\geq 1}\frac{(-1)^{k+1}}{k}\operatorname{tr}(A^k),$$

which converges for $\rho(A) < 1$. Hence,

$$-\log\det C = -\operatorname{tr}\log(I+A) = \frac{1}{2}\operatorname{tr}(A^2) - \frac{1}{3}\operatorname{tr}(A^3) + \cdots = \frac{1}{2}\|A\|_F^2 + O(\|A\|_F^3).$$

Therefore, to second order,

$$-\log\det C \approx \frac{1}{2}\|C - I\|_F^2,$$

which establishes the stated relation between the KL redundancy and the quadratic approximation. □

A.7 Proof of Definition/Proposition 2–6 (Spectral Redundancy)

Proof. Let $\lambda_1, \ldots, \lambda_D$ be the eigenvalues of the covariance matrix $\Sigma_Z \succeq 0$, and define the normalized spectrum

$$\tilde{\lambda}_i = \frac{\lambda_i}{\sum_{j=1}^D \lambda_j}, \qquad \sum_{i=1}^D \tilde{\lambda}_i = 1.$$

The spectral entropy is

$$H_\lambda(Z) = -\sum_{i=1}^D \tilde{\lambda}_i \log\tilde{\lambda}_i,$$

which satisfies $H_\lambda(Z) \in [0, \log D]$, with $H_\lambda(Z) = \log D$ if and only if $\tilde{\lambda}$ is uniform and $H_\lambda(Z) = 0$ if and only if the spectrum is rank one. Setting $r_{\text{eff}}(Z) = \exp(H_\lambda(Z)) \in [1, D]$ and

$$\mathcal{R}_{\text{spec}}(Z) = 1 - \frac{r_{\text{eff}}(Z)}{D},$$

we obtain

$$\mathcal{R}_{\text{spec}}(Z) \in \left[0,\, 1 - \frac{1}{D}\right].$$

The lower and upper extremes correspond, respectively, to the uniform (maximally spread) and rank-one (completely collapsed) spectra. These bounds follow directly from the standard entropy inequalities, completing the proof. □

A.8 Proof of Lemma 1 (Spectral Norm Control)

Proof. Let $Z = WX$ with $\|W\|_2 \leq M$ and $\mathbb{E}\|X\|^2 \leq \sigma^2$. Then

$$\operatorname{tr}(\Sigma_Z) = \mathbb{E}\|Z\|^2 = \mathbb{E}\|WX\|^2 \leq \|W\|_2^2 \mathbb{E}\|X\|^2 \leq M^2\sigma^2.$$

Therefore, the total spectral energy of Z is bounded, implying $r_{\text{eff}}(Z) \in [1, D]$ and $\mathcal{R}_{\text{spec}}(Z) \in [0, 1 - \frac{1}{D}]$ by Definition 2. This establishes the claim. □

A.9 Proof of Lemma 2 (KL vs. Frobenius Near Independence)

Proof. Let $C = I + A$ with spectral radius $\rho(A) < 1$. Then the logarithmic determinant admits the series expansion

$$-\log \det C \;=\; -\mathrm{tr}\log(I+A) \;=\; \sum_{k \geq 2} \frac{(-1)^k}{k}\,\mathrm{tr}(A^k).$$

Using the Hölder–trace inequality $|\mathrm{tr}(A^k)| \leq \|A\|_F^k$ for $k \geq 3$, we obtain

$$-\log \det C \;\geq\; \tfrac{1}{2}\mathrm{tr}(A^2) \;-\; \tfrac{1}{3}\|A\|_F^3 \;=\; \tfrac{1}{2}\|A\|_F^2 \;-\; \tfrac{1}{3}\|A\|_F^3.$$

Thus, for small $\|A\|_F$,

$$\mathcal{R}_{\mathrm{KL}}(X) \;\gtrsim\; \tfrac{1}{2}\|C - I\|_F^2,$$

which proves the stated near-independence bound. $\qquad\square$

A.10 Proof of Lemma 3 (Capacity-Side Bound)

Proof. We have

$$H(Z) \;=\; \sum_i H(Z_i) - \mathrm{TC}(Z) \;=\; \sum_i H(Z_i) - \mathcal{R}_{\mathrm{KL}}(Z).$$

Under the entropy budget $\sum_i H(Z_i) \leq C_0$, it follows that

$$H(Z) \;\leq\; C_0 - \mathcal{R}_{\mathrm{KL}}(Z).$$

Since $I(Z;S) \leq H(Z)$, and standard information–risk inequalities are monotone in $-I(Z;S)$ (e.g., Fano's inequality for classification and the I–MMSE relation for Gaussian regression), there exists an increasing function g_{info} such that

$$\mathcal{E}(R) \;\geq\; g_{\mathrm{info}}(C_0 - R).$$

This establishes the capacity-side bound. $\qquad\square$

A.11 Proof of Lemma 4 (Robustness-Side Bound)

Proof. Consider a channel $X \to Z \to \hat{S}$ operating under bounded stochastic noise, dropout, or adversarial perturbations. When redundancy is nearly zero ($R \approx 0$), the coordinates of Z are almost independent, and the system lacks correlated backups: missing or corrupted components cannot be reconstructed from others. Rate–distortion and coding-theoretic arguments show that introducing controlled correlation (replication or redundancy across coordinates) reduces the expected reconstruction or classification error. Hence, for small R, increasing redundancy strictly improves robustness, implying the existence of a continuous function g_{robust} that is strictly decreasing on $[0, R_1)$ with

$$\mathcal{E}(R) \;\leq\; g_{\mathrm{robust}}(R).$$

This proves the robustness-side bound. $\qquad\square$

A.12 Proof of Theorem 2 (Interior Optimum and Stability Band)

Proof. By assumption, for R near $\underline{R}$ we have $g_{\mathrm{robust}}(R) \leq \mathcal{E}(R)$ with g_{robust} strictly decreasing, and for R near $\overline{R}$ we have $\mathcal{E}(R) \leq g_{\mathrm{info}}(C_0 - R)$ with g_{info} strictly increasing. Hence, for sufficiently small $\varepsilon > 0$,

$$\mathcal{E}'(\underline{R}+\varepsilon) < 0, \qquad \mathcal{E}'(\overline{R}-\varepsilon) > 0.$$

By the continuity of $\mathcal{E}'$, the intermediate value theorem guarantees the existence of $R^* \in (\underline{R}, \overline{R})$ such that $\mathcal{E}'(R^*) = 0$, i.e., a stationary point. Local convexity of the bounding functions g_{info} and g_{robust} transfers to $\mathcal{E}$, implying $\mathcal{E}''(R^*) > 0$; thus R^* is a strict local minimizer.

For the stability band, consider gradient-based dynamics on the model parameters θ_t that monotonically decrease $\mathcal{E}$ under small additive noise (e.g., stochastic gradient descent). Composing with $R(\theta) = \mathcal{R}(f_\theta(X))$, which is locally Lipschitz in θ, yields a one-dimensional stochastic approximation with Lyapunov potential $\mathcal{E}(R)$. Classical results from stochastic approximation theory (e.g., Robbins–Monro or Kushner–Yin) then imply that $R(\theta_t)$ converges to an $O(\sigma)$ neighborhood $[R_{\min}, R_{\max}]$ of the stable minimum R^*, whose width is proportional to the noise scale σ. This establishes both the existence of the interior optimum and the stability band. $\square$

A.13 Proof of Theorem 1 (Existence of Redundancy Equilibrium)

Proof. Let $\Phi(R) = D(R) + \lambda R$ on the compact interval $[\underline{R}, \overline{R}]$. By assumption, D is continuous; hence Φ is continuous and attains a minimum on this interval by the Weierstrass theorem. To show that the minimizer lies in the interior, note that $D'(R)$ exists almost everywhere and is strictly negative on $[\underline{R}, R_1)$ and strictly positive on $(R_2, \overline{R}]$. Thus, for $\lambda > 0$,

$$\Phi'(R) = D'(R) + \lambda$$

is negative sufficiently close to $\underline{R}$ and positive sufficiently close to $\overline{R}$. By the intermediate value theorem, there exists $R^* \in (\underline{R}, \overline{R})$ such that $\Phi'(R^*) = 0$. If $D \in C^2$ with $D''(R^*) > 0$, then $\Phi''(R^*) = D''(R^*) > 0$, so R^* is a strict local minimizer, i.e., a locally stable equilibrium point. This completes the proof. $\square$

A.14 Proof of Corollary 1 (Banded Convergence)

Proof. Consider stochastic gradient descent of the form

$$\theta_{t+1} = \theta_t - \eta_t \big(\nabla_\theta \mathcal{L}(\theta_t) + \xi_t \big),$$

where $\mathbb{E}[\xi_t | \mathcal{F}_t] = 0$, $\mathbb{E}\|\xi_t\|^2 \le \sigma^2$, and the step sizes η_t satisfy the Robbins–Monro conditions. Let $R_t = \mathcal{R}(f_{\theta_t}(X))$. Under local Lipschitz regularity of $R(\theta)$ and strict local convexity of Φ at R^*, the ODE method for stochastic approximation implies that R_t tracks the deterministic flow $\dot{R} = -\Phi'(R)$. The limiting invariant distribution of R_t is therefore concentrated in a neighborhood of R^* whose width scales with σ. Quantitative bounds on this concentration follow from standard Lyapunov drift arguments, establishing the claimed banded convergence. $\square$

A.15 Operational Lemmas Used in Section 3

Total correlation identity. For $f(t) = t \log t$, we have

$$\mathcal{R}_{\mathrm{KL}}(X) = D_{\mathrm{KL}}(P_X \| \Pi_X) = \sum_i H(X_i) - H(X),$$

by the chain rule of entropy.

Quadratic proxy. In near-Gaussian regimes,

$$D_{\chi^2}(P_X \| \Pi_X) = \tfrac{1}{2} \int \frac{(p - \prod_i p_i)^2}{\prod_i p_i} \, dx \approx \tfrac{1}{4} \| C - I \|_F^2,$$

to second order in weak correlations. The proof follows the same cumulant (Isserlis) expansion used in Proposition 5.

Attention-head redundancy. For attention maps $A_h \in \mathbb{R}^{N \times N}$, define

$$\mathcal{R}_{\mathrm{head}} = \frac{1}{H(H-1)} \sum_{h \ne h'} \frac{\langle \mathrm{vec}(A_h), \mathrm{vec}(A_{h'}) \rangle^2}{\|A_h\|_F^2 \, \|A_{h'}\|_F^2}.$$

Since each term is a squared cosine similarity, $\mathcal{R}_{\mathrm{head}} \in [0, 1]$. For any head-wise post-processing represented by a linear Markov operator (e.g., column/row-stochastic smoothing or projections), the data processing

inequality for χ^2-like functionals implies contraction—the squared correlation cannot increase under averaging.

Remark on regularity. All continuity and compactness arguments above can be made fully rigorous by restricting to feasible parameter sets with $R \in [\underline{R}, \overline{R}]$, and by invoking standard envelope theorems for

$$D(R) \;=\; \inf_{\theta:\, \mathcal{R}(\theta)=R} \mathbb{E}\, \ell_{\mathrm{rec}}(\theta).$$

No pathological measurability issues arise under the usual assumptions of bounded loss, dominated model families, or coercive parametrizations.