A CROSS-ENTROPY NOISE PROPAGATION OPERATOR AND THE EFFECTIVE RANK OF THE CLASS-AVERAGED BLOCK AT NEURAL COLLAPSE

LIGHTMAN CHANG

ABSTRACT. We extend the noise-propagation framework of [1]—originally formulated for the squared loss in regression—to multi-class classification under the cross-entropy loss. The technical step is to derive a Fisher-corrected noise propagation operator $\mathbf{M}_{\mathrm{CE}}$ directly from the cross-entropy Hessian $\nabla^2_\theta \ell_{\mathrm{CE}} = \mathbf{J}^\top \mathbf{H} \mathbf{J}$, where $\mathbf{H}(\mathbf{x}) = \mathrm{diag}(\mathbf{p}(\mathbf{x})) - \mathbf{p}(\mathbf{x})\mathbf{p}(\mathbf{x})^\top$ is the output-space Fisher matrix at the softmax probabilities $\mathbf{p}(\mathbf{x})$. Our first result (Theorem 4.1) is that $\mathrm{tr}(\mathbf{M}_{\mathrm{CE}}) < \infty$ if and only if the underlying power-law exponents satisfy $\beta > \alpha + 1/2$, identical to the trichotomy for MSE. We then study the limit of $\mathbf{M}_{\mathrm{CE}}$ along a sequence of trained predictors approaching the Neural Collapse fixed point of Papyan–Han–Donoho. Although $\mathbf{H}$ becomes singular in the strict one-hot limit, the limiting operator is well-defined on the $(K-1)$-dimensional zero-sum subspace fixed by the simplex equiangular tight frame. Theorem 5.5 establishes two facts about the effective rank on this subspace: (a) for the *class-averaged* block $\overline{\mathbf{M}}_{\mathrm{CE}}$, the S_K-symmetry of the simplex ETF combined with the irreducibility of $\mathcal{Z}_K$ as an S_K-representation forces a scalar operator, so $r_{\mathrm{eff}}(\overline{\mathbf{M}}_{\mathrm{CE}})$ equals exactly $K-1$; (b) for an individual per-sample block $\mathbf{M}^{(i)}_{\mathrm{CE}}$, the stabilizer is only S_{K-1} and $\mathcal{Z}_K$ splits into a 1-dimensional "target vs. mean non-target" direction and a $(K-2)$-dimensional "differences among non-targets" subspace, on which $\mathbf{H}$ has two distinct eigenvalues Kab and b; consequently $r_{\mathrm{eff}}(\mathbf{M}^{(i)}_{\mathrm{CE}}) \in [1, K-1]$ in general, with equality only in the symmetric edge case where the two eigenvalue blocks coincide. The class-averaged statement is dimension-free for K-class classification at the terminal phase.

Date: May 7, 2026.

2020 *Mathematics Subject Classification.* Primary 68T07; Secondary 62G05, 41A25, 60J60, 62H30.

Key words and phrases. Cross-entropy loss, Neural Collapse, Fisher information, noise propagation, benign overfitting, effective rank, multi-class classification.

1. Introduction

The unified bound of [1] on the test risk of overparameterized regression involves a noise-propagation operator $\mathbf{M} = \mathbf{\Sigma}^{-1}\mathbf{\Gamma}\mathbf{\Sigma}^{-1}$ built from the SVD of the training Jacobian $\mathbf{J}$ and the test–train cross-correlation $\mathbf{\Gamma}$ of gradient features. The trichotomy $\operatorname{tr}(\mathbf{M}) < \infty \iff \beta > \alpha + 1/2$, under power-law decay $\sigma_j \asymp j^{-\alpha}$ and $\varrho_j := \sqrt{\mathbf{\Gamma}_{jj}} \asymp j^{-\beta}$, characterizes benign, tempered, and catastrophic regimes for squared loss; the power-law model is standard in the random-feature analysis of [8], with bias–variance refinements in the companion [2].

The same generalization story should apply to classification, but the squared-loss derivation does not directly transfer: cross-entropy introduces a nonlinear softmax link, and its parameter Hessian is $\mathbf{J}^\top\mathbf{H}\mathbf{J}$ rather than $\mathbf{J}^\top\mathbf{J}$. The natural question is whether a Fisher-corrected operator preserves the trichotomy and admits an exact effective-rank calculation at the Neural Collapse (NC) fixed point of [10].

This paper answers both questions affirmatively. We define $\mathbf{M}_{\mathrm{CE}}$ as the natural generalization of $\mathbf{M}$ obtained by replacing $\mathbf{J}^\top\mathbf{J}$ with the cross-entropy Hessian (see Definition 3.1 and the derivation that precedes it). We then show:

(i) The CE trichotomy $\operatorname{tr}(\mathbf{M}_{\mathrm{CE}}) < \infty \iff \beta > \alpha + 1/2$ holds whenever the predicted probabilities are bounded away from the one-hot vertices uniformly along the relevant directions (Theorem 4.1).

(ii) Approaching the NC simplex equiangular tight frame fixed point, the *class-averaged* block $\overline{\mathbf{M}}_{\mathrm{CE}}$ on the zero-sum subspace $\mathcal{Z}_K$ has effective rank exactly $K-1$ (Theorem 5.5 (b)). The proof uses Schur's lemma: $\mathcal{Z}_K$ is the $(K-1)$-dimensional standard irreducible S_K-representation, and class-averaging produces an S_K-equivariant operator, hence a scalar.

(iii) For a fixed *per-sample* block $\mathbf{M}_{\mathrm{CE}}^{(i)}$ the stabilizer of class c_i is only S_{K-1}, which splits $\mathcal{Z}_K$ into a 1-dimensional and a $(K-2)$-dimensional isotypic component carrying generically distinct eigenvalues, so $r_{\mathrm{eff}}(\mathbf{M}_{\mathrm{CE}}^{(i)}) \in [1, K-1]$ in general; we give the explicit two-eigenvalue formula (Theorem 5.5 (a)).

The first result extends the trichotomy across loss functions. The second and third together identify the right object whose effective rank is dimension-free at NC: the class-averaged block, not the per-sample block. The averaged statement is what controls generalization at the terminal phase, since the noise-leakage variance $\sigma^2 \operatorname{tr}(\mathbf{M}_{\mathrm{CE}})$ in

Lemma 3.4, and the corresponding effective rank, aggregate over training samples; class-balance averages the per-sample contributions back into a fully S_K-symmetric object. We work conditional on the presence of NC and characterize what this implies for the effective rank, rather than proving NC.

Relation to prior work. The Neural Collapse phenomenon was identified empirically in [10] and characterized mathematically in the unconstrained-features setting under cross-entropy by [7, 9] and under squared loss / central path by [6]. The class-imbalanced extension of unconstrained features is [5]. The cross-entropy Fisher matrix $\mathbf{H} = \mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top$ and its degeneracy at one-hot vertices is classical (see [4]); we use only the finite-rank facts about $\mathbf{H}$ and the simplex ETF geometry.

Organization. Section 2 fixes notation and recalls the classification setting. Section 3 derives $\mathbf{M}_{\mathrm{CE}}$ from the CE Hessian. Section 4 proves the CE trichotomy. Section 5 derives the effective-rank formula at the NC fixed point. Section 6 discusses limitations.

2. Setting and notation

Let $K \geq 2$ be the number of classes. The predictor has K-dimensional output $f(\mathbf{x};\theta) = (f_1(\mathbf{x};\theta), \ldots, f_K(\mathbf{x};\theta)) \in \mathbb{R}^K$, softmax probabilities $p_k(\mathbf{x};\theta) = e^{f_k}/\sum_{k'} e^{f_{k'}}$, and one-hot labels $\mathbf{y}_i \in \{e_1, \ldots, e_K\}$, where e_k is the k-th standard basis vector. Training data $S = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ are i.i.d. The cross-entropy loss is

$$\ell_{\mathrm{CE}}(f, \mathbf{y}) = -\sum_{k=1}^{K} y_k \log p_k(f).$$

A direct computation gives the well-known output-space Hessian

$$\nabla_f^2 \ell_{\mathrm{CE}}(f, \mathbf{y}) = \mathbf{H}(\mathbf{x}) := \mathrm{diag}(\mathbf{p}(\mathbf{x})) - \mathbf{p}(\mathbf{x})\mathbf{p}(\mathbf{x})^\top, \tag{1}$$

which is symmetric positive semidefinite, has rank $K-1$ for any $\mathbf{p}$ in the open simplex (it has a single zero eigenvalue along $\mathbf{1}$), and rank 0 at every one-hot vertex $\mathbf{p} = e_k$ (since at $\mathbf{p} = e_k$ we have $\mathrm{diag}(e_k) = e_k e_k^\top$, hence $\mathbf{H} = \mathrm{diag}(e_k) - e_k e_k^\top = 0$). The Fisher matrix $\mathbf{H}(\mathbf{x})$ is the building block of the parameter Hessian:

$$\nabla_\theta^2 \ell_{\mathrm{CE}}(\mathbf{x}, \mathbf{y}; \theta) = \mathbf{J}(\mathbf{x})^\top \mathbf{H}(\mathbf{x}) \mathbf{J}(\mathbf{x}) + (\text{label-dependent rank-1 term}), \tag{2}$$

where $\mathbf{J}(\mathbf{x}) \in \mathbb{R}^{K \times p}$ is the per-sample Jacobian of $f(\mathbf{x};\cdot)$ with respect to θ. The label-dependent term vanishes in expectation over labels at any predictor that is exact on the population (or at any limit point with vanishing gradient), and we drop it from this point onward.

We adopt the notation of [1]. Let $\mathbf{J}_{\mathrm{CE}} \in \mathbb{R}^{nK\times p}$ denote the stacked training Jacobian with rows $\nabla_\theta f_k(\mathbf{x}_i;\theta^*)^\top$ ranging over $(i,k) \in [n] \times [K]$, and let $\mathbf{H} = \bigoplus_{i=1}^n \mathbf{H}(\mathbf{x}_i)$ denote the block-diagonal sample Fisher. Denote the cross-class test–train cross-correlation

$$(\mathbf{\Gamma}_{\mathrm{CE}})_{(j,k),(j',k')} = \mathbb{E}_{\mathbf{x}}[\psi_{j,k}(\mathbf{x})\psi_{j',k'}(\mathbf{x})], \quad \psi_{j,k}(\mathbf{x}) = \nabla_\theta f_k(\mathbf{x};\theta^*)^\top \mathbf{v}_j, \tag{3}$$

where $\{\mathbf{v}_j\}$ is the right SVD basis of an appropriate factor of $\mathbf{J}_{\mathrm{CE}}$ (see Section 3).

3. Deriving $\mathbf{M}_{\mathrm{CE}}$ from the cross-entropy Hessian

In MSE, the operator $\mathbf{M} = \mathbf{\Sigma}^{-1}\mathbf{\Gamma}\mathbf{\Sigma}^{-1}$ arises because the test risk decomposes as a quadratic form in the predicted residuals, weighted by the inverse of $\mathbf{J}^\top\mathbf{J}$ at the interpolant. Concretely, $\mathbf{\Sigma}$ comes from the SVD of the training Jacobian, and the noise-leakage variance equals $\sigma^2 \operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{\Gamma}\mathbf{\Sigma}^{-1})$.

For cross-entropy at a (near-)interpolating predictor, the linearized test risk is governed by the inverse of the parameter Hessian (2), that is $(\mathbf{J}_{\mathrm{CE}}^\top \mathbf{H} \mathbf{J}_{\mathrm{CE}})^{-1}$. Two natural factorizations arise:

(1) Define the *Fisher-weighted Jacobian*

$$\widetilde{\mathbf{J}} := \mathbf{H}^{1/2}\mathbf{J}_{\mathrm{CE}} \in \mathbb{R}^{nK\times p},$$

with SVD $\widetilde{\mathbf{J}} = \widetilde{\mathbf{U}}\,\widetilde{\mathbf{\Sigma}}\,\widetilde{\mathbf{V}}^\top$. Then $\mathbf{J}_{\mathrm{CE}}^\top\mathbf{H}\mathbf{J}_{\mathrm{CE}} = \widetilde{\mathbf{V}}\,\widetilde{\mathbf{\Sigma}}^2\,\widetilde{\mathbf{V}}^\top$, so the parameter Hessian's eigenstructure coincides with that of $\widetilde{\mathbf{J}}^\top\widetilde{\mathbf{J}}$.

(2) The right singular vectors $\widetilde{\mathbf{V}}$ provide the natural basis for gradient features $\widetilde{\psi}_{j,k}(\mathbf{x}) := \nabla_\theta f_k(\mathbf{x};\theta^*)^\top\widetilde{\mathbf{v}}_j$, and the corresponding cross-correlation matrix is $\widetilde{\mathbf{\Gamma}}_{(j,k),(j',k')} = \mathbb{E}_{\mathbf{x}}[\widetilde{\psi}_{j,k}(\mathbf{x})\widetilde{\psi}_{j',k'}(\mathbf{x})]$. The matrix $\widetilde{\mathbf{\Sigma}}_{\mathrm{CE}}$ refers throughout to the singular value matrix $\widetilde{\mathbf{\Sigma}}$ above; these are the square roots of the eigenvalues of the parameter Hessian.

These two ingredients give the operator that propagates label noise through the linearized CE risk:

Definition 3.1 (Cross-entropy noise propagation operator)**.** The *cross-entropy noise propagation operator* is

$$\mathbf{M}_{\mathrm{CE}} := \widetilde{\mathbf{\Sigma}}^{-1}\,\widetilde{\mathbf{\Gamma}}\,\widetilde{\mathbf{\Sigma}}^{-1}, \qquad \widetilde{\mathbf{J}} = \mathbf{H}^{1/2}\mathbf{J}_{\mathrm{CE}} = \widetilde{\mathbf{U}}\widetilde{\mathbf{\Sigma}}\widetilde{\mathbf{V}}^\top, \tag{4}$$

where the inverse on the right is the Moore–Penrose pseudoinverse restricted to the range of $\widetilde{\mathbf{\Sigma}}$.

Definition 3.2 (Effective rank)**.** For any non-zero positive semidefinite operator $\mathbf{A}$ on a finite-dimensional inner-product space, the *effective rank* is

$$r_{\text{eff}}(\mathbf{A}) \;:=\; \frac{(\operatorname{tr}\mathbf{A})^2}{\|\mathbf{A}\|_F^2}.$$

This quantity satisfies $1 \le r_{\text{eff}}(\mathbf{A}) \le \operatorname{rank}(\mathbf{A})$, with equality on the right iff all nonzero eigenvalues are equal, and is invariant under positive scalar rescaling of $\mathbf{A}$.

Remark 3.3 (Equivalence with the form $\mathbf{F}^{-1}\mathbf{\Gamma}\mathbf{F}^{-1}$). Some prior write-ups state $\mathbf{M}_{\text{CE}} = \mathbf{F}^{-1}\mathbf{\Gamma}_{\text{CE}}\mathbf{F}^{-1}$ with $\mathbf{F} = \mathbf{H}^{1/2}\mathbf{\Sigma}_{\text{CE}}$, where $\mathbf{\Sigma}_{\text{CE}}$ denotes the singular values of $\mathbf{J}_{\text{CE}}$. This is the same operator as (4) written in the basis of $\mathbf{J}_{\text{CE}}$ rather than $\widetilde{\mathbf{J}}$: $\widetilde{\mathbf{\Sigma}}$ are precisely the singular values of $\mathbf{H}^{1/2}\mathbf{J}_{\text{CE}}$, which equal the square roots of the eigenvalues of $\mathbf{J}_{\text{CE}}^\top\mathbf{H}\mathbf{J}_{\text{CE}}$. We retain Definition 3.1 as the canonical form because the factorization is unambiguous.

Cross-reference to [3]. The symbols $\mathbf{M}, \mathbf{\Sigma}, \mathbf{\Gamma}$ used in this paper are the direct cross-entropy analogs of the squared-loss operators in the companion paper [3]. The dictionary is: $\mathbf{\Sigma}$ (singular values of $\mathbf{J}$) $\leftrightarrow$ $\widetilde{\mathbf{\Sigma}}$ (singular values of $\widetilde{\mathbf{J}} = \mathbf{H}^{1/2}\mathbf{J}_{\text{CE}}$); $\mathbf{\Gamma}$ (test–train cross-correlation in the right singular basis of $\mathbf{J}$) $\leftrightarrow$ $\widetilde{\mathbf{\Gamma}}$ (in the right singular basis of $\widetilde{\mathbf{J}}$); and $\mathbf{M} = \mathbf{\Sigma}^{-1}\mathbf{\Gamma}\mathbf{\Sigma}^{-1} \leftrightarrow \mathbf{M}_{\text{CE}} = \widetilde{\mathbf{\Sigma}}^{-1}\widetilde{\mathbf{\Gamma}}\widetilde{\mathbf{\Sigma}}^{-1}$. Setting $\mathbf{H} \equiv \mathbf{I}$ in the cross-entropy definitions recovers the squared-loss objects of [3] verbatim.

The justification for Definition 3.1 is the following linearization lemma, which derives the noise-leakage variance from a Taylor expansion of the empirical CE risk near θ^*. The argument is the cross-entropy analog of the MSE template in [1, §2]; the loss-specific ingredient is the substitution of the CE Hessian $\mathbf{J}^\top\mathbf{H}\mathbf{J}$ for the squared-loss Hessian $\mathbf{J}^\top\mathbf{J}$, equivalently the replacement of $\mathbf{J}$ by $\widetilde{\mathbf{J}} = \mathbf{H}^{1/2}\mathbf{J}$.

Lemma 3.4 (Linearized noise leakage under CE)**.** *Let θ^* be a stationary point of the empirical CE risk on the training set S, with parameter Hessian $\mathbf{F} := \mathbf{J}_{\text{CE}}^\top\mathbf{H}\mathbf{J}_{\text{CE}}$ invertible on its range. Suppose label perturbations $\boldsymbol{\xi}_i \in \mathbb{R}^K$ satisfy $\mathbb{E}[\boldsymbol{\xi}_i] = 0$ and $\mathbb{E}[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top] = \sigma^2\mathbf{I}_K$, so that $\mathbf{H}^{1/2}\boldsymbol{\xi}_i$ is the effective Fisher-weighted label noise that matches the CE link. Let $\theta^*_{\boldsymbol{\xi}}$ denote the perturbed stationary point. Then to linear order in $\boldsymbol{\xi}$,*

- *(i) the parameter perturbation is $\delta\theta = \mathbf{F}^\dagger\mathbf{J}_{\text{CE}}^\top\mathbf{H}^{1/2}\boldsymbol{\xi} = \widetilde{\mathbf{V}}\widetilde{\mathbf{\Sigma}}^{-1}\widetilde{\mathbf{U}}^\top\boldsymbol{\xi}$;*
- *(ii) along any gradient feature direction $\widetilde{\mathbf{v}}_j$ the variance contribution to the test residual at $\mathbf{x}_*$ equals $\sigma^2\widetilde{\mathbf{\Gamma}}_{jj}/\widetilde{\sigma}_j^2$;*
- *(iii) summing,* $\operatorname{Var}(\textit{linearized test residual}) = \sigma^2 \operatorname{tr}(\mathbf{M}_{\text{CE}})$.

Proof. Linearize the empirical-risk gradient near θ^*:

$$\nabla_\theta \widehat{\mathcal{R}}_{\mathrm{CE}}(\theta^* + \delta\theta;\, \mathbf{y} + \boldsymbol{\xi}) = \mathbf{F}\,\delta\theta - \mathbf{J}_{\mathrm{CE}}^\top \mathbf{H}^{1/2}\boldsymbol{\xi} + O(\|\delta\theta\|^2 + \|\boldsymbol{\xi}\|^2),$$

using $\nabla_\theta \ell_{\mathrm{CE}}(\mathbf{x}, \mathbf{y}; \theta) = \mathbf{J}(\mathbf{x})^\top(\mathbf{p}(\mathbf{x};\theta) - \mathbf{y})$ and the noise model. Setting the gradient to zero gives $\delta\theta = \mathbf{F}^\dagger \mathbf{J}_{\mathrm{CE}}^\top \mathbf{H}^{1/2}\boldsymbol{\xi}$, which is (i); substituting the SVD $\widetilde{\mathbf{J}} = \widetilde{\mathbf{U}}\widetilde{\boldsymbol{\Sigma}}\widetilde{\mathbf{V}}^\top$ into $\mathbf{F} = \widetilde{\mathbf{V}}\widetilde{\boldsymbol{\Sigma}}^2\widetilde{\mathbf{V}}^\top$ yields the explicit form. For (ii), the test residual contribution along $\widetilde{\mathbf{v}}_j$ is $\widetilde{\psi}_{j,\cdot}(\mathbf{x}_*)^\top \delta\theta$, whose variance is $\sigma^2 \widetilde{\sigma}_j^{-2} \widetilde{\boldsymbol{\Gamma}}_{jj}$ by orthonormality of $\widetilde{\mathbf{U}}$. For (iii), sum to recognize $\sum_j \widetilde{\boldsymbol{\Gamma}}_{jj}/\widetilde{\sigma}_j^2 = \mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{\Gamma}}\widetilde{\boldsymbol{\Sigma}}^{-1}) = \mathrm{tr}(\mathbf{M}_{\mathrm{CE}})$. □

This lemma plays the same role as the squared-loss derivation in [1]: $\mathbf{M}_{\mathrm{CE}}$ is exactly the quadratic form that controls noise-leakage variance under CE, with $\mathbf{H}$ replacing the identity in the output-space inner product.

4. Cross-entropy trichotomy

Theorem 4.1 (Cross-entropy trichotomy)**.** *Suppose the singular values $\widetilde{\sigma}_j$ of $\widetilde{\mathbf{J}} = \mathbf{H}^{1/2}\mathbf{J}_{\mathrm{CE}}$ and the diagonal entries $\widetilde{\boldsymbol{\Gamma}}_{jj}$ (in the right singular basis of $\widetilde{\mathbf{J}}$) exhibit power-law decay $\widetilde{\sigma}_j \asymp j^{-\alpha}$, $\widetilde{\varrho}_j := \sqrt{\widetilde{\boldsymbol{\Gamma}}_{jj}} \asymp j^{-\beta}$. Suppose furthermore that the Fisher block is uniformly non-degenerate in the sense of the operator inequality*

$$\mathbf{H}(\mathbf{x}_i) \succeq h_{\min}\, \mathbf{P}_{\mathcal{Z}_K} \qquad \text{for all } i = 1, \dots, n, \tag{5}$$

for some $h_{\min} > 0$, where $\mathbf{P}_{\mathcal{Z}_K}$ is the orthogonal projection onto the zero-sum subspace $\mathcal{Z}_K = \{u \in \mathbb{R}^K : \mathbf{1}^\top u = 0\}$. Then

$$\mathrm{tr}(\mathbf{M}_{\mathrm{CE}}) < \infty \iff \beta > \alpha + \tfrac{1}{2}.$$

Proof. By Definition 3.1,

$$\mathrm{tr}(\mathbf{M}_{\mathrm{CE}}) = \sum_j \widetilde{\boldsymbol{\Gamma}}_{jj}/\widetilde{\sigma}_j^2 \asymp \sum_j j^{-2\beta}/j^{-2\alpha} = \sum_j j^{2(\alpha-\beta)},$$

which converges iff $2(\alpha - \beta) < -1$, i.e. $\beta > \alpha + 1/2$. This is a statement about the singular values of $\widetilde{\mathbf{J}}$ and the diagonal of $\widetilde{\boldsymbol{\Gamma}}$ in $\widetilde{\mathbf{J}}$'s right singular basis, so no basis conversion is needed. Hypothesis (5) ensures that $\widetilde{\mathbf{J}} = \mathbf{H}^{1/2}\mathbf{J}_{\mathrm{CE}}$ has full rank on the zero-sum subspace whenever $\mathbf{J}_{\mathrm{CE}}$ does, with smallest nonzero singular value at least $h_{\min}^{1/2}$ times that of $\mathbf{J}_{\mathrm{CE}}$ on $\mathcal{Z}_K$, so the power-law hypotheses are equivalent to power laws on $\mathbf{J}_{\mathrm{CE}}$ up to a multiplicative constant. □

Remark 4.2 (On the uniform Fisher lower bound). Hypothesis (5) is the cross-entropy analog of the "well-conditioned predictor" assumption:

at any predictor whose softmax outputs $\mathbf{p}(\mathbf{x})$ avoid the boundary of the simplex, the nonzero eigenvalues of $\mathbf{H}(\mathbf{x})$ are bounded below by a positive constant on $\mathcal{Z}_K$. For a stable training trajectory away from the zero-margin classifier, this is automatic with $h_{\min}$ controlled by the entropy of the outputs. The hypothesis fails in the strict NC limit; we treat that limit separately in Section 5 via a different argument.

Remark 4.3. The matrix $\mathbf{\Sigma}_{\mathrm{CE}}$ in the form $\mathbf{M}_{\mathrm{CE}} = \mathbf{F}^{-1}\mathbf{\Gamma}_{\mathrm{CE}}\mathbf{F}^{-1}$ of Remark 3.3 denotes the diagonal matrix of singular values of the (unweighted) Jacobian $\mathbf{J}_{\mathrm{CE}}$; under hypothesis (5), the singular values of $\widetilde{\mathbf{J}}$ relate to those of $\mathbf{J}_{\mathrm{CE}}$ via $\widetilde{\sigma}_j \in [h_{\min}^{1/2}\sigma_j^{\mathrm{CE}},\, h_{\max}^{1/2}\sigma_j^{\mathrm{CE}}]$ where $h_{\max} = \|\mathbf{H}\|_{\mathrm{op}}$.

5. Effective rank at the Neural Collapse fixed point

We now compute the effective rank $r_{\mathrm{eff}}(\mathbf{M}_{\mathrm{CE}})$ in the limit where the trained predictor approaches the Neural Collapse fixed point of [10]. A subtlety: in the strict infinite-margin limit (logits diverging in the simplex ETF directions), the softmax output for each sample becomes one-hot, $\mathbf{H}(\mathbf{x}) \to 0$, and $\mathbf{M}_{\mathrm{CE}}$ degenerates. We address this by considering a sequence of predictors whose last-layer features approach the NC fixed point with *uniformly bounded logit norms*; in this regime the limiting $\mathbf{p}(\mathbf{x})$ is a probability vector strictly in the open simplex, peaked on the correct class but with positive mass on every coordinate. This is the regime relevant to NC phenomenology at any finite training time, and the one in which the unconstrained-features analyses of [7, 9, 5] yield well-defined limits. The strict-margin limit is treated separately by rescaling (Remark 5.2).

5.1. Geometry of the simplex equiangular tight frame.

Recall the NC fixed point: the last-layer features satisfy $h(\mathbf{x}_i) = \mu_{c(\mathbf{x}_i)}$ where $\{\mu_c\}_{c=1}^K$ forms a simplex equiangular tight frame in $\mathbb{R}^{K-1}$ with $\sum_c \mu_c = 0$ and $\mu_c^\top \mu_{c'} = \frac{K\delta_{cc'}-1}{K-1}\,\|\mu\|^2$ for some common norm $\|\mu\|$. The classifier weights $\{\mathbf{w}_c\}$ likewise satisfy $\mathbf{w}_c = (\|\mathbf{w}\| \,/\, \|\mu\|)\mu_c$ up to a scalar.

The image of the per-class logit difference map $h \mapsto (\mathbf{w}_1^\top h, \dots, \mathbf{w}_K^\top h)$ at the NC fixed point therefore lies in the $(K-1)$-dimensional zero-sum subspace

$$\mathcal{Z}_K := \{u \in \mathbb{R}^K : \mathbf{1}^\top u = 0\}.$$

Let $\mathbf{P}_{\mathcal{Z}}$ be the orthogonal projection onto $\mathcal{Z}_K$, that is $\mathbf{P}_{\mathcal{Z}} = \mathbf{I}_K - \frac{1}{K}\mathbf{1}\mathbf{1}^\top$.

5.2. Limiting Fisher and limiting cross-correlation.

Lemma 5.1 (Fisher matrix under bounded-logit NC)**.** *Consider a sequence of predictors $\{\theta^{(t)}\}$ whose last-layer features approach the NC fixed point and whose logit vectors $f(\mathbf{x}_i;\theta^{(t)}) \in \mathbb{R}^K$ satisfy $\left\|f(\mathbf{x}_i;\theta^{(t)})\right\|_\infty \le$*

L for some finite $L > 0$. Then for any sample $\mathbf{x}_i$ and along any limit point of $\{\theta^{(t)}\}$, every coordinate of the predicted probability $\mathbf{p}(\mathbf{x}_i; \theta^{(t)})$ is bounded below by $p_{\min} := e^{-2L}/K > 0$, the matrix $\mathbf{H}(\mathbf{x}_i)$ has rank exactly $K-1$ with kernel $\operatorname{span}(\mathbf{1})$, and the operator inequality

$$\mathbf{H}(\mathbf{x}_i) \;\succeq\; h_{\min}\,\mathbf{P}_{\mathcal{Z}_K}, \qquad h_{\min} := (K-1)\,p_{\min}^2,$$

holds uniformly in i and t.

Proof. For any logit vector f with $\|f\|_\infty \le L$, the softmax probabilities satisfy $p_k = e^{f_k}/\sum_{k'} e^{f_{k'}} \ge e^{-L}/(Ke^{L}) = e^{-2L}/K$, giving $p_{\min} > 0$. The Fisher matrix $\mathbf{H} = \operatorname{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top$ satisfies $\mathbf{H}\mathbf{1} = \mathbf{p} - \mathbf{p}(\mathbf{1}^\top\mathbf{p}) = 0$, so $\operatorname{span}(\mathbf{1}) \subseteq \ker \mathbf{H}$, with equality on the open simplex (a standard fact about the multinomial Fisher matrix; see e.g. [4, §3]). For the lower eigenvalue bound on $\mathcal{Z}_K$, the classical estimate is

$$\lambda_{\min}\big(\mathbf{H}|_{\mathcal{Z}_K}\big) \;\ge\; p_{\min}\,(1 - p_{\max}),$$

which follows from the variational characterization. Since $p_{\max} = 1 - \sum_{k \ne \arg\max} p_k \le 1 - (K-1)p_{\min}$, we obtain

$$\lambda_{\min}\big(\mathbf{H}|_{\mathcal{Z}_K}\big) \;\ge\; p_{\min}\,\big(1 - (1 - (K-1)p_{\min})\big) \;=\; (K-1)\,p_{\min}^2.$$

(The chain "$p_{\max} \le 1-(K-1)p_{\min} \le 1-1/K$" that might be tempting requires $p_{\min} \ge 1/K$ and is generally false: $p_{\min} = e^{-2L}/K < 1/K$ whenever $L > 0$. Hence we retain the sharper bound $h_{\min} = (K-1)p_{\min}^2$, which is correct for all $L > 0$.) □

Remark 5.2 (Strict-one-hot limit and rescaling). In the strict infinite-margin limit ($L \to \infty$), the bounded-logit condition fails, $\mathbf{p} \to e_c$, and $\mathbf{H} \to 0$, so $\widetilde{\mathbf{J}} \to 0$ and Definition 3.1 degenerates as written. The natural rescaled object is

$$\widetilde{\mathbf{J}}_{\text{rsc}} := h_{\min}^{-1/2}\widetilde{\mathbf{J}} \;=\; (h_{\min}^{-1/2}\mathbf{H}^{1/2})\,\mathbf{J}_{\text{CE}}.$$

The operator $\mathbf{M}_{\text{CE}}$ rebuilt from $\widetilde{\mathbf{J}}_{\text{rsc}}$ has the same effective rank as the unrescaled $\mathbf{M}_{\text{CE}}$, since r_{eff} is invariant under positive scalar rescaling of $\widetilde{\mathbf{\Sigma}}$. We caution that $h_{\min}^{-1/2}\mathbf{H}^{1/2}$ does *not* converge to the projector $\mathbf{P}_{\mathcal{Z}_K}$: at NC the per-sample probabilities $\mathbf{p}(\mathbf{x}_i)$ are of the form $(a, b, \ldots, b)$ with $a + (K-1)b = 1$ and $a > b$, and the eigenvalues of $\mathbf{H}(\mathbf{x}_i)|_{\mathcal{Z}_K}$ are Kab on the 1-dimensional "target vs. mean non-target" direction and b on the $(K-2)$-dimensional "differences among non-targets" subspace (direct calculation, see Step 2 of the proof of Theorem 5.5). The ratio $Kab/b = Ka$ remains finite (and equals K in the strict $a \to 1$ limit), so $h_{\min}^{-1/2}\mathbf{H}^{1/2}$ rescales the trace but not the eigenvalue ratio: rescaling by $h_{\min}$ does not symmetrize the spectrum. This is what forces us to track the two eigenvalue blocks separately in Theorem 5.5, rather than

collapsing them by a strict-NC limit. Below we work with finite logit scale throughout.

5.3. **Effective rank formula.** Before stating the theorem, we make the structural NC hypotheses explicit. The simplex-ETF geometry of [10] fixes the last-layer features and classifier weights, but the per-sample Jacobian also depends on the backbone (everything below the last layer). The following assumption, standard in the unconstrained-features literature [7, 9, 5], isolates what is needed.

Assumption 5.3 (NC backbone non-degeneracy)**.** At θ^*, with last-layer feature map $h(\mathbf{x};\theta^*) \in \mathbb{R}^{K-1}$ and last-layer classifier $W \in \mathbb{R}^{K\times(K-1)}$ in simplex-ETF configuration:

(a) (Class collapse / NC1) For every sample $\mathbf{x}_i$ in class $c_i \in \{1,\dots,K\}$, $h(\mathbf{x}_i;\theta^*) = \mu_{c_i}$ where $\{\mu_c\}$ is the simplex ETF ($\sum_c \mu_c = 0$, $\mu_c^\top \mu_{c'} = (K\delta_{cc'} - 1)/(K-1) \cdot \|\mu\|^2$).
(b) (Self-duality / NC2) $W = (\|W\| / \|\mu\|)\, M$ where $M^\top = [\mu_1 \cdots \mu_K]$.
(c) (Backbone full-rank in ETF directions) The backbone Jacobian $\partial h(\mathbf{x}_i;\theta^*)/\partial\theta \in \mathbb{R}^{(K-1)\times p}$ has full row rank $K-1$ for every i.

Hypothesis (c) is the chain-rule input to the per-sample Jacobian $\mathbf{J}(\mathbf{x}_i) \in \mathbb{R}^{K\times p}$; it follows in the unconstrained-features regime where h is a free variable [9], and is the operative assumption in feature-learning analyses of NC. We do not derive it from NC1–NC4; we assume it.

We measure the effective rank on the per-sample block (equivalently, on the output-averaged operator), which is the geometric object fixed by simplex-ETF symmetry and is independent of n.

Definition 5.4 (Per-sample / averaged $\mathbf{M}_{\mathrm{CE}}$ block)**.** Let $\mathbf{M}_{\mathrm{CE}}^{(i)} \in \mathbb{R}^{K\times K}$ denote the per-sample restriction of $\mathbf{M}_{\mathrm{CE}}$ to the output block indexed by sample i, viewed as an operator on $\mathbb{R}^K$. Define the *averaged operator*

$$\overline{\mathbf{M}}_{\mathrm{CE}} := \frac{1}{n}\sum_{i=1}^{n} \mathbf{M}_{\mathrm{CE}}^{(i)}.$$

Theorem 5.5 (Effective rank at NC)**.** *Let θ^* be a limit point of the training trajectory of Lemma 5.1 satisfying Assumption 5.3, and assume the training set is class-balanced (n/K samples per class).*

*(a) (**Per-sample block, two-eigenvalue formula.**) For each sample i in class c_i, the per-sample block $\mathbf{M}_{\mathrm{CE}}^{(i)}$ vanishes on* $\mathrm{span}(\mathbf{1})$ *and on $\mathcal{Z}_K$ has at most two distinct eigenvalues:*

$$\lambda_1^{(i)} = \frac{\gamma_1^{(i)}}{K\,a\,b} \quad \textit{(multiplicity 1, on the target-vs-non-target axis)},$$

$\lambda_2^{(i)} = \dfrac{\gamma_2^{(i)}}{b}$ *(multiplicity $K-2$, on differences among non-targets), where $a, b > 0$ are the target / non-target softmax probabilities ($a + (K-1)b = 1$), and $\gamma_1^{(i)}, \gamma_2^{(i)} > 0$ are the corresponding eigenvalues of the per-sample $\widetilde{\mathbf{\Gamma}}$-block under the S_{K-1}-isotypic decomposition. Consequently*

$$r_{\text{eff}}\big(\mathbf{M}_{\text{CE}}^{(i)}\big) \;=\; \frac{\big(\lambda_1^{(i)} + (K-2)\lambda_2^{(i)}\big)^2}{(\lambda_1^{(i)})^2 + (K-2)(\lambda_2^{(i)})^2} \;\in\; [1,\, K-1],$$

with the upper bound $K-1$ attained iff $\lambda_1^{(i)} = \lambda_2^{(i)}$ and the lower bound approached as one eigenvalue dominates the other. (For $K = 2$ the second block is absent and $r_{\text{eff}} = 1$; for $K \geq 3$ the range is non-degenerate.)

(b) *(**Class-averaged block, exact** $K-1$.) The class-averaged operator $\overline{\mathbf{M}}_{\text{CE}}$ on $\mathcal{Z}_K$ is a scalar:*

$$\overline{\mathbf{M}}_{\text{CE}}\big|_{\mathcal{Z}_K} = \kappa\, \mathbf{P}_{\mathcal{Z}_K}, \qquad \kappa > 0,$$

hence

$$r_{\text{eff}}\big(\overline{\mathbf{M}}_{\text{CE}}\big) \;=\; K-1.$$

Proof. Step 1: Per-sample block has rank exactly $K-1$ on $\mathcal{Z}_K$. By Lemma 5.1, $\mathbf{H}(\mathbf{x}_i)$ is positive definite on $\mathcal{Z}_K$ and zero on span($\mathbf{1}$). By Assumption 5.3(b)–(c), the model output map $f(\mathbf{x}_i;\theta) = W\,h(\mathbf{x}_i;\theta)$ has the property that varying W along the simplex-ETF directions sweeps out the full $(K-1)$-dimensional zero-sum subspace $\mathcal{Z}_K$ in output space (since the rescaled rows of W, namely $(\|W\| \,/\, \|\mu\|)\mu_c$ for $c \in [K]$, are exactly the K vertices of the simplex ETF in $\mathcal{Z}_K$ and span it). Hence the projected Jacobian $\mathbf{P}_{\mathcal{Z}_K}\mathbf{J}(\mathbf{x}_i)$ has row rank exactly $K-1$ at any predictor satisfying the NC backbone hypothesis, and the same holds for $\widetilde{\mathbf{J}}(\mathbf{x}_i) = \mathbf{H}(\mathbf{x}_i)^{1/2}\mathbf{J}(\mathbf{x}_i)$.

Step 2: Eigenstructure of $\mathbf{H}(\mathbf{x}_i)$ on $\mathcal{Z}_K$. At NC, the predicted probability for sample $\mathbf{x}_i$ in class c_i is permutation-symmetric across the $K-1$ non-target classes (since the corresponding logits $\mathbf{w}_c^\top \mu_{c_i} = (\|\mathbf{w}\| \,/\, \|\mu\|)\mu_c^\top \mu_{c_i}$ take the common value $-(\|\mathbf{w}\|\,\|\mu\|)/(K-1)$ for all $c \neq c_i$ by the simplex-ETF inner product). Thus $\mathbf{p}(\mathbf{x}_i)$ has the form $(\ldots, a, \ldots)$ with the target coordinate a at position c_i and every non-target coordinate equal to b, with $a + (K-1)b = 1$ and $a > b > 0$ (strict inequalities by the bounded-logit Lemma 5.1 and the assumption that NC is non-trivial, i.e., not the uniform prediction $a = b = 1/K$).

Decompose $\mathcal{Z}_K$ into the two S_{K-1}-isotypic components fixed by permutations of the non-target indices:

- $V_1^{(i)} := \operatorname{span}\big((K-1)\,e_{c_i} - \sum_{c\neq c_i} e_c\big)$, the 1-dimensional "target vs. mean of non-targets" direction;
- $V_2^{(i)} := \{u \in \mathcal{Z}_K : u_{c_i} = 0\}$, the $(K-2)$-dimensional "differences among non-target coordinates" subspace.

A direct computation (using $a+(K-1)b = 1$, equivalently $1-(a-b) = Kb$ and $1+(K-1)(a-b) = Ka$) gives

$$\mathbf{H}(\mathbf{x}_i)\,v \;=\; Kab\,v \quad \forall v \in V_1^{(i)}, \qquad \mathbf{H}(\mathbf{x}_i)\,v \;=\; b\,v \quad \forall v \in V_2^{(i)}.$$

For $v \in V_2^{(i)}$ this is immediate from $\mathbf{H}v = \operatorname{diag}(\mathbf{p})v - \mathbf{p}(\mathbf{p}^\top v) = bv - 0$ since $\mathbf{p}^\top v = b\sum_{c\neq c_i} v_c = 0$ on $V_2^{(i)}$. For $v_1 = (K-1)e_{c_i} - \sum_{c\neq c_i} e_c \in V_1^{(i)}$, $\operatorname{diag}(\mathbf{p})v_1 = (K-1)a\,e_{c_i} - b\sum_{c\neq c_i} e_c$ and $\mathbf{p}^\top v_1 = (K-1)(a-b)$, so

$$\begin{aligned}
\mathbf{H}(\mathbf{x}_i)v_1 &= (K-1)a\,e_{c_i} - b\sum_{c\neq c_i} e_c \;-\; (K-1)(a-b)\big[a\,e_{c_i} + b\sum_{c\neq c_i} e_c\big] \\
&= e_{c_i}\,(K-1)a\big[1-(a-b)\big] \;+\; \sum_{c\neq c_i} e_c\,(-b)\big[1+(K-1)(a-b)\big] \\
&= e_{c_i}\,(K-1)a\,(Kb) \;-\; \sum_{c\neq c_i} e_c\,b\,(Ka) = Kab\,v_1.
\end{aligned}$$

The two eigenvalues Kab (on $V_1^{(i)}$) and b (on $V_2^{(i)}$) are equal iff $Ka = 1$, i.e., $a = 1/K$ (uniform prediction), which is excluded at non-trivial NC. The earlier draft of this paper claimed these eigenvalues are equal by simplex-ETF symmetry; that claim is incorrect because the stabilizer of class c_i is only S_{K-1}, not transitive on $\mathcal{Z}_K$.

Step 3: Eigenstructure of the per-sample $\widetilde{\mathbf{\Gamma}}$ block. By Assumption 5.3, the gradient features $\widetilde{\psi}_{j,k}(\mathbf{x})$ commute with the S_{K-1}-action that permutes the non-target indices $c \neq c_i$ (the simplex ETF and the backbone are both invariant under this action; only the target index c_i is distinguished). Hence the per-sample block of $\widetilde{\mathbf{\Gamma}}$ on $\mathcal{Z}_K$ commutes with this S_{K-1}-action and, by Schur's lemma applied to the S_{K-1}-isotypic decomposition $\mathcal{Z}_K = V_1^{(i)} \oplus V_2^{(i)}$ (both isotypic components carry inequivalent S_{K-1}-representations: $V_1^{(i)}$ is the trivial rep and $V_2^{(i)}$ is the standard $(K-2)$-dimensional irreducible rep of S_{K-1}), is block-diagonal with eigenvalues $\gamma_1^{(i)} \geq 0$ on $V_1^{(i)}$ (multiplicity 1) and $\gamma_2^{(i)} \geq 0$ on $V_2^{(i)}$ (multiplicity $K-2$). We assume $\gamma_1^{(i)}, \gamma_2^{(i)} > 0$ (i.e., the gradient features genuinely populate both blocks); this is generic.

Step 4: Per-sample r_{eff} formula. The per-sample block $\mathbf{M}_{\text{CE}}^{(i)} = \widetilde{\mathbf{\Sigma}}^{-1}\widetilde{\mathbf{\Gamma}}\widetilde{\mathbf{\Sigma}}^{-1}$ restricted to $\mathcal{Z}_K$ is, in the basis $V_1^{(i)} \oplus V_2^{(i)}$, the diagonal

matrix with entries $\gamma_1^{(i)}/(Kab)$ on $V_1^{(i)}$ and $\gamma_2^{(i)}/b$ on $V_2^{(i)}$. Calling these $\lambda_1^{(i)}$ and $\lambda_2^{(i)}$ respectively,

$$r_{\text{eff}}(\mathbf{M}_{\text{CE}}^{(i)}) = \frac{(\lambda_1^{(i)} + (K-2)\lambda_2^{(i)})^2}{(\lambda_1^{(i)})^2 + (K-2)(\lambda_2^{(i)})^2}.$$

Cauchy–Schwarz applied to the vectors $(1, \sqrt{K-2})$ and $(\lambda_1^{(i)}, \sqrt{K-2}\,\lambda_2^{(i)})$ gives the upper bound $1 + (K-2) = K-1$, attained iff $\lambda_1^{(i)} = \lambda_2^{(i)}$. As $\lambda_1^{(i)}/\lambda_2^{(i)} \to \infty$ the ratio tends to 1; as $\lambda_1^{(i)}/\lambda_2^{(i)} \to 0$ it tends to $K-2$. Hence the lower bound is 1 for $K \geq 3$. This proves (a).

Step 5: Averaged operator is scalar by Schur's lemma. The class-averaged block is $\overline{\mathbf{M}}_{\text{CE}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{M}_{\text{CE}}^{(i)}$ with class-balanced training data (n/K samples per class). By the simplex-ETF symmetry of NC (Assumption 5.3 (a)–(b)), the data distribution and the predictor are invariant under the full permutation group S_K acting simultaneously on classes, last-layer features, and rows of W. The per-sample blocks $\mathbf{M}_{\text{CE}}^{(i)}$ depend on i only through the class c_i, and the averaged operator $\overline{\mathbf{M}}_{\text{CE}}$ commutes with the diagonal S_K-action on $\mathbb{R}^K$.

The standard fact about the $(K-1)$-dimensional zero-sum subspace $\mathcal{Z}_K \subset \mathbb{R}^K$ is that it is the *standard irreducible representation* of S_K (see, e.g., [4, §2.3] or any text on symmetric-group representations). By Schur's lemma, every S_K-equivariant endomorphism of an irreducible representation is a scalar. Hence $\overline{\mathbf{M}}_{\text{CE}}|_{\mathcal{Z}_K} = \kappa\,\mathbf{P}_{\mathcal{Z}_K}$ for some scalar $\kappa > 0$ (positivity follows from the fact that each $\mathbf{M}_{\text{CE}}^{(i)}$ is positive semidefinite and at least one per-sample block has $\gamma_j^{(i)} > 0$). Therefore all $K-1$ nonzero eigenvalues of $\overline{\mathbf{M}}_{\text{CE}}$ are equal, and by Definition 3.2,

$$r_{\text{eff}}(\overline{\mathbf{M}}_{\text{CE}}) = \frac{((K-1)\kappa)^2}{(K-1)\kappa^2} = K-1.$$

This proves (b). □

Remark 5.6 (How does class-averaging restore $K-1$?). The mechanism is purely group-theoretic. Each per-sample block $\mathbf{M}_{\text{CE}}^{(i)}$ has only an S_{K-1} symmetry (the stabilizer of class c_i), and $\mathcal{Z}_K$ decomposes as $V_1^{(i)} \oplus V_2^{(i)}$ with two distinct eigenvalues. But the target direction $V_1^{(c)}$ moves as c varies over the K classes: explicitly, $V_1^{(c)} = \text{span}(Ke_c - \mathbf{1})$ and one computes

$$\sum_{c=1}^{K} \frac{(Ke_c - \mathbf{1})(Ke_c - \mathbf{1})^\top}{\|Ke_c - \mathbf{1}\|^2}\bigg|_{\mathcal{Z}_K} = \frac{K}{K-1}\,\mathbf{P}_{\mathcal{Z}_K}.$$

That is, summing the rank-one projectors onto target directions across all classes yields a scalar on $\mathcal{Z}_K$. Combined with the corresponding sum over the $V_2^{(c)}$-blocks (also scalar by symmetry), the class-average produces an operator with a single eigenvalue across all $K-1$ directions. Schur's lemma is the clean abstract way to phrase this.

Remark 5.7 (Dimension-free scaling for classification). Theorem 5.5 (b) states that, at the NC fixed point, the effective rank of the *class-averaged* block $\overline{\mathbf{M}}_{\mathrm{CE}}$ depends only on K and not on the input dimension d or the parameter count p. This is the classification analog of the dimension-free regression rate of [3] (rank-one collapse onto a single target direction): both rest on a residual symmetry of the trained predictor at its fixed point that quotients out the ambient parameter dimension. In classification the residual symmetry is S_K acting on the simplex-ETF directions, and the target subspace $\mathcal{Z}_K = \mathbb{R}^K/\operatorname{span}(\mathbf{1})$ is irreducible of dimension $K-1$.

Remark 5.8 (Full-operator effective rank). With n training samples class-balanced and the S_K-symmetry of NC, the full operator $\mathbf{M}_{\mathrm{CE}}$ on the joint sample-class space has rank $n(K-1)$ on the zero-sum subspace and $r_{\mathrm{eff}}(\mathbf{M}_{\mathrm{CE}}) = n(K-1)$ at the limit point (once both the per-sample two-block structure and the S_K- averaging across classes are simultaneously taken into account). The dimension-free scaling we care about for generalization is the class-averaged quantity, which captures the noise leakage *per training example averaged over class identity*; this is the right object because the noise-leakage variance $\sigma^2 \operatorname{tr}(\mathbf{M}_{\mathrm{CE}})$ from Lemma 3.4 aggregates over both samples and classes and is invariant under the diagonal S_K-action.

5.4. **Comparison with prior work.** The result of Theorem 5.5 is positioned at the intersection of three prior literatures, and the contribution lies precisely in the operator-framework angle that none of them addresses.

- *NC geometry (Lu–Steinerberger [7], Mixon–Parshall–Pi [9]).* These works characterize the NC fixed point itself in the unconstrained-features model under cross-entropy: the simplex ETF, self-duality, and within-class collapse. They establish that NC is the unique global minimizer of the layer-peeled / unconstrained-features objective, and they quantify how rapidly NC is approached. We use their geometric characterization as input (Assumption 5.3) and do not re-derive it.

- *Central path under MSE (Han–Papyan–Donoho [6]).* This work studies the dynamical approach to NC under squared loss, identifying a central path in feature/classifier space. Our treatment is for the cross-entropy noise-propagation operator $\mathbf{M}_{\mathrm{CE}}$, not for the central-path dynamics; the loss-specific ingredient (Fisher matrix $\mathbf{H}$ instead of identity) is what forces the two-eigenvalue per-sample structure of Theorem 5.5 (a) and the S_K-equivariant Schur argument of Theorem 5.5 (b).
- *Class-imbalanced extension (Fang et al. [5]).* The unconstrained-features model under class imbalance shows that the simplex ETF deforms to a frequency-weighted variant. Our S_K-symmetry argument relies on class balance; the imbalanced case would replace S_K-equivariance by a smaller invariance group and is left to future work.

The contribution here is the r_{eff}**-of-**$\mathbf{M}_{\mathrm{CE}}$ **characterization within the noise-propagation operator framework of** [1]: identifying the right object (the class-averaged block on $\mathcal{Z}_K$, not the per-sample block), exhibiting the exact two-eigenvalue per-sample formula, and using S_K-irreducibility of $\mathcal{Z}_K$ to obtain $r_{\mathrm{eff}} = K - 1$ for the averaged operator. The earlier draft of this paper conflated the per-sample and averaged statements; the corrected statement delineates them, and is what is actually proven by simplex-ETF geometry plus Schur's lemma.

6. Discussion

Status of the NC hypothesis. The conclusion $r_{\mathrm{eff}}(\overline{\mathbf{M}}_{\mathrm{CE}}) = K-1$ (class-averaged block) is conditional on the NC fixed point being attained or approached at the end of training. NC is observed empirically across diverse networks [10], and is rigorously established in the unconstrained-features model of [7, 9], which idealizes the last-layer features as free variables. The result of this paper is therefore a statement of the form "NC + backbone non-degeneracy (Assumption 5.3(c)) $\Rightarrow$ rank-$(K-1)$ effective rank for the class-averaged block"; we do not prove that NC holds for finite networks trained end-to-end with cross-entropy. Per-sample blocks have $r_{\mathrm{eff}}^{(i)} \in [1, K-1]$ in general (see Theorem 5.5 (a) and Step 2 of the proof); the dimension-free $K-1$ scaling is a property of the class-averaged object only. Limitations in the strict one-hot limit are addressed by rescaling (Remark 5.2); the strict limit does *not* flatten the per-sample two-block spectrum, but the class-averaged quantity is unaffected.

Off-fixed-point behavior. A natural follow-up question is: how does $r_{\mathrm{eff}}(\mathbf{M}_{\mathrm{CE}})$ behave during the approach to the NC fixed point? We

expect a gap term whose decay is governed by the convergence rate of NC, which would yield a rate of $r_{\text{eff}}(\mathbf{M}_{\text{CE}})(t) - (K-1)$ in terms of the training time t. Quantifying this gap is open.

Class-imbalanced data. Theorem 5.5 assumes class-balanced gradient features. For imbalanced classes, the simplex ETF deforms to a class-frequency-weighted variant (see e.g. [5] for the unconstrained-features analysis), and the effective rank gains a class-frequency-dependent correction. We do not pursue this here.

Connection to the regression result of [3]. The companion paper [3] (this paper's regression counterpart) shows that, after feature-direction alignment, the effective rank of the regression operator $\mathbf{M} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma}\boldsymbol{\Sigma}^{-1}$ collapses to a single dimension along the target direction. In classification, the analogous collapse for the class-averaged block $\overline{\mathbf{M}}_{\text{CE}}$ is to $K-1$ dimensions along the simplex faces: each class identifies its corresponding simplex vertex, and the rank-one regression collapse becomes a rank-$(K-1)$ classification collapse. Both reflect the same abstract mechanism — a residual symmetry of the trained predictor at its fixed point reduces an a priori $\Theta(d)$ rank to a small representation-theoretic invariant — but the per-sample behavior differs: paper [3]'s per-sample rank-one collapse is exact for each sample, whereas in classification the per-sample blocks have only the smaller stabilizer S_{K-1} and the dimension-free $K-1$ statement holds only after S_K-averaging across class identities. We retain distinct notation $\mathbf{M}$ vs. $\mathbf{M}_{\text{CE}}$ throughout to disambiguate the two operators.

Acknowledgements. The exploratory analysis underlying Theorems 4.1 and 5.5, including the derivation of the Fisher-corrected operator from the cross-entropy Hessian, was conducted with the assistance of Claude (Anthropic). All mathematical statements and proofs have been verified by the author.

References

[1] L. Chang. Noise propagation operators and a two-parameter characterization of benign overfitting. *Companion paper*, 2026.

[2] L. Chang. A bias–variance decomposition for the noise propagation operator and the role of feature learning. *Companion paper*, 2026.

[3] L. Chang. Feature learning, effective dimension, and architecture-specific critical depth. *Companion paper*, 2026.

[4] S.-i. Amari. *Information Geometry and Its Applications*. Springer, 2016.

[5] C. Fang, H. He, Q. Long, and W. J. Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43):e2103091118, 2021.

[6] X. Y. Han, V. Papyan, and D. L. Donoho. Neural collapse under MSE loss: proximity to and dynamics on the central path. In *International Conference on Learning Representations (ICLR)*, 2022.
[7] J. Lu and S. Steinerberger. Neural collapse under cross-entropy loss. *Applied and Computational Harmonic Analysis*, 59:224–241, 2022.
[8] S. Mei, T. Misiakiewicz, and A. Montanari. Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, 2022.
[9] D. G. Mixon, H. Parshall, and J. Pi. Neural collapse with unconstrained features. *Sampling Theory, Signal Processing, and Data Analysis*, 20(2):11, 2022.
[10] V. Papyan, X. Y. Han, and D. L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

Independent Researcher
Email address: lightman.chang@gmail.com