Effective Dimension Ratios under Symmetry Augmentation

Hikaru Matsuoka matsuokah@sun.ac.jp
Department of Information Security, Faculty of Information Systems, University of Nagasaki, Nagasaki, Japan
RIKEN Center for Advanced Intelligence Project (AIP), Tokyo, Japan

Abstract

We study how finite-group symmetry augmentation changes the effective dimension d_{eff} of learned representations in deep networks. For a group action whose feature representation is regular-representation-like (Assumption R), the Reynolds decomposition splits the feature covariance into a G-invariant block and a variant block. Under an idealized isotropy assumption this yields a closed form for d_{eff} in terms of a single *symmetry index* $S \in [0, 1]$, the share of feature variance lying in the G-invariant subspace (Lemma 1). Trained representations, however, are typically anisotropic: their covariance concentrates variance in a low-rank, task-induced *spike* on top of an approximately isotropic *bulk*. We summarize this spike–bulk structure with a scalar *anisotropy index* $\rho \in [0, 1]$, the fraction of feature variance carried by the isotropic bulk, so that ρ=1 is fully isotropic and recovers Lemma 1. Our main result (Lemma 2) varies the *augmentation probability* $p \in [0, 1]$, the chance that a training sample is hit by a random group transform: p=0 is no augmentation and p=1 transforms every sample. Writing $S(p)$ and $\rho(p)$ for these indices measured at augmentation level p, and assuming only a mild conditional-invariance property of the spike, the *ratio* of effective dimensions between the two endpoints is fixed by the participation-ratio prefactor ψ: $d_{\text{eff}}(p{=}1)/d_{\text{eff}}(p{=}0) = \psi(S(1))/\psi(S(0))$ up to an $O(|\Delta\rho|)$ correction, where $\Delta\rho := \rho(1) - \rho(0)$. On CIFAR-10 with three groups (Z_2, D_2, D_4) and 20 width–data cells per group, the prediction holds with median error 9.2%, 10.0%, and 5.6% respectively, and a maximum error of 15.4%. This ratio is a representation-level signature of symmetry that holds independently of whether augmentation is helpful or harmful at the loss level, and constitutes a robust empirical regularity across the finite groups we tested.

1. Introduction

Two threads in deep learning theory have evolved largely independently. The first is the theory of *neural scaling laws*, which describes test loss as a sum $L(N, W) \approx L_\infty + AN^{-\alpha} + BW^{-\beta}$ in dataset size N and model width W [2, 8, 10, 17]. The second is the theory of *symmetry-aware learning*, which examines how invariance and equivariance affect sample complexity and generalization [3, 4, 6, 13, 15, 18]. A natural question at their intersection is: how does symmetry augmentation modulate the *capacity* of learned representations, and does this admit a clean, group-theoretic prediction?

This paper proposes a representation-level answer. We focus on the *participation ratio* of the feature covariance, $d_{\text{eff}} := (\operatorname{tr}\Sigma)^2/\operatorname{tr}(\Sigma^2)$, as a smooth proxy for the effective dimensionality of learned features. Following the Reynolds-decomposition framework recently popularized in Matsuoka [14] as a lens on equivariant learning dynamics, we write $\Sigma = \Sigma_{\text{inv}} \oplus \Sigma_{\text{var}}$ and, under an

© H. Matsuoka.

idealized *isotypic isotropy* assumption, obtain a closed form for d_{eff} in terms of a single *symmetry index* $S \in [0, 1]$ — the share of feature variance carried by the G-invariant subspace (Eq. (1); Section 3, Lemma 1). This first result isolates the group-theoretic mechanism, but it is too absolute for trained networks: empirical covariances often have a few dominant eigenvalues, or *spikes*, on top of an approximately isotropic bulk [1, 5, 7, 16]. As a result, Lemma 1 can miss the observed value of d_{eff} by an order of magnitude.

Two scalars drive the rest of the paper, and because both symbols are overloaded elsewhere we fix their meaning here. The *augmentation probability* $p \in [0, 1]$ is the chance that a randomly chosen group transform is applied to a training sample: $p{=}0$ is training without augmentation and $p{=}1$ transforms every sample. It is a single scalar knob, *not* the data distribution, which we write p_X. The *anisotropy index* $\rho \in [0, 1]$ summarizes the spike–bulk structure of the trained covariance as the fraction of feature variance carried by the isotropic bulk, so $\rho{=}1$ is fully isotropic (recovering Lemma 1) and small ρ is spike-dominated. We never use ρ for a group representation: group actions are written U_g throughout, and ρ always denotes this scalar anisotropy index.

Our main contribution (Section 4) is to show that, even under such anisotropy, the *ratio* of effective dimensions across two augmentation conditions remains predictable: under an anisotropic isotypic model and a mild, empirically verifiable conditional-invariance assumption on the task-induced spike, $d_{\text{eff}}(p{=}1)/d_{\text{eff}}(p{=}0) = \psi(S(1))/\psi(S(0))$ up to a controlled $O(|\Delta\rho|)$ correction (Lemma 2).

We test this prediction on CIFAR-10 [11] with three groups of varying alignment to the data: Z_2 (horizontal flip; nearly a true symmetry), D_2 (Klein four), and D_4 (full dihedral; not a true symmetry). Across 60 width-by-data cells and 5 augmentation strengths, the predicted ratio holds with median error of 9.2%, 10.0%, and 5.6% for Z_2, D_2, D_4 respectively, independent of whether augmentation helps or hurts loss.

Contributions. (i) We give a closed-form expression for d_{eff} under isotypic isotropy (Lemma 1). (ii) We prove a conditional ratio invariance result (Lemma 2) that survives the empirically observed spike-bulk anisotropy. (iii) We verify the prediction on three groups across 900 training runs on CIFAR-10. (iv) We discuss practical implications for choosing augmentation strength and comparing equivariant architectures with augmentation.

2. Setup

Let G be a finite group acting unitarily on the feature space $\mathbb{R}^D$ via operators U_g. The Reynolds projection $P_G := |G|^{-1} \sum_{g \in G} U_g$ is idempotent and self-adjoint; its image is the G-invariant subspace $V_{\text{inv}} \subseteq \mathbb{R}^D$, with orthogonal complement V_{var}.

Assumption (R) [regular-representation-like decomposition]. We assume that the action of G on $\mathbb{R}^D$ decomposes into $D/|G|$ copies of the regular representation, so that $\dim V_{\text{inv}} = D/|G|$ and $\dim V_{\text{var}} = D(|G| - 1)/|G|$.[1]

Let Σ denote the centered feature covariance. We define the *symmetry index* and the *effective dimension* as

$$S \;:=\; \frac{\operatorname{tr}(P_G \Sigma P_G)}{\operatorname{tr}(\Sigma)} \in [0, 1], \qquad d_{\text{eff}} \;:=\; \frac{(\operatorname{tr}\Sigma)^2}{\operatorname{tr}(\Sigma^2)}. \tag{1}$$

1. In general, $\dim V_{\text{inv}}$ equals the multiplicity of the trivial representation, which need not equal $D/|G|$. For our small CNN experiments, where G acts on each channel independently, (R) is the natural assumption; for general architectures (e.g. group convolutions [4]) one should verify this explicitly.

The participation ratio d_{eff} is bounded by $\min(B, D)$ for an empirical covariance from B samples and equals D when all eigenvalues are equal.

3. Lemma 1: A Closed Form, and Why It Is Brittle

Lemma 1 (Effective dimension under isotypic isotropy) *Assume (R), and additionally: (A) G-invariant data distribution and G-equivariant learning map (so that $U_g \Sigma U_g^\top = \Sigma$ for all $g \in G$); (B) within each isotypic component, eigenvalues concentrate around their bulk values (*isotypic isotropy*). Then*

$$d_{\text{eff}} \;=\; D \cdot \psi(S, |G|) \cdot (1 + o(1)), \qquad \psi(S, |G|) \;:=\; \frac{1}{|G|\big(S^2 + (1-S)^2/(|G|-1)\big)}. \tag{2}$$

Proof sketch By (A), Σ commutes with each U_g, hence with P_G. Combined with $P_G^2 = P_G$, this gives $P_G \Sigma (I - P_G) = 0$, so Σ is block-diagonal in the $V_{\text{inv}} \oplus V_{\text{var}}$ basis. By (R), the dimensions of the two blocks are $D/|G|$ and $D(|G|-1)/|G|$. By (B), the bulk eigenvalues are $\mu_{\text{inv}} = |G| S \operatorname{tr}(\Sigma)/D$ and $\mu_{\text{var}} = |G|(1-S)\operatorname{tr}(\Sigma)/[D(|G|-1)]$. Substituting into $d_{\text{eff}} = (\operatorname{tr}\Sigma)^2/[\operatorname{tr}(\Sigma_{\text{inv}}^2) + \operatorname{tr}(\Sigma_{\text{var}}^2)]$ yields (2). ■

Why the absolute prediction is brittle. Assumption (B) does not hold for trained representations. Empirically, Σ is strongly anisotropic: a small number of dominant eigenvalues (*spikes*) sit on top of an isotropic bulk [5, 7, 16]. Consequently, the absolute prediction (2) can over-estimate the observed d_{eff} by one to two orders of magnitude on CIFAR-10. We address this in Section 4.

4. Main Result: Lemma 1′ (Conditional Ratio Invariance)

To accommodate the empirically observed spike–bulk structure, we model the trained covariance as

$$\Sigma \;=\; \rho \cdot \Sigma_{\text{iso}} \;+\; (1-\rho) \cdot \Sigma_{\text{spike}}, \tag{3}$$

where Σ_{iso} satisfies isotypic isotropy (Assumption B), Σ_{spike} is a low-rank task-induced principal component, and both components are trace-normalized so that $\rho \in [0, 1]$ is the bulk's share of the feature variance.

Lemma 2 (Conditional ratio invariance) *Consider two augmentation conditions $p \in \{0, 1\}$ and let $\Delta\rho := \rho(1) - \rho(0)$. Assume (R), (A), the model (3), and the following* conditional-invariance *assumption:*

(CI) *the task-induced spike Σ_{spike} and its normalized overlap with the isotropic bulk are empirically stable across the two augmentation conditions, and the residual change is absorbed into $\Delta\rho$.*

Then

$$\frac{d_{\text{eff}}(p{=}1)}{d_{\text{eff}}(p{=}0)} \;=\; \frac{\psi(S(1), |G|)}{\psi(S(0), |G|)} \cdot \big(1 + O(|\Delta\rho|)\big). \tag{4}$$

Proof sketch By (CI), the spike component and its overlap with the isotropic bulk are common to the two endpoint conditions up to the scalar change $\Delta\rho$. The Reynolds decomposition therefore changes the isotropic component through $S(0)$ and $S(1)$, while the spike-dependent terms enter as

a shared sensitivity factor. For fixed ρ, direct substitution of (3) into the participation ratio leaves the leading S-dependence as the prefactor ratio $\psi(S(1), |G|)/\psi(S(0), |G|)$. Varying ρ between the two endpoints gives the first-order correction $O(|\Delta\rho|)$. ■

Remark on (CI). We emphasize that (CI) is *not* the strong claim that orbit averaging preserves class labels; for example, a 90° rotation in D_4 is clearly not class-preserving on CIFAR-10. Rather, (CI) is the much weaker requirement that the task-induced spike—which encodes class-relevant structure—is empirically stable across the two training conditions, with any drift absorbed into $\Delta\rho$. This is directly testable: we verify it in Section 5 for groups that are aligned with the data (Z_2) as well as misaligned (D_4).

Implications. Lemma 2 provides a representation-level invariant that is independent of whether augmentation is helpful or harmful at the loss level. The same prediction applies to true symmetries (where augmentation typically helps; cf. Elesedy and Zaidi 6, Mei et al. 15) and to broken symmetries (where augmentation can hurt; cf. Lawrence et al. 12). It is therefore a candidate observable signature of symmetry that can be measured directly from trained features, without reference to loss.

5. Experiments

Setup. On CIFAR-10 we sweep $G \in \{Z_2, D_2, D_4\}$, width $W \in \{16, 32, 64, 128, 256\}$, training set size $N \in \{500, 2000, 8000, 30000\}$, augmentation probability $p \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$, with 3 random seeds, totaling 900 training runs. We use a small 3-layer CNN trained for 30 epochs with cosine-annealed SGD ($\sim$ 80K parameters at $W = 64$). We measure d_{eff} as the participation ratio of the flattened third-conv-layer features and S via the empirical Reynolds projection.

5.1. $S(p{=}0)$ as a Natural Symmetry Indicator

We use $S(p{=}0)$ (the symmetry index of features trained *without* augmentation) as a representation-level analog of distributional symmetry. The values, averaged over W and N, are: $S(p{=}0) = 0.840$ for Z_2, 0.669 for D_2, and 0.573 for D_4, monotonically tracking the alignment between G and the actual CIFAR-10 distribution.[2] Z_2 being closest to 1 reflects that CIFAR-10 is approximately left–right symmetric, while D_4's low value reflects that 90° rotations are clearly not a symmetry of natural images.

5.2. Lemma 2 Verification

For each (W, N) cell we compare the observed ratio $d_{\text{eff}}(p{=}1)/d_{\text{eff}}(p{=}0)$ with the theoretical prediction $\psi(S(1))/\psi(S(0))$ from (4). Across 20 cells per group and three groups (**60 cells total**):

- Z_2 ($|G| = 2$): median relative error $\mathbf{9.22}\%$, max 12.39%.
- D_2 ($|G| = 4$): median $\mathbf{10.04}\%$, max 15.36%.
- D_4 ($|G| = 8$): median $\mathbf{5.64}\%$, max 13.54%.

2. We separately verified that the input-space metric $m(p_X)$ of Lawrence et al. [12] produces the same monotonic ordering: $m(Z_2) = 0.503$, $m(D_2) = 0.637$, $m(D_4) = 0.728$ (chance level 0.5).

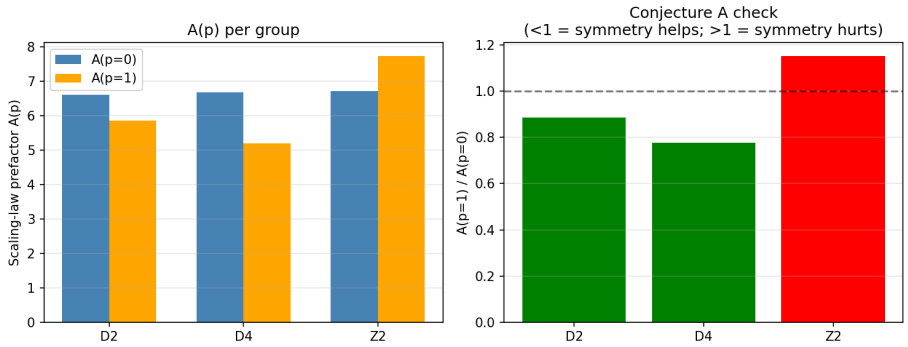

Figure 1: Scaling-law fits per group. Left: prefactors $A(p{=}0)$ and $A(p{=}1)$ from $L = AN^{-\alpha} + BW^{-\beta}$. Right: ratio $A(p{=}1)/A(p{=}0)$. The ratio alone is misleading: D_4 has the largest A reduction yet performs *worst*, because the exponent α also degrades (Section 5.3). Lemma 2 (representation-level) holds independently of these loss-level effects.

All three groups satisfy median error $\leq 10.1\%$ and maximum $\leq 16\%$, despite $|G|$ varying by a factor of four and the alignment to the data varying from near-perfect (Z_2) to clearly broken (D_4). Lemma 2 thus exhibits a robust empirical regularity across the finite groups we tested.

5.3. Loss-Level Effects (For Reference)

The loss-level fit confirms that single-coefficient analyses can be deceiving (Figure 1). From $p{=}0$ to $p{=}1$, Z_2 improves at large N (A: $6.71 \to 7.73$, α: $0.242 \to 0.266$), D_2 stays roughly flat (A: $6.61 \to 5.86$, α stable), and D_4 deteriorates (A: $6.67 \to 5.19$, α: $0.241 \to 0.186$). Crucially, Lemma 2 (a representation-level claim) is unaffected by these loss-level subtleties.

5.4. Practical Implications

Choosing augmentation strength. Two measurements ($S(p{=}0)$ and $S(p{=}1)$) suffice to predict representation capacity at any intermediate p via Lemma 2, eliminating expensive grid search. **Equivariant architectures vs. augmentation.** An equivariant architecture forces $S = 1$, fixing $d_{\text{eff}} = D/|G|$; augmentation gives a continuous ψ trade-off. Lemma 2 quantifies the capacity gap. **Diagnostics.** If S moves but d_{eff} deviates from Lemma 2, $\Delta\rho$ is large—a signal that the spike structure is shifting (e.g. overfitting or distribution drift). **Effective width.** Plug $W_{\text{eff}} = W \cdot \psi(S, |G|)$ into capacity calculations.

6. Discussion

Related work. The benefits of equivariance for sample complexity are by now well-studied at the loss level [3, 4, 6, 13, 15, 18]. Lawrence et al. [12] document that augmentation can in fact hurt when the data distribution is asymmetric. Independently, neural scaling-law theory [2, 8, 10, 17] aims to predict loss as a function of N and W. Our work is complementary: rather than asking when augmentation helps, we identify a representation-level invariant that is unaffected by that question. Our analytical setup—spike–bulk decomposition—draws on random matrix theory [1, 5, 7, 16] and is consistent with the lazy/rich regime distinction [9].

Limitations. (a) Tightening the $O(|\Delta\rho|)$ correction in Lemma 2 is open. (b) Assumption (R) may not hold for general equivariant architectures; the multiplicity of the trivial representation should be verified case-by-case. (c) We test discrete groups only; continuous groups (e.g. $SO(d)$, S_n) involve a regime where exponents themselves can change. (d) Our claims are an empirical regularity on CIFAR-10 with a small CNN; broader validation on point clouds, sequences, and larger models is left to future work.

References

[1] Alexander Atanasov, Blake Bordelon, Sabarish Sainathan, and Cengiz Pehlevan. The onset of variance-limited behavior for networks in the lazy and rich regimes. In *International Conference on Learning Representations*, 2024.

[2] Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024.

[3] Alberto Bietti, Luca Venturi, and Joan Bruna. On the sample complexity of learning under geometric stability. In *Advances in Neural Information Processing Systems*, 2021.

[4] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning*, pages 2990–2999, 2016.

[5] Stéphane D'Ascoli, Levent Sagun, and Giulio Biroli. Triple descent and the two kinds of overfitting: Where & why do they appear? In *Advances in Neural Information Processing Systems*, 2020.

[6] Bryn Elesedy and Sheheryar Zaidi. Provably strict generalisation benefit for invariant models. In *International Conference on Machine Learning*, pages 2959–2969, 2021.

[7] Boris Hanin and Mihai Nica. Products of many large random matrices and gradients in deep neural networks. *Communications in Mathematical Physics*, 376:287–322, 2020.

[8] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, 2022.

[9] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, 2018.

[10] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

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

[12] Hannah Lawrence, Elyssa Hofgard, Vasco Portilheiro, Yuxuan Chen, Tess Smidt, and Robin Walters. To augment or not to augment? diagnosing distributional symmetry breaking. In *International Conference on Learning Representations*, 2026.

[13] Clare Lyle, Mark van der Wilk, Marta Kwiatkowska, Yarin Gal, and Benjamin Bloem-Reddy. On the benefits of invariance in neural networks. In *arXiv preprint arXiv:2005.00178*, 2020.

[14] H. Matsuoka. Algebraic geometry of deep learning dynamics. Blog post, GRAM Blogposts, 2026. https://gram-blogposts.github.io/2026/blog/2026/algebraic-geometry-deep-learning-dynamics/.

[15] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Learning with invariances in random features and kernel models. In *Conference on Learning Theory*, pages 3351–3418, 2021.

[16] Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems*, 2017.

[17] Utkarsh Sharma and Jared Kaplan. Scaling laws from the data manifold dimension. *Journal of Machine Learning Research*, 23(9):1–34, 2022.

[18] Behrooz Tahmasebi and Stefanie Jegelka. The exact sample complexity gain from invariances for kernel regression. In *Advances in Neural Information Processing Systems*, 2023.