On the Geometry of Difference in Representation Space

Opposite ≠ Different: The Orthogonality Thesis

Taha Bouhsine

azetta.ai

22 February 2026

Abstract

For unit vectors $\mathbf{u}, \mathbf{v} \in \mathbb{S}^{d-1}$ with $\cos\theta(\mathbf{u}, \mathbf{v}) = -1$, the relation $\mathbf{v} = -\mathbf{u}$ holds, so antiparallel vectors are linearly dependent and span a single line. The maximally distinct configuration in the sense of linear independence is orthogonality, $\langle \mathbf{u}, \mathbf{v} \rangle = 0$, at which the span has dimension two and neither vector projects onto the other. We collect the algebraic, information-theoretic, and packing-theoretic statements of this fact: the squared cosine $\cos^2\theta$ measures shared variance and is maximised at $\theta \in \{0, \pi\}$ and minimised at $\theta = \pi/2$; the regular simplex of n unit vectors on $\mathbb{S}^{d-1}$ for $n \leq d+1$ has pairwise cosine $-1/(n-1)$, which tends to zero as $n \to \infty$; the uniform distribution on $\mathbb{S}^{d-1}$ concentrates near orthogonality with $\mathrm{Var}[\cos\theta] = 1/d$. We further show that cross-entropy, in the sense of the singularity $H(p, q) \to +\infty$ at $\mathrm{supp}(p) \cap \mathrm{supp}(q) = \varnothing$, registers disjoint support—the probabilistic analog of vector orthogonality—as an unbounded penalty, and that on the unit sphere the softmax classifier's equilibrium is the regular simplex. A closing discussion (Section 8) examines what these statements imply for contrastive losses used in practice; that material is offered as motivation and consequence rather than as further theorems.

Contents

1 Introduction

Cosine similarity is the standard measure of directional relationship between unit vectors on $\mathbb{S}^{d-1}$. It ranges from $+1$ (parallel) through 0 (orthogonal) to -1 (antiparallel). A widely held convention treats the extremes of this range as a single "difference" axis, with -1 taken as the maximally distinct configuration.

Algebraically, this convention is incorrect. Two unit vectors $\mathbf{u}, \mathbf{v} \in \mathbb{S}^{d-1}$ satisfying $\cos\theta(\mathbf{u}, \mathbf{v}) = -1$ also satisfy $\mathbf{v} = -\mathbf{u}$; they are linearly dependent and span a single line. Genuine independence requires the span to have dimension at least two, and the cleanest realisation of that condition is orthogonality, $\langle \mathbf{u}, \mathbf{v} \rangle = 0$. The aim of this paper is to collect the algebraic, information-theoretic, packing-theoretic, and probabilistic statements of this fact, and to discuss in a final section their implications for loss functions used in modern representation learning.

The paper is organised as follows. Section 2 establishes the algebra of linear dependence and independence on the sphere. Section 7 treats the information-theoretic counterpart in terms of $\cos^2\theta$. Section **??** records the concentration of $\cos\theta$ for uniform unit vectors on $\mathbb{S}^{d-1}$. Section 6 states the simplex packing result. Section 5 treats cross-entropy and the singularity at disjoint support. Section 8 is a closing discussion of contrastive losses in machine learning; that material is offered as motivation and consequence rather than as further theorems.

2 The Geometry of Difference

We begin by formalising what "different" means in a vector space; the answer is not the answer most contrastive losses act on.

Definition 1 (Linear dependence)**.** *Two nonzero vectors* $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ *are* linearly dependent *if there exists* $\alpha \in \mathbb{R} \setminus \{0\}$ *such that* $\mathbf{v} = \alpha\mathbf{u}$*. The span satisfies* $\operatorname{span}(\mathbf{u}, \mathbf{v}) = \operatorname{span}(\mathbf{u})$*, so* $\dim \operatorname{span}(\mathbf{u}, \mathbf{v}) = 1$.

Theorem 2 (Antiparallel vectors are linearly dependent)**.** *Let* $\mathbf{u}, \mathbf{v} \in \mathbb{S}^{d-1}$ *with* $\cos\theta(\mathbf{u}, \mathbf{v}) = -1$*. Then* $\mathbf{v} = -\mathbf{u}$ *and:*

(a) $\operatorname{span}(\mathbf{u}, \mathbf{v}) = \operatorname{span}(\mathbf{u})$*, so* $\dim = 1$.

(b) The projection of $\mathbf{v}$ *onto* $\mathbf{u}$ *is* $\operatorname{proj}_{\mathbf{u}}(\mathbf{v}) = -\mathbf{u}$*, with* $|\operatorname{proj}_{\mathbf{u}}(\mathbf{v})| = \|\mathbf{u}\| = 1$.

(c) Knowing $\mathbf{u}$ *determines* $\mathbf{v}$ *completely:* $H(\mathbf{v} \mid \mathbf{u}) = 0$.

Proof. Since $\mathbf{u}, \mathbf{v} \in \mathbb{S}^{d-1}$, we have $\|\mathbf{u}\| = \|\mathbf{v}\| = 1$. The condition $\cos\theta = -1$ gives $\langle \mathbf{u}, \mathbf{v} \rangle = -1$. Now consider $\|\mathbf{u} + \mathbf{v}\|^2 = \|\mathbf{u}\|^2 + 2\langle \mathbf{u}, \mathbf{v} \rangle + \|\mathbf{v}\|^2 = 1 - 2 + 1 = 0$. Since $\|\mathbf{u} + \mathbf{v}\|^2 = 0$ implies $\mathbf{u} + \mathbf{v} = \mathbf{0}$, we have $\mathbf{v} = -\mathbf{u}$. Parts (a)–(c) follow immediately: $\mathbf{v}$ is a scalar multiple of $\mathbf{u}$ with $\alpha = -1$. □

Theorem 3 (Orthogonal vectors are maximally independent)**.** *Let* $\mathbf{u}, \mathbf{v} \in \mathbb{S}^{d-1}$ *with* $\cos\theta(\mathbf{u}, \mathbf{v}) = 0$*. Then:*

(a) $\operatorname{span}(\mathbf{u}, \mathbf{v})$ *has* $\dim = 2$.

(b) $\operatorname{proj}_{\mathbf{u}}(\mathbf{v}) = \mathbf{0}$*:* $\mathbf{v}$ *has zero component along* $\mathbf{u}$.

(c) $\mathbf{v}$ *cannot be reconstructed from* $\mathbf{u}$ *and vice versa.*

Proof. $\langle \mathbf{u}, \mathbf{v} \rangle = 0$ with $\mathbf{u}, \mathbf{v} \neq \mathbf{0}$ implies linear independence (standard result). Hence $\dim \operatorname{span}(\mathbf{u}, \mathbf{v}) = 2$. The projection $\operatorname{proj}_{\mathbf{u}}(\mathbf{v}) = \frac{\langle \mathbf{v}, \mathbf{u} \rangle}{\|\mathbf{u}\|^2}\mathbf{u} = 0 \cdot \mathbf{u} = \mathbf{0}$. □

Remark. The cosine scale has three landmarks, not two:

$\cos\theta$	Geometric status	Information content
$+1$	Parallel (identical direction)	Maximal redundancy
-1	Antiparallel (opposite direction)	Maximal redundancy (sign-flipped)
0	Orthogonal (perpendicular)	Zero shared information

The "difference" axis runs from ± 1 (dependent) to 0 (independent), not from $+1$ to -1.

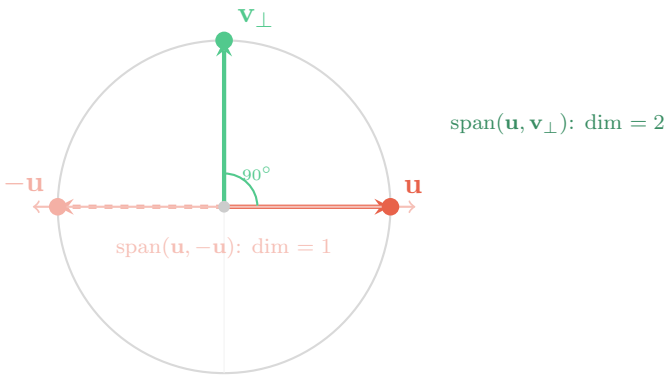

Figure 1: Antiparallel vectors $\mathbf{u}$ and $-\mathbf{u}$ span a single line (dim = 1): they are linearly dependent. The orthogonal vector $\mathbf{v}_\perp$ spans an independent direction (dim = 2): it carries genuinely new information.

3 Softmax Contrastive Losses and Their Targets

CLIP [1] is the foundational model for modern vision–language systems, and its training objective is InfoNCE [4]: a softmax contrastive loss over a batch of N image–text pairs.

Definition 4 (InfoNCE / CLIP loss)**.** *Given a batch of N image-text pairs with embeddings $\{(\mathbf{z}_i^{img}, \mathbf{z}_i^{txt})\}_{i=1}^N$, the InfoNCE loss for the image side is:*

$$\mathcal{L}_{CLIP} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\big(\mathrm{sim}(\mathbf{z}_i^{img}, \mathbf{z}_i^{txt})/\tau\big)}{\sum_{k=1}^{N} \exp\big(\mathrm{sim}(\mathbf{z}_i^{img}, \mathbf{z}_k^{txt})/\tau\big)}, \tag{1}$$

where $\mathrm{sim}(\mathbf{a}, \mathbf{b}) = \mathbf{a}^\top \mathbf{b}/(\|\mathbf{a}\|\|\mathbf{b}\|)$ is cosine similarity and $\tau > 0$ is a temperature.

Proposition 5 (InfoNCE implicitly targets antiparallel alignment)**.** *For a fixed anchor $\mathbf{z}_i^{img}$ and temperature $\tau > 0$, the InfoNCE gradient with respect to a negative embedding $\mathbf{z}_k^{txt}$ $(k \neq i)$ pushes:*

$$\frac{\partial \mathcal{L}_{CLIP}}{\partial \mathrm{sim}(\mathbf{z}_i^{img}, \mathbf{z}_k^{txt})} \propto \frac{\exp\big(\mathrm{sim}(\mathbf{z}_i^{img}, \mathbf{z}_k^{txt})/\tau\big)}{\sum_{j=1}^{N} \exp\big(\mathrm{sim}(\mathbf{z}_i^{img}, \mathbf{z}_j^{txt})/\tau\big)} > 0. \tag{2}$$

Since the gradient is strictly positive, the loss monotonically decreases as $\mathrm{sim}(\mathbf{z}_i^{img}, \mathbf{z}_k^{txt})$ decreases. The global minimum of each negative similarity is -1 (antiparallel), so InfoNCE drives negatives toward linear dependence with the anchor.

Remark. CLIP's InfoNCE loss wants "a photo of a dog" to be *different* from "a photo of a cat." But by driving their embeddings to $\cos = -1$, it makes the cat embedding a simple negation of the dog embedding. The network doesn't learn that dogs and cats are different concepts; it learns they are opposite poles of the same axis. One dimension consumed, zero new information.

3.1 Geometric Capacity Bound

Proposition 6 (Antiparallel capacity limit)**.** *In $\mathbb{R}^d$, the maximum number of mutually antiparallel unit vector pairs is d. Each pair $(\mathbf{e}_i, -\mathbf{e}_i)$ consumes one basis direction.*

Proof. Antiparallel pairs lie along one-dimensional subspaces. Distinct one-dimensional subspaces of $\mathbb{R}^d$ correspond to points in the projective space $\mathbb{R}\mathbf{P}^{d-1}$. For n pairs to be mutually antiparallel, they must lie along n distinct coordinate axes, and at most d such orthogonal axes exist. □

Corollary 7. *CLIP's 512-dimensional embedding space can represent at most 512 binary oppositions. But ImageNet has 1,000 classes, and the real world has millions of concepts. Any loss that targets* $\cos = -1$ *for all negative pairs faces a fundamental geometric impossibility.*

3.2 The Computational Catastrophe

The geometric error compounds into a computational one.

Proposition 8 (InfoNCE batch-size dependence)**.** *The gradient signal for negative k in InfoNCE is weighted by its softmax probability:*

$$w_k = \frac{\exp(\text{sim}_k/\tau)}{\sum_j \exp(\text{sim}_j/\tau)}. \tag{3}$$

In high dimensions, random unit vectors have $\text{sim} \approx 0$ *with variance* $1/d$ *(Theorem 11). Consequently, most negatives contribute approximately equal, near-zero gradient, and the loss requires $N \gg 1$ to accumulate sufficient signal.*

Remark. CLIP was trained with batch sizes of $N = 32{,}768$ image-text pairs. Each batch requires computing an $N \times N$ similarity matrix ($\sim$ 4 GB at float32). This extreme cost is a direct consequence of the wrong geometric target:

- The loss pushes negatives toward $\cos = -1$, but random embeddings are already near $\cos \approx 0$.
- Most of the batch contributes near-zero gradient—the loss has to see thousands of negatives to find the rare pairs near antiparallel alignment.
- The scaling law is brutal: doubling accuracy requires roughly quadrupling batch size.

The field spent years engineering distributed training, gradient caching, and memory-efficient attention—all to compensate for a geometric mistake in the loss function.

4 Sigmoid Pairwise Contrastive Losses

SigLIP [2] replaces InfoNCE's softmax with a pairwise sigmoid. The change is small in code and structural in geometry.

Definition 9 (SigLIP loss)**.** *Given image-text pairs with labels $y_{ij} = +1$ (matched) or -1 (mismatched):*

$$\mathcal{L}_{SigLIP} = \sum_{i,j} \log\big(1 + \exp\big(-y_{ij}\big(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau - b\big)\big)\big), \tag{4}$$

where b is a learnable bias parameter.

Proposition 10 (Sigmoid equilibrium at orthogonality)**.** *For a mismatched pair ($y_{ij} = -1$), the SigLIP gradient with respect to similarity is:*

$$\frac{\partial \mathcal{L}}{\partial \mathrm{sim}_{ij}} = \frac{1}{\tau} \cdot \sigma\big(\mathrm{sim}_{ij}/\tau - b\big), \tag{5}$$

where σ is the logistic sigmoid. This gradient vanishes as $\mathrm{sim}_{ij} \to -\infty$, but in practice the sigmoid saturates once $\mathrm{sim}_{ij} < b$ by a few multiples of τ. Unlike InfoNCE, SigLIP does not *push similarity toward -1—it only requires $\mathrm{sim}_{ij} < b$.*

Theorem 11 (Concentration of measure on $\mathbb{S}^{d-1}$)**.** *For two independent uniform random vectors $\mathbf{x}, \mathbf{y} \sim \mathrm{Uniform}(\mathbb{S}^{d-1})$:*

$$\mathbb{E}[\cos\theta(\mathbf{x}, \mathbf{y})] = 0, \qquad \mathrm{Var}[\cos\theta(\mathbf{x}, \mathbf{y})] = \frac{1}{d}. \tag{6}$$

For $d = 512$ (CLIP's embedding dimension), $\mathrm{Std}[\cos\theta] \approx 0.044$. Almost all pairs of random unit vectors are nearly orthogonal.

Proof. Let $\mathbf{y} = (Y_1, \ldots, Y_d)^\top$ be uniform on $\mathbb{S}^{d-1}$. Without loss of generality, fix $\mathbf{x} = \mathbf{e}_1$. Then $\cos\theta = Y_1$. The marginal distribution of Y_1 for the uniform distribution on $\mathbb{S}^{d-1}$ has $\mathbb{E}[Y_1] = 0$ (by symmetry) and $\mathrm{Var}[Y_1] = \mathbb{E}[Y_1^2] = \frac{1}{d}$ (since $\sum_i Y_i^2 = 1$ and all components have equal variance). □

Remark. SigLIP's sigmoid loss aligns with the natural geometry of high-dimensional spheres:

- Random embeddings are already near $\cos \approx 0$ (orthogonal).
- The sigmoid only needs mismatched similarities below the bias b—it doesn't fight the geometry by pushing toward -1.
- The pairwise structure eliminates the $N \times N$ softmax competition, enabling smaller batch sizes.

Result: better zero-shot accuracy with less compute [2]. Not because of better engineering, but because the geometry is right.

5 Cross-Entropy at Disjoint Support

Cross-entropy is the standard classification loss. The geometry it has always been driving toward is orthogonality, not opposition—a fact obscured by the way the loss is usually described, but visible immediately in its singularity structure.

Definition 12 (Cross-entropy)**.** *For discrete distributions p and q over alphabet $\mathcal{X}$:*

$$H(p, q) = -\sum_{x \in \mathcal{X}} p(x) \log q(x). \tag{7}$$

Theorem 13 (Singularity at disjoint support)**.** *If $\mathrm{supp}(p) \cap \mathrm{supp}(q) = \varnothing$, then $H(p, q) = +\infty$.*

Proof. If the supports are disjoint, then for every x with $p(x) > 0$ we have $q(x) = 0$, so $\log q(x) = -\infty$. Since $p(x) > 0$ for at least one such x:

$$H(p, q) = -\sum_x p(x) \log q(x) \geq -p(x_0) \log q(x_0) = -p(x_0) \cdot (-\infty) = +\infty.$$

□

Remark. Disjoint support is the probabilistic analog of orthogonality. Two distributions whose supports don't intersect are like two vectors with zero dot product: they share no information, they occupy independent regions of the sample space. The singularity $H(p, q) \to +\infty$ is cross-entropy's way of encoding complete orthogonality.

5.1 Why Classifiers Learn Orthogonal Representations

Proposition 14 (Softmax drives orthogonal weight vectors)**.** *In a softmax classifier with logits $z_k = \mathbf{w}_k^\top \mathbf{h}$ and cross-entropy loss, the gradient with respect to the wrong-class logit z_k ($k \neq y$) is:*

$$\frac{\partial \mathcal{L}}{\partial z_k} = \hat{p}(k \mid \mathbf{h}) = \frac{\exp(z_k)}{\sum_j \exp(z_j)} > 0. \tag{8}$$

This gradient pushes $z_k = \mathbf{w}_k^\top \mathbf{h}$ toward $-\infty$. On the unit sphere ($\|\mathbf{w}_k\| = \|\mathbf{h}\| = 1$), the loss has competing pressures: it wants $\mathbf{w}_y$ parallel to $\mathbf{h}$ (to maximise z_y) and each $\mathbf{w}_k$ with $k \neq y$ antiparallel to $\mathbf{h}$ (to minimise z_k). With n classes sharing the same $\mathbf{h}$, the antiparallel target cannot be reached simultaneously by all $\mathbf{w}_k$, so the equilibrium is determined by the configuration that minimises the wrong-class logits subject to mutual diversity of the $\mathbf{w}_k$. That configuration is the regular simplex (Theorem 15): $\langle \mathbf{w}_k, \mathbf{h} \rangle = -1/(n-1)$ for all $k \neq y$, which tends to 0 as n grows. The equilibrium is therefore approximately *orthogonal for any non-trivial multi-class setting and* exactly *orthogonal in the limit $n \to \infty$; the binary case $n = 2$ is the only regime in which the equilibrium is genuinely antiparallel.*

Example. This explains why plain cross-entropy classifiers produce well-separated representations even without any contrastive mechanism:

- The loss's singularity structure is geometrically informed by orthogonality.
- SigLIP's sigmoid loss shares this singularity structure (binary cross-entropy per pair), which is why it naturally drives mismatched pairs to orthogonality.
- CLIP's softmax InfoNCE does *not* share this structure—its softmax competition pushes beyond orthogonality toward opposition.

6 Optimal Packing on the Sphere

What is the optimal arrangement of n class representations on $\mathbb{S}^{d-1}$ when $n \leq d+1$?

Theorem 15 (Regular simplex packing)**.** *The arrangement of n unit vectors on $\mathbb{S}^{d-1}$ that maximizes the minimum pairwise angle (the Tammes problem for $n \leq d+1$) is the regular simplex, where all pairwise cosine similarities are equal:*

$$\cos\theta_{ij} = -\frac{1}{n-1} \quad \textit{for all } i \neq j. \tag{9}$$

Proof sketch. The n vertices of a regular simplex centered at the origin in $\mathbb{R}^n$ satisfy $\sum_{i=1}^n \mathbf{v}_i = \mathbf{0}$ and $\|\mathbf{v}_i\| = c$ for all i. Computing:

$$\mathbf{0} = \Big\|\sum_i \mathbf{v}_i\Big\|^2 = \sum_i \|\mathbf{v}_i\|^2 + 2\sum_{i<j}\langle \mathbf{v}_i, \mathbf{v}_j\rangle = nc^2 + 2\binom{n}{2}\langle \mathbf{v}_i, \mathbf{v}_j\rangle,$$

where the last equality uses equi-angularity. Solving: $\cos\theta = \langle \mathbf{v}_i, \mathbf{v}_j\rangle / c^2 = -\frac{1}{n-1}$. For $n \leq d+1$, this configuration embeds in $\mathbb{S}^{d-1}$. □

Corollary 16 (Simplex approaches orthogonality)**.** *As the number of classes grows:*

$$\lim_{n\to\infty} \cos\theta_{simplex} = \lim_{n\to\infty}\left(-\frac{1}{n-1}\right) = 0.$$

The optimal packing converges to orthogonality, not opposition.

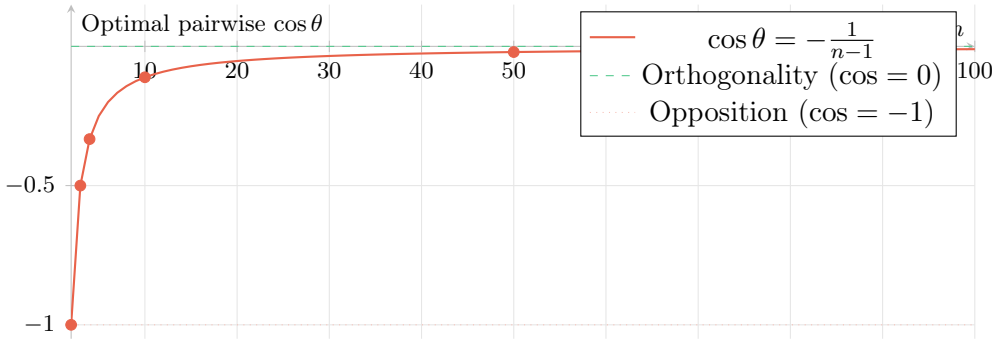

Figure 2: The optimal inter-class cosine similarity for n classes on the simplex. For $n = 2$ (binary), the optimal angle is 180° ($\cos = -1$)—opposition works. But for $n \geq 3$, the optimal configuration rapidly approaches orthogonality. By $n = 10$, $\cos\theta \approx -0.11$; by $n = 50$, $\cos\theta \approx -0.02$. Contrastive losses that target $\cos = -1$ for all negatives are fighting the geometry.

Remark 1. The simplex result explains a striking empirical observation: for binary classification ($n = 2$), the optimal arrangement *is* antiparallel ($\cos = -1$), and standard contrastive losses work well. The confusion arises from over-generalizing the binary case to multi-class settings, where opposition becomes increasingly wrong.

7 Directional Mutual Information

We can also frame the distinction in information-theoretic terms.

Definition 17 (Directional mutual information)**.** *For unit vectors* $\mathbf{u}, \mathbf{v} \in \mathbb{S}^{d-1}$*, define the directional mutual information as:*

$$I_{dir}(\mathbf{u}; \mathbf{v}) = 1 - \frac{\|\mathbf{v} - \text{proj}_{\mathbf{u}}(\mathbf{v})\|^2}{\|\mathbf{v}\|^2} = \cos^2\theta(\mathbf{u}, \mathbf{v}). \tag{10}$$

This measures the fraction of $\mathbf{v}$*'s variance explained by* $\mathbf{u}$*.*

Proposition 18 (Antiparallel maximizes shared information)**.**

$$\cos\theta = -1 \implies I_{dir} = \cos^2\theta = 1 \qquad \textit{(maximal redundancy)}, \tag{11}$$

$$\cos\theta = 0 \implies I_{dir} = \cos^2\theta = 0 \qquad \textit{(zero shared info)}, \tag{12}$$

$$\cos\theta = +1 \implies I_{dir} = \cos^2\theta = 1 \qquad \textit{(maximal redundancy)}. \tag{13}$$

Both parallel and *antiparallel vectors share all information. Only orthogonal vectors share none.*

Remark. The true "difference" metric is $\cos^2\theta$, not $\cos\theta$. It is minimized at orthogonality, not at opposition:

$$\boxed{\text{max difference} \iff \cos^2\theta = 0 \iff \cos\theta = 0 \iff \mathbf{u} \perp \mathbf{v}} \tag{14}$$

8 Discussion

Remark. The methods that target orthogonality (SigLIP, cross-entropy) share two properties:

1. They align with the natural concentration of measure on $\mathbb{S}^{d-1}$.

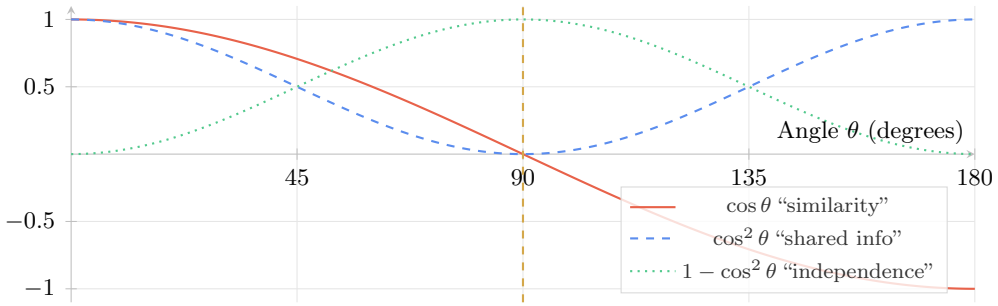

Figure 3: The similarity $\cos\theta$ (red) reaches its minimum at $\theta = 180°$, but the shared information $\cos^2\theta$ (blue dashed) is *maximized* at both 0° and 180°. True independence $1 - \cos^2\theta$ (green dotted) peaks at $\theta = 90°$: **orthogonality**. The gold line marks the point of maximum difference.

Table 1: Comparison of loss functions and their geometric targets.

Method	**Loss type**	**Negative target**	**Batch req.**	**Geometry**
SimCLR [3]	Softmax InfoNCE	$\cos \to -1$	$N \geq 4096$	Opposition
CLIP [1]	Softmax InfoNCE	$\cos \to -1$	$N = 32768$	Opposition
SigLIP [2]	Sigmoid pairwise	$\cos < b$	$N \geq 1024$	Orthogonality
Cross-entropy	Softmax CE	$z_k \to 0$	N/A	Orthogonality
SupCon [5]	Supervised InfoNCE	$\cos \to -1$	$N \geq 2048$	Opposition

2. They achieve comparable or better accuracy with less compute.

The methods that target opposition (SimCLR, CLIP, SupCon) require enormous batch sizes to overcome the geometric tension between their objective and the high-dimensional geometry of the sphere.

The point of the present paper is that the algebraic distinction between dependence and independence (Section 2), the information-theoretic counterpart (Section 7), the packing geometry of the regular simplex on $\mathbb{S}^{d-1}$ (Section 6), and the disjoint-support singularity of cross-entropy (Section 5) all identify the same configuration: orthogonality is the maximally distinct configuration in every formulation under consideration, and opposition is its redundant counterpart. This identification is mathematically straightforward; the consequences for losses used in practice are interpretive, and we have tried to mark the boundary throughout.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of ICML*, 2021. https://arxiv.org/abs/2103.00020

[2] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. Sigmoid loss for language image pre-training. In *Proceedings of ICCV*, 2023. https://arxiv.org/abs/2303.15343

[3] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of ICML*, 2020. https://arxiv.org/abs/2002.05709

[4] A. van den Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. https://arxiv.org/abs/1807.03748

[5] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. Supervised contrastive learning. In *Advances in NeurIPS*, 2020. https://arxiv.org/abs/2004.11362

[6] H. Steck, C. Ekanadham, and N. Kallus. Is cosine-similarity of learned representations all you need? *arXiv preprint arXiv:2403.05440*, 2024. https://arxiv.org/abs/2403.05440

[7] T. Wang and P. Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of ICML*, 2020. https://arxiv.org/abs/2005.10242