On the Singularity Structure of Information-Theoretic Losses

Beyond Binary Finiteness: The Geometry of Cross-Entropy Divergence

Taha Bouhsine

azetta.ai

23 February 2026

Abstract

For discrete distributions p, q on a finite alphabet $\mathcal{X}$, the cross-entropy $H(p,q) = -\sum_x p(x)\log q(x)$ is infinite precisely when the violating set $V_{p,q} = \{x : p(x) > 0,\ q(x) = 0\}$ is non-empty, a strictly weaker condition than disjointness of supports. We classify this singularity by the combinatorial size $|V_{p,q}|$, the mass severity $S(p,q) = \sum_{x \in V_{p,q}} p(x)$, and the asymptotic growth rate. Under uniform smoothing $q^{(\varepsilon)} = (1-\varepsilon)q + \varepsilon u$, where u is uniform on $|\mathcal{X}| = n$,

$$H(p, q^{(\varepsilon)}) = S(p,q)\,(-\log \varepsilon) + S(p,q)\log n + C_{p,q} + O(\varepsilon),$$

so the rate of divergence equals the violating mass. The Kullback–Leibler divergence inherits the singularity directly and is undefined at disjoint support—the probabilistic analog of vector orthogonality on $\mathbb{S}^{d-1}$, where the Euclidean distance is finite but the information-theoretic distance is not. We further establish a per-coordinate asymmetry: contributions of the form $-p(x)\log q(x)$ are unbounded at $q(x) = 0,\ p(x) > 0$ and vanish at $p(x) = 0,\ q(x) > 0$. Section 6 surveys divergences that are finite at the boundary. A closing discussion (Section 7) examines how these properties bear on the design and behaviour of contemporary losses, treating that material as motivation and consequence rather than as further theorems.

Contents

1 Introduction

The cross-entropy $H(p,q) = -\sum_x p(x)\log q(x)$ between discrete distributions is the standard divergence used to train statistical models, with the well-known property that minimisation in q is equivalent to maximum likelihood and admits a unique minimiser $q = p$. On the interior of the probability simplex it is a smooth and well-behaved functional. On the boundary it is not: if q assigns zero mass to a point where p has positive mass, $H(p,q)$ is infinite.

Standard treatments record this only as a dichotomy: $H(p,q)$ is either finite or infinite, well-defined or undefined. The present paper studies the structure of the singularity beyond this binary. We show that the condition for divergence is strictly weaker than disjoint support; that the rate of divergence under smoothing is a real-valued function of the violating mass; that the per-coordinate contributions are infinitely asymmetric; and that the Kullback–Leibler divergence inherits these properties and fails to extend continuously to the configuration of disjoint support, which is the probabilistic analog of vector orthogonality.

The paper is organised as follows. Section 2 fixes notation. Section 3 states the divergence condition and the growth-rate identity under uniform smoothing. Section 4 establishes the per-coordinate asymmetry. Section 5 treats the Kullback–Leibler divergence and its singularity at disjoint support, including the parallel with vector orthogonality. Section 6 surveys divergence measures that remain finite at the boundary. A closing discussion (Section 7) examines how the structure bears on losses used in contrastive and multimodal learning; that material is offered as motivation and consequence rather than as further theorems.

2 Preliminaries

Definition 1 (Cross-entropy)**.** *For discrete distributions p and q over a finite alphabet $\mathcal{X}$:*

$$H(p,q) = -\sum_{x\in\mathcal{X}} p(x)\log q(x), \tag{1}$$

with the convention $0\log 0 = 0$ and $p(x)\log 0 = +\infty$ when $p(x) > 0$.

Definition 2 (KL divergence)**.** *The Kullback–Leibler divergence from q to p is:*

$$D_{\mathrm{KL}}(p\|q) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} = H(p,q) - H(p), \tag{2}$$

where $H(p) = -\sum_x p(x)\log p(x)$ is the Shannon entropy.

Definition 3 (Support and violating set)**.** *The* support *of a distribution p is* $\mathrm{supp}(p) = \{x : p(x) > 0\}$*. The* violating set *of p with respect to q is:*

$$V_{p,q} = \{x \in \mathcal{X} : p(x) > 0 \text{ and } q(x) = 0\} = \mathrm{supp}(p) \setminus \mathrm{supp}(q). \tag{3}$$

The violating mass *is $S(p,q) = \sum_{x\in V_{p,q}} p(x)$.*

3 The Taxonomy of Singularities

Theorem 4 (Necessary and sufficient condition for divergence)**.** *$H(p,q) = +\infty$ if and only if $V_{p,q} \neq \varnothing$, i.e., $\operatorname{supp}(p) \not\subseteq \operatorname{supp}(q)$.*

Proof. ($\Rightarrow$) If $V_{p,q} \neq \varnothing$, there exists x_0 with $p(x_0) > 0$ and $q(x_0) = 0$. Then $-p(x_0)\log q(x_0) = +\infty$, so $H(p,q) = +\infty$.

($\Leftarrow$) If $V_{p,q} = \varnothing$, then $\operatorname{supp}(p) \subseteq \operatorname{supp}(q)$, so $q(x) > 0$ whenever $p(x) > 0$. Each term $-p(x)\log q(x)$ is finite, and the sum over finite $\mathcal{X}$ is finite. □

Remark. The condition $V_{p,q} \neq \varnothing$ is strictly weaker than $\operatorname{supp}(p) \cap \operatorname{supp}(q) = \varnothing$. Two distributions can share 99% of their mass and still give $H(p,q) = +\infty$ provided a single coordinate exists where p has positive mass and q has none. The textbook gloss "cross-entropy diverges at disjoint support" is a sufficient condition presented as if it were necessary and sufficient—and the distinction matters in practice, because models routinely produce q with broadly overlapping but not subsuming support.

3.1 Example: Overlapping but Infinite

Example. Consider distributions over $\{A, B, C\}$:

$$p = (0.5,\ 0.3,\ 0.2), \qquad q = (0.7,\ 0.3,\ 0).$$

The supports overlap on $\{A, B\}$, sharing 80% of p's mass. But $V_{p,q} = \{C\}$ with $p(C) = 0.2 > 0$ and $q(C) = 0$:

$$H(p,q) = -0.5\log 0.7 - 0.3\log 0.3 - 0.2\log 0 = +\infty.$$

The supports are far from disjoint. The infinity comes from a single uncovered coordinate.

3.2 The Three Tiers

We propose a tiered classification of the singularity, going beyond the binary finite/infinite distinction.

Definition 5 (Singularity tiers)**.** *For distributions p, q with $V_{p,q} \neq \varnothing$, define:*

(i) ***Combinatorial tier****: $|V_{p,q}|$ — the number of violating coordinates.*

(ii) ***Mass tier****: $S(p,q) = \sum_{x \in V_{p,q}} p(x)$ — the total target mass on violating coordinates.*

(iii) ***Rate tier****: the asymptotic growth rate of $H(p, q^{(\varepsilon)})$ as $\varepsilon \to 0^+$, where $q^{(\varepsilon)}$ is a smoothed approximation.*

Theorem 6 (Growth rate under uniform smoothing)**.** *Let $q^{(\varepsilon)} = (1-\varepsilon)q + \varepsilon u$, where u is the uniform distribution over $\mathcal{X}$ with $|\mathcal{X}| = n$. Then as $\varepsilon \to 0^+$:*

$$\boxed{H(p, q^{(\varepsilon)}) = S(p,q)(-\log \varepsilon) + S(p,q)\log n + C_{p,q} + O(\varepsilon)} \tag{4}$$

where $C_{p,q} = -\sum_{x \notin V_{p,q}} p(x)\log q(x)$ is the finite part of the cross-entropy.

Proof. For $x \in V_{p,q}$: $q^{(\varepsilon)}(x) = \varepsilon/n$, so $-p(x)\log q^{(\varepsilon)}(x) = p(x)(-\log\varepsilon + \log n)$. Summing over $V_{p,q}$: $S(p,q)(-\log\varepsilon + \log n)$.

For $x \notin V_{p,q}$: $q^{(\varepsilon)}(x) = (1-\varepsilon)q(x) + \varepsilon/n \to q(x)$ as $\varepsilon \to 0$, and $-p(x)\log q^{(\varepsilon)}(x) \to -p(x)\log q(x) = C_{p,q}^{(x)}$ with $O(\varepsilon)$ error. □

Remark. The divergence rate is exactly $S(p, q)$, the violating mass. Not all infinities are equal:

Scenario	$\lvert V\rvert$	$S(p, q)$	Growth as $\varepsilon \to 0$
Single tail miss	1	0.01	$0.01(-\log \varepsilon)$
Single mode miss	1	0.90	$0.90(-\log \varepsilon)$
Complete disjoint	n	1.00	$1.00(-\log \varepsilon)$

A model that misses a tail with 1% mass diverges 100× slower than a model with completely disjoint support. The violating mass $S(p, q)$ is the precise measure of singularity severity.

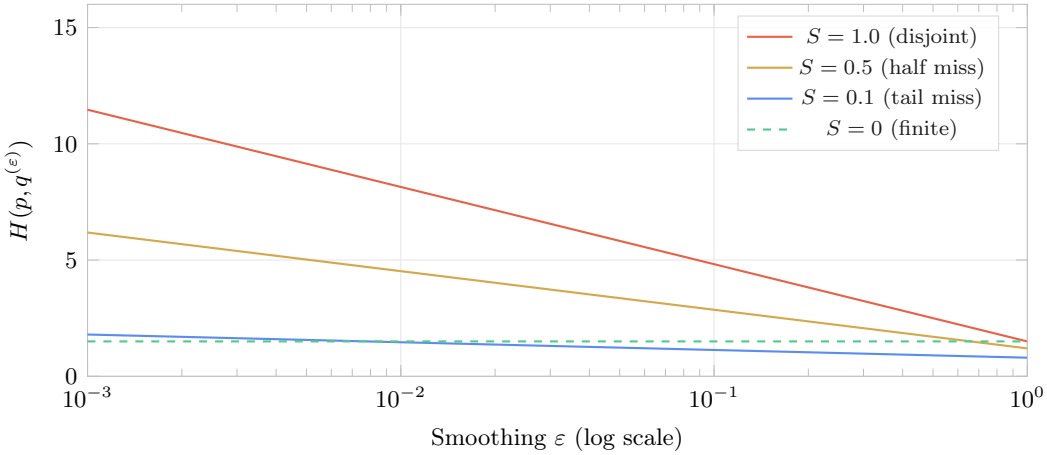

Figure 1: Cross-entropy under uniform smoothing $q^{(\varepsilon)}$ for varying violating mass $S(p, q)$. The slope of the divergence as $\varepsilon \to 0$ is exactly $S(p, q)$. Complete disjoint support ($S = 1$) diverges fastest; a small tail miss ($S = 0.1$) diverges slowly but still reaches $+\infty$. When $S = 0$ (green dashed), the loss is finite and independent of ε.

4 The Asymmetry of Cross-Entropy

Cross-entropy is not symmetric in its arguments: $H(p, q) \neq H(q, p)$ in general. The asymmetry is not a notational curiosity. It produces a sharp operational distinction at the boundary of the simplex, and it is the mechanism by which models trained under cross-entropy learn to prefer overcoverage to silence.

Proposition 7 (Directional asymmetry)**.** *For distributions p and q:*

(a) If $q(x) = 0$ and $p(x) > 0$: $\quad -p(x) \log q(x) = +\infty \quad$ *(infinite penalty).*

(b) If $p(x) = 0$ and $q(x) > 0$: $\quad -p(x) \log q(x) = 0 \quad$ *(zero penalty).*

Proof. (a) follows from $\log 0 = -\infty$ and $p(x) > 0$. (b) follows from $0 \cdot \log q(x) = 0$ by convention. □

Remark. This asymmetry has a clear operational interpretation:

- **Missing truth ($q = 0$ where $p > 0$)**: The model is overconfident—it assigns zero probability to something that can happen. This is an infinite penalty. One missing coordinate is enough.

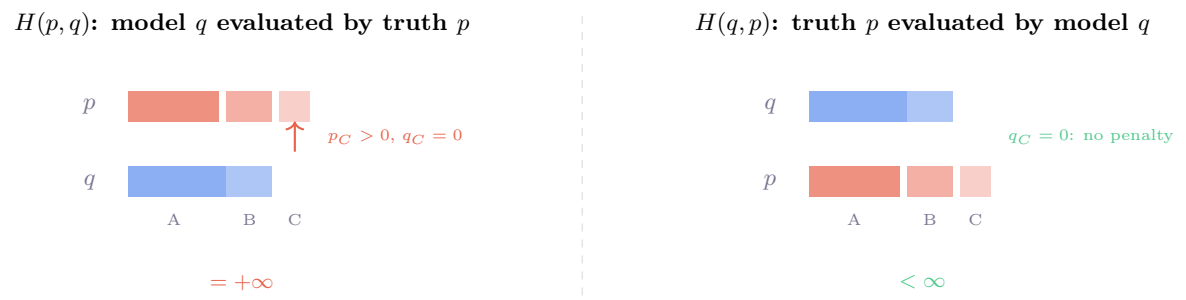

Figure 2: The asymmetry of cross-entropy. **Left**: $H(p, q)$ diverges because the model q assigns zero mass to coordinate C where the truth p has positive mass. **Right**: $H(q, p)$ is finite because the truth p covers all of q's support—the extra mass p_C costs nothing. Missing truth is catastrophic; adding falsehood is free.

- **Adding falsehood ($q > 0$ where $p = 0$)**: The model is underconfident—it assigns probability to something that cannot happen. This is *zero* penalty. The model can hallucinate with impunity.

Cross-entropy enforces *coverage* with infinite force but imposes *no constraint* on precision.

5 KL Divergence and the Impossible Zero

Because $D_{\mathrm{KL}}(p\|q) = H(p, q) - H(p)$ and $H(p)$ is finite for any distribution on a finite alphabet, the singularity structure of KL divergence inherits directly from that of cross-entropy.

Theorem 8 (KL divergence is undefined at disjoint support)**.** *If* $\mathrm{supp}(p) \cap \mathrm{supp}(q) = \varnothing$*, then* $D_{\mathrm{KL}}(p\|q) = +\infty$*. More generally,* $D_{\mathrm{KL}}(p\|q) = +\infty$ *if and only if* $V_{p,q} \neq \varnothing$*.*

Proposition 9 (Gradient explosion near the boundary)**.** *The gradient of* $D_{\mathrm{KL}}(p\|q)$ *with respect to* q_i *(under the simplex constraint) is:*

$$\frac{\partial D_{\mathrm{KL}}}{\partial q_i} = -\frac{p_i}{q_i}. \tag{5}$$

As $q_i \to 0^+$ *with* $p_i > 0$*, this gradient diverges:* $|\nabla_{q_i}| \to +\infty$*.*

Proof. $D_{\mathrm{KL}}(p\|q) = \sum_x p(x) \log p(x) - \sum_x p(x) \log q(x)$. Differentiating with respect to q_i: $\partial D_{\mathrm{KL}}/\partial q_i = -p_i/q_i$. □

5.1 The Geometric Interpretation

Disjoint support is the probabilistic analog of vector orthogonality. Two distributions whose supports don't intersect are like two vectors with zero dot product: they share no common events, they occupy independent regions of the sample space.

Remark.

	Vector space	**Probability space**
Same	$\cos\theta = +1$	$p = q$
Independent	$\cos\theta = 0$ (orthogonal)	$\mathrm{supp}(p) \cap \mathrm{supp}(q) = \varnothing$
Distance at "same"	$d = 0$ (well-defined)	$D_{\mathrm{KL}} = 0$ (well-defined)
Distance at "independent"	$d = \sqrt{2}$ (well-defined)	$D_{\mathrm{KL}} = +\infty$ (**undefined**)

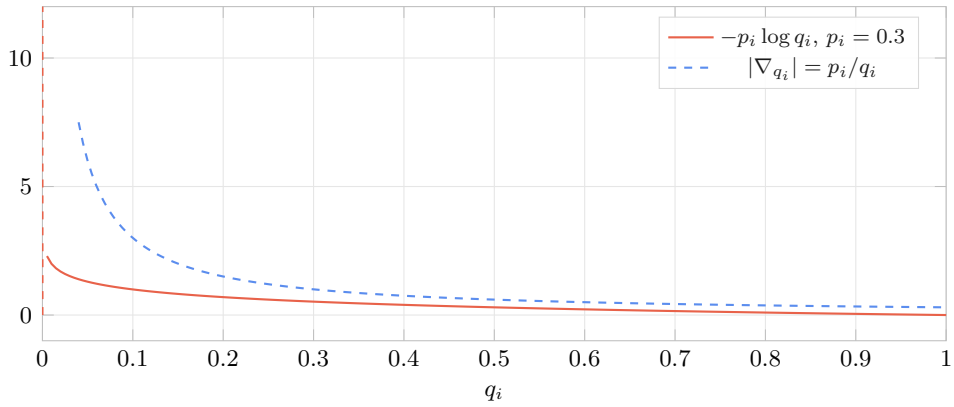

Figure 3: The contribution $-p_i \log q_i$ (solid red) and its gradient magnitude p_i/q_i (dashed blue) as functions of q_i. Both diverge as $q_i \to 0^+$: the loss becomes infinite and the gradient explodes. Any optimization that pushes q_i toward zero faces unbounded gradients near the boundary.

Geometric orthogonality is a perfectly reachable configuration with a well-defined distance. Probabilistic orthogonality is a singularity: the distance metric breaks at the very point that represents genuine independence.

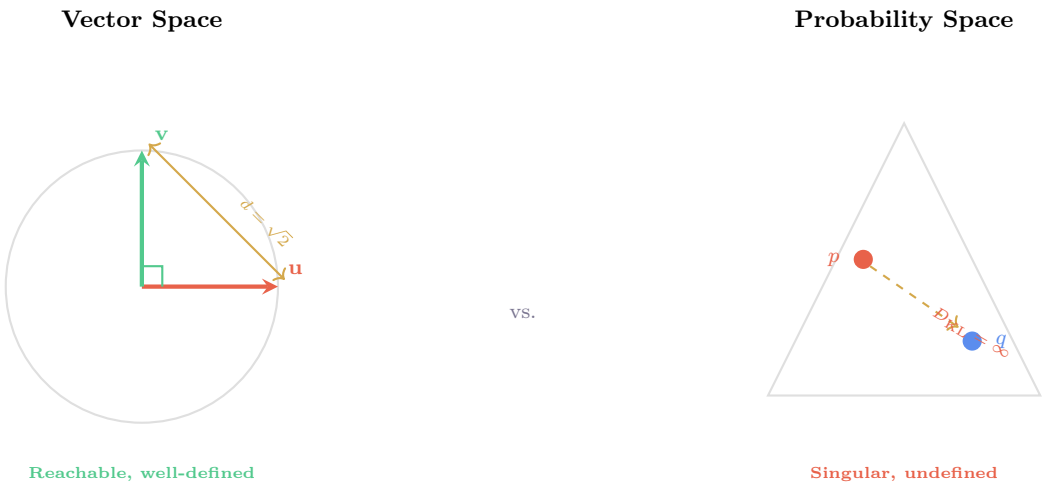

Figure 4: The fundamental asymmetry between vector geometry and probability geometry. **Left**: Orthogonal vectors have a well-defined Euclidean distance ($\sqrt{2}$ on the unit circle). **Right**: Distributions with disjoint support (the probabilistic analog of orthogonality) have $D_{\mathrm{KL}} = +\infty$. The geometric framework handles independence naturally; the information-theoretic framework breaks at the same point.

6 Divergences Finite at the Boundary

The singularity at disjoint support is not an inevitable feature of all divergence measures. Several alternatives are well-defined at the boundary; we record their behaviour for reference.

Jensen–Shannon divergence is bounded by $\log 2$ and is well-defined everywhere; it was used in the original GAN formulation [6]. Wasserstein distance [7] does not require overlapping support and metrises weak convergence, making it the natural candidate for losses that need to measure distances at the boundary. Maximum Mean Discrepancy operates in a reproducing kernel Hilbert space and is finite for any pair of distributions. The cost paid for finiteness at the boundary

Table 1: Divergence measures and their behaviour at disjoint support.

Divergence	**At disjoint supp.**	**Symmetric**	**Diff'able**
Cross-entropy $H(p, q)$	$+\infty$	No	Yes (interior)
$D_{\mathrm{KL}}(p\|q)$	$+\infty$	No	Yes (interior)
Jensen–Shannon JSD	$\leq \log 2$	Yes	Yes
Wasserstein $W_1(p, q)$	Finite	Yes	Subgradient
MMD $\mathrm{MMD}^2(p, q)$	Finite	Yes	Yes
Total variation $\mathrm{TV}(p, q)$	≤ 1	Yes	Subgradient

differs across these alternatives—computational, statistical, and in terms of the gradient signal available to first-order optimisation—and a thorough comparison is beyond the scope of the present paper.

7 Discussion

The preceding sections are mathematical content; this one is interpretive. Three of the most persistent failure modes of modern multimodal and contrastive systems—the modality gap, hallucination, and InfoNCE's batch-size hunger—admit a natural reading in terms of that structure. We present each as a *hypothesis* unifying observed behaviour with the singularity, not as a derivation that displaces the existing accounts in the literature. Each of these phenomena has competing explanations that any complete analysis must engage with; the singularity-shadow lens is one factor among them, offered here for its unifying value rather than as sole cause.

7.1 The Modality Gap

CLIP [1] and SigLIP [2] train image and text encoders to produce embeddings in a shared space. Empirically, image and text embeddings do not overlap—they form two disjoint clusters separated by a persistent *modality gap* [5].

Remark 1 (The modality gap). Let p_{img} and p_{txt} denote the distributions of image and text embeddings on $\mathbb{S}^{d-1}$. The contrastive loss wants mismatched pairs to have orthogonal embeddings, which in the distributional sense corresponds to pushing toward $\mathrm{supp}(p_{\mathrm{img}}) \cap \mathrm{supp}(p_{\mathrm{txt}}) = \varnothing$. The cross-entropy-based loss diverges at this configuration (Theorem 4). The optimisation therefore cannot reach orthogonal separation as a limit point; the residual "safe distance" it settles for is consistent with the empirically observed modality gap.

Remark. The reading above is offered as a unifying hypothesis, not a sole-cause explanation. The modality gap has at least three other live accounts in the literature. **(i)** Liang et al. [5] document a gap present at *initialisation*, before any contrastive training, attributable to the geometry of randomly initialised deep encoders. **(ii)** Cone effects in deep networks concentrate embeddings in narrow regions of the sphere irrespective of the contrastive objective. **(iii)** Optimisation dynamics around the temperature parameter shape how aggressively the loss drives negatives across the equator. The singularity-shadow account is consistent with each of these and may compose with them; a clean separation is, at this stage, an open empirical question.

7.2 Hallucination Under the Asymmetric Loss

Large multimodal models hallucinate: they describe objects that do not exist in the image, invent facts, and produce confidently wrong outputs.

Remark 2 (Overcoverage bias). Let p be the true distribution over outputs and q the model's distribution. At the coordinate level, cross-entropy contributions decompose into:

(a) if $q(x) = 0$ and $p(x) > 0$, the penalty is $+\infty$;

(b) if $q(x) > 0$ and $p(x) = 0$, the penalty is 0.

Per coordinate, putting mass on a real outcome that the model had zeroed eliminates the only unbounded term in the loss, while putting mass on an impossible outcome is, term by term, free. This produces a systematic bias toward overcoverage under uncertainty.

Remark 3 (The simplex constraint qualifies the "optimal strategy" framing). The per-coordinate result above is exact, but it does not on its own characterise an unconstrained optimum. Under the softmax that produces q, the coordinates are coupled by the simplex constraint $\sum_x q(x) = 1$. Mass placed on a falsehood necessarily comes from somewhere, including, in general, from real outcomes; "adding falsehood" is not literally free once one accounts for the displaced mass on truth. The proposition therefore characterises a *bias* rather than an *optimum*: the direction in which gradients push under coverage-vs-precision tradeoffs, not the loss-minimising configuration of the model. Stronger versions of the claim—such as "hallucination is the optimal strategy under the loss"—require either dropping the simplex constraint or coupling the loss with regularisation that breaks the symmetry.

Example. The bias is recognisable in practice. A model that distributes its output mass across 10 described objects in an image of 5 real objects pays a smaller marginal cross-entropy for the 5 hallucinated descriptions than the cost of zeroing out any of the 5 real ones. The asymmetry is one mechanism among the empirical reasons multimodal models tend to overgenerate rather than undergenerate; it is consistent with the observed pattern but does not exhaust it.

7.3 The Computational Cost of Fighting the Singularity

InfoNCE [3]—the loss behind CLIP and SimCLR [4]—computes a softmax cross-entropy over the batch similarity matrix. For negative pairs, the loss pushes similarities toward the boundary where the singularity lives.

Remark 4 (Singularity flatness and batch-size requirements). The InfoNCE gradient for negative pair k is weighted by:

$$w_k = \frac{\exp(\text{sim}_k/\tau)}{\sum_j \exp(\text{sim}_j/\tau)}. \tag{6}$$

As negative similarities approach the singularity boundary (near orthogonality, sim ≈ 0), the softmax weights become approximately uniform, $w_k \approx 1/N$, and each negative contributes $O(1/N)$ gradient. Accumulating useful signal in this regime requires N large. This is consistent with the large batches used in practice—for example, CLIP's $N = 32{,}768$—though it is one contributing factor among several.

Remark 5. Other factors known to drive contrastive batch-size requirements compose with the singularity-flatness effect described above. Hard-negative mining benefits from larger batches because the chance of seeing genuinely informative negatives grows with N; gradient variance reduction benefits from larger batches because the InfoNCE estimator's variance scales with $1/N$; the temperature schedule, the embedding dimension, and the choice of similarity function all shape the regime in which the softmax operates. The proposition above identifies one mechanism within this composition; it does not claim to be the dominant one.

Remark. The three phenomena—modality gap, hallucination bias, contrastive batch-size hunger—admit a unifying reading:

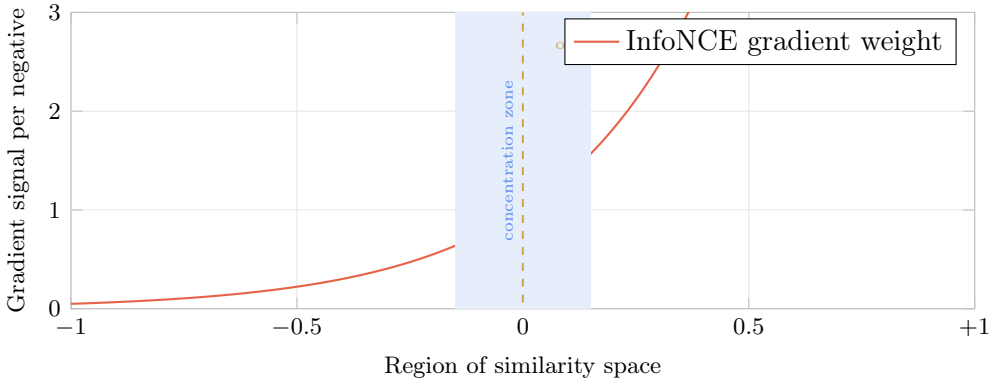

Figure 5: InfoNCE gradient weight as a function of pairwise similarity. Near orthogonality ($\cos \approx 0$), where most high-dimensional random vectors concentrate, the gradient weight is exponentially small. The loss needs enormous batches to accumulate signal from this flat region.

Cross-entropy diverges at the configuration its consumers are trying to achieve.

The modality gap is consistent with the optimisation maintaining a safe distance from the singularity, the asymmetry of cross-entropy biases models toward overcoverage under uncertainty, and the singularity-flatness effect is one driver of the gradient-signal pressure that demands large batches. These are hypothesised structural causes; each phenomenon also has independent contributing accounts in the literature, with which the singularity reading composes rather than competes.

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] 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

[4] 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

[5] V. W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Y. Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in NeurIPS*, 2022. https://arxiv.org/abs/2203.02053

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in NeurIPS*, 2014.

[7] C. Villani. *Optimal Transport: Old and New*. Springer, 2009.