The Geometry of High-Dimensional Representations: Large Concentration, Orthogonality, and Metric Degeneracy

Chima Emmanuel and Angel Ezeahurukwe

Department of Software Engineering

Department of Mechatronics Engineering

FUTO Research Intensive Student Experience

Abstract

This paper presents a self-contained, mathematically rigorous treatment of the geometry of high-dimensional Euclidean spaces, aimed at readers with a background in probability theory and linear algebra but limited prior exposure to concentration-of-measure phenomena. We develop, from first principles, four interconnected results: (i) the concentration of norms around a deterministic value; (ii) the concentration of pairwise distances; (iii) the asymptotic orthogonality of independent random vectors; and (iv) the resulting collapse of relative distance structure, which we term metric degeneracy. For each result we provide explicit probability bounds, full proof details, and intuitive commentary. We then apply these theoretical findings to explain several widely observed failure modes in machine learning including the breakdown of nearest-neighbour classifiers, the degeneration of radial-basis-function kernels, the necessity of temperature scaling in contrastive learning, and the hubness phenomenon. The overarching message is that high-dimensional spaces are not simply "bigger" than low-dimensional ones; they are governed by a fundamentally different probabilistic geometry that practitioners must understand in order to design robust algorithms.

1 Introduction

When students first encounter Euclidean geometry they learn that space is isotropic: all directions are equivalent, objects spread out uniformly, and distance is a reliable indicator of similarity. This intuition, refined over centuries of human experience, applies faithfully in two or three dimensions. But modern machine learning models operate in a very different regime. Word embeddings live in 300 dimensions. Transformer hidden states inhabit spaces of dimension 768, 1024, or higher. Image features extracted by deep convolutional networks may reside in spaces of dimension 2048 or more.

In all of these settings, geometric intuition that is built on low-dimensional experience becomes not merely imprecise but actively misleading. The culprit is a collection of related phenomena grouped under the umbrella term **concentration of measure**. Loosely, concentration of measure refers to the fact that in high dimensions, a smooth function of many independent random variables is very close to its mean with overwhelmingly high probability. The randomness, rather than spreading values across a wide range, squeezes them into a narrow band. This has dramatic consequences for geometry: norms concentrate, pairwise distances concentrate, angles between random vectors shrink toward 90 degrees, and ultimately all inter-point distances become approximately equal.

This paper provides a precise, beginner-friendly treatment of these phenomena. We proceed carefully, stating every assumption, unfolding every proof step, and providing intuitive commentary alongside the mathematics. Our goal is not merely to state classical results but to make them fully transparent, so that the reader comes away with a working understanding of why high-dimensional geometry is structurally different and what that means for the machine learning algorithms that operate in it.

1.1 Roadmap

The paper is divided into sections. Section 1 gives the general introduction. Section 2 fixes notation and recalls the sub-Gaussian family of distributions, which is the natural probabilistic setting for our results. Section 3 proves norm concentration. Section 4 proves distance concentration. Section 5 establishes asymptotic orthogonality. Section 6 formalises metric degeneracy and quantifies the rate of collapse. Section 7 unifies these results into a single master theorem. Section 8 draws out the implications for machine learning. Section 9 concludes with a discussion of design principles that follow from our analysis.

2 Notation and Preliminaries

We write vectors in bold: $\mathbf{x} \in \mathbb{R}^d$. The Euclidean norm is denoted $\|\mathbf{x}\| = (\sum_i x_i^2)^{1/2}$. The inner product of $\mathbf{x}$ and $\mathbf{y}$ is $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_i x_i y_i$. The cosine similarity between $\mathbf{x}$ and $\mathbf{y}$ is $\cos(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle / (\|\mathbf{x}\| \|\mathbf{y}\|)$.

We use O_p and o_p for the standard probabilistic order notation: $X_n = O_p(a_n)$ means that for every $\varepsilon > 0$ there exists M such that $P(|X_n/a_n| > M) < \varepsilon$ for all n, and $X_n = o_p(a_n)$ means $X_n/a_n \xrightarrow{P} 0$.

2.1 Sub-Gaussian Random Variables

Our results hold for a broad and practically important class of distributions called sub-Gaussian distributions. This class includes Gaussian, Rademacher (symmetric ± 1 Bernoulli), and any bounded random variable, making our theorems widely applicable.

Definition 2.1 (Sub-Gaussian Random Variable)**.** *A mean-zero random variable X is called sub-Gaussian with parameter $K > 0$ if its moment generating function satisfies:*

$$\mathbb{E}[\exp(tX)] \leq \exp(K^2 t^2/2) \quad \textit{for all } t \in \mathbb{R}.$$

Equivalently, X is sub-Gaussian if and only if there exists a constant C such that $P(|X| \geq u) \leq 2\exp(-u^2/C^2)$ for all $u \geq 0$. The smallest K for which Definition 2.1 holds is called the sub-Gaussian norm of X and is written $\|X\|_{\psi_2}$. Gaussian $\mathcal{N}(0, \sigma^2)$ random variables are sub-Gaussian with $K = \sigma$. Bounded random variables $|X| \leq B$ are sub-Gaussian with $K = B$ (by Hoeffding's lemma).

Definition 2.2 (Sub-Exponential Random Variable)**.** *A mean-zero random variable Z is sub-exponential if $\mathbb{E}[\exp(tZ)] \leq \exp(K^2 t^2/2)$ for $|t| \leq 1/K$. Sub-exponential variables have heavier tails than sub-Gaussian ones but still lighter than polynomial tails. Crucially, if X is sub-Gaussian then X^2 is sub-exponential; a fact we will use in the proof of norm concentration.*

Definition 2.3 (Isotropic Vector)**.** *A random vector $\mathbf{X} \in \mathbb{R}^d$ is called isotropic if $\mathbb{E}[\mathbf{X}] = 0$ and $Cov(\mathbf{X}) = I_d$, i.e., the components are uncorrelated and each has unit variance. Throughout this paper we assume $\mathbf{X}$ is isotropic with i.i.d. sub-Gaussian components with sub-Gaussian norm at most K. The isotropic assumption is not a restriction in practice: any random vector with non-degenerate covariance matrix Σ can be whitened i.e. multiplied by $\Sigma^{-1/2}$ to become isotropic, after which our results apply.*

2.2 Key Concentration Inequality

Our proofs rest on the following classical inequality.

Theorem 2.4 (Bernstein's Inequality)**.** *Let $Z_1, \ldots, Z_d$ be independent mean-zero sub-exponential random variables with sub-exponential norms at most K. Then for every $t > 0$:*

$$P\left(\left|\sum_{i=1}^{d} Z_i/d - \mathbb{E}[Z_1]\right| \geq t\right) \leq 2\exp\left(-c \cdot d \cdot \min\left(\frac{t^2}{K^4}, \frac{t}{K^2}\right)\right)$$

where $c > 0$ is a universal constant. In the regime $t \leq K^2$ (which covers all our applications), this simplifies to $2\exp(-cdt^2/K^4)$, a purely Gaussian tail.

3 Norm Concentration: Vectors Live on a Thin Shell

3.1 Intuition

In one dimension, a scalar random variable can take any value in its support. In two dimensions, a random vector can point in any direction and have any length within a moderate range. What happens as d grows? Consider $\|\mathbf{X}\|^2 = X_1^2 + X_2^2 + \cdots + X_d^2$. Each term X_i^2 has mean 1 and some variance. The sum is an average of d i.i.d. quantities scaled by d. By the law of large numbers, $\|\mathbf{X}\|^2/d \to 1$. By concentration inequalities, the fluctuations around this limit are exponentially small.

Consequently, $\|\mathbf{X}\| \approx \sqrt{d}$ with very high probability implies that all random vectors of dimension d are approximately the same length! One can visualise this as follows. In $\mathbb{R}^2$, a uniform distribution over a disc places probability mass throughout the interior. As d increases, the analogous distribution over the d-dimensional ball places almost all its mass near the surface — the thin shell of radius $\sqrt{d}$. This is the thin shell phenomenon, and it is the foundational observation of high-dimensional geometry.

3.2 Formal Statement and Proof

Theorem 3.1 (Non-Asymptotic Norm Concentration)**.** *Let $\mathbf{X} = (X_1, \ldots, X_d) \in \mathbb{R}^d$ be an isotropic random vector with i.i.d. sub-Gaussian components, each with sub-Gaussian norm at most K. Then for any $\varepsilon > 0$:*

$$P\left(\left|\|\mathbf{X}\|^2/d - 1\right| \geq \varepsilon\right) \leq 2\exp(-cd\varepsilon^2/K^4)$$

where $c > 0$ is a universal constant. Equivalently, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$:

$$\|\mathbf{X}\|^2 \in [d - \varepsilon d, d + \varepsilon d]$$

provided $d \geq (K^4/c\varepsilon^2)\log(2/\delta)$.

Proof. Define $Z_i = X_i^2$ for $i = 1, \ldots, d$. Since X_i is sub-Gaussian with parameter K, the squared variable Z_i is sub-exponential. Explicitly, Z_i has mean $\mathbb{E}[X_i^2] = 1$ (by the isotropic assumption) and variance $\mathrm{Var}(X_i^2) = \mathbb{E}[X_i^4] - 1 \leq CK^4$ for a universal constant C, by the moment bounds for sub-Gaussian variables. We write:

$$\|\mathbf{X}\|^2/d = (1/d)\sum_{i=1}^{d} Z_i.$$

This is an empirical average of d i.i.d. sub-exponential random variables with mean 1. Applying Bernstein's inequality (Theorem 2.4) with $t = \varepsilon$ gives:

$$P\left(\left|(1/d)\sum_i Z_i - 1\right| \geq \varepsilon\right) \leq 2\exp(-cd\varepsilon^2/K^4)$$

which is the claimed bound. □

Corollary 3.2 (Thin Shell Phenomenon)**.** *Under the conditions of Theorem 3.1, taking square roots via the delta method:*

$$\|\mathbf{X}\| = \sqrt{d} \pm O(1) \quad \textit{with high probability.}$$

More precisely, for any $\eta > 0$, with probability at least $1 - 2\exp(-c\eta^2)$:

$$\sqrt{d}\cdot\left(1 - \frac{\eta}{2\sqrt{d}}\right) \leq \|\mathbf{X}\| \leq \sqrt{d}\cdot\left(1 + \frac{\eta}{2\sqrt{d}}\right).$$

The fluctuation is $O(1)$ — constant order while the typical norm is $\sqrt{d}$. The relative fluctuation $\|\mathbf{X}\|/\sqrt{d} - 1$ is therefore of order $1/\sqrt{d}$, vanishing as $d \to \infty$. In words: virtually all of the probability mass of an isotropic distribution lies within a thin shell of radius $\sqrt{d}$ and width $O(1)$ around the origin. The thickness of the shell does not grow with d; only its radius does.

4 Distance Concentration: All Points Are Equally Far Apart

4.1 Intuition

Having established that individual norms concentrate, we now turn to pairwise distances. Suppose $\mathbf{X}$ and $\mathbf{Y}$ are two independent isotropic vectors. Their difference $\mathbf{Z} = \mathbf{X} - \mathbf{Y}$ is also a random vector. Since $\mathbf{X}$ and $\mathbf{Y}$ are independent and isotropic, $\mathbf{Z}$ has mean zero and covariance $\mathbb{E}[\mathbf{Z}\mathbf{Z}^T] = \mathbb{E}[\mathbf{X}\mathbf{X}^T] + \mathbb{E}[\mathbf{Y}\mathbf{Y}^T] = 2I_d$. In other words $\mathbf{Z}/\sqrt{2}$ is isotropic.

Applying Theorem 3.1 to $\mathbf{Z}/\sqrt{2}$ immediately gives concentration of $\|\mathbf{X} - \mathbf{Y}\|^2$ around $2d$. The practical consequence is startling: any two randomly drawn high-dimensional vectors will be approximately the same distance apart, namely $\sqrt{2d}$. As the dimension increases, the spread of the distribution of pairwise distances shrinks relative to the typical distance. There are no "close" neighbours and "far" neighbours; just the single typical distance.

4.2 Formal Statement and Proof

Theorem 4.1 (Pairwise Distance Concentration)**.** *Let $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^d$ be independent isotropic random vectors with i.i.d. sub-Gaussian components with sub-Gaussian norm at most K. Then for any $\varepsilon > 0$:*

$$P\left(\left|\|\mathbf{X} - \mathbf{Y}\|^2/(2d) - 1\right| \geq \varepsilon\right) \leq 2\exp(-cd\varepsilon^2/K^4)$$

where $c > 0$ is the same universal constant as in Theorem 3.1.

Proof. Let $\mathbf{Z} = \mathbf{X} - \mathbf{Y}$. Since $\mathbf{X}$ and $\mathbf{Y}$ are independent:

$$\mathbb{E}[\mathbf{Z}] = \mathbb{E}[\mathbf{X}] - \mathbb{E}[\mathbf{Y}] = 0$$

$$\mathrm{Cov}(\mathbf{Z}) = \mathrm{Cov}(\mathbf{X}) + \mathrm{Cov}(\mathbf{Y}) = I_d + I_d = 2I_d.$$

Define $\mathbf{W} = \mathbf{Z}/\sqrt{2}$. Then $\mathbf{W}$ is isotropic: $\mathbb{E}[\mathbf{W}] = 0$ and $\mathrm{Cov}(\mathbf{W}) = I_d$. Furthermore the components $W_i = (X_i - Y_i)/\sqrt{2}$ are independent; since X_i and Y_i are sub-Gaussian with parameter K, their difference is sub-Gaussian with parameter at most $\sqrt{2}K$ (using the triangle inequality for sub-Gaussian norms: $\|X - Y\|_{\psi_2} \le \|X\|_{\psi_2} + \|Y\|_{\psi_2}$), and dividing by $\sqrt{2}$ brings the parameter back to $K\sqrt{2}/\sqrt{2} = K$. Therefore the components of $\mathbf{W}$ are sub-Gaussian with parameter K, and $\mathbf{W}$ satisfies the hypotheses of Theorem 3.1. Applying Theorem 3.1 to $\mathbf{W}$:

$$P(|\|\mathbf{W}\|^2/d - 1| \ge \varepsilon) \le 2\exp(-cd\varepsilon^2/K^4).$$

Since $\|\mathbf{W}\|^2 = \|\mathbf{Z}\|^2/2 = \|\mathbf{X}-\mathbf{Y}\|^2/2$, we have $\|\mathbf{W}\|^2/d = \|\mathbf{X}-\mathbf{Y}\|^2/(2d)$, and the result follows. □

Corollary 4.2. *Applying the delta method as in Corollary 3.2:*

$$\|\mathbf{X} - \mathbf{Y}\| = \sqrt{2d} \pm O(1) \quad \textit{with high probability.}$$

The absolute spread of pairwise distances is $O(1)$ — constant — while the typical distance grows as $\sqrt{2d}$. The ratio of fluctuation to typical distance is therefore $O(1/\sqrt{d})$, which tends to zero.

5 Asymptotic Orthogonality of Random Vectors

5.1 Intuition

We have seen that norms and distances concentrate. Now consider the angle between two random vectors. Cosine similarity measures $\cos(\theta) = \langle \mathbf{X}, \mathbf{Y}\rangle/(\|\mathbf{X}\|\|\mathbf{Y}\|)$. In two dimensions, two random unit vectors drawn uniformly on the circle can have cosine similarities ranging from -1 to 1. In high dimensions, we will show that $\cos(\theta)$ collapses to zero: all pairs of random vectors are nearly perpendicular.

To see why, consider the dot product $\langle \mathbf{X}, \mathbf{Y}\rangle = \sum_i X_i Y_i$. Each term $X_i Y_i$ has mean zero and variance 1. There are d such terms. By the Central Limit Theorem, $\langle \mathbf{X}, \mathbf{Y}\rangle$ behaves like a sum of d mean-zero, unit-variance random variables — its typical magnitude is $\sqrt{d}$. Meanwhile, $\|\mathbf{X}\| \approx \sqrt{d}$ and $\|\mathbf{Y}\| \approx \sqrt{d}$, so $\|\mathbf{X}\|\|\mathbf{Y}\| \approx d$. Therefore $\cos(\theta) \approx \sqrt{d}/d = 1/\sqrt{d} \to 0$.

5.2 Formal Statement and Proof

Theorem 5.1 (Vanishing Cosine Similarity)**.** *Let $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^d$ be independent isotropic sub-Gaussian vectors. Then:*

$$\langle \mathbf{X}, \mathbf{Y}\rangle/(\|\mathbf{X}\|\|\mathbf{Y}\|) = O_p(1/\sqrt{d})$$

and in particular the cosine similarity converges in probability to zero:

$$\langle \mathbf{X}, \mathbf{Y}\rangle/(\|\mathbf{X}\|\|\mathbf{Y}\|) \xrightarrow{P} 0 \quad \textit{as } d \to \infty.$$

Proof. We analyse numerator and denominator separately.

Step 1: The dot product. Define $S_d = \langle \mathbf{X}, \mathbf{Y}\rangle = \sum_{i=1}^d X_i Y_i$. The summands $U_i = X_i Y_i$ are i.i.d. with:

$$\mathbb{E}[U_i] = \mathbb{E}[X_i]\mathbb{E}[Y_i] = 0 \cdot 0 = 0$$

$$\mathrm{Var}(U_i) = \mathbb{E}[X_i^2 Y_i^2] - 0 = \mathbb{E}[X_i^2]\mathbb{E}[Y_i^2] = 1 \cdot 1 = 1.$$

Thus $\mathrm{Var}(S_d) = d$, and $S_d/\sqrt{d} \xrightarrow{d} \mathcal{N}(0,1)$ by the Central Limit Theorem. In O_p notation: $\langle \mathbf{X}, \mathbf{Y} \rangle = O_p(\sqrt{d})$.

Step 2: The norm product. From Theorem 3.1, $\|\mathbf{X}\|^2/d \xrightarrow{P} 1$, so $\|\mathbf{X}\|/\sqrt{d} \xrightarrow{P} 1$, i.e., $\|\mathbf{X}\| = \sqrt{d}(1 + o_p(1))$. The same holds for $\mathbf{Y}$. Therefore:

$$\|\mathbf{X}\|\|\mathbf{Y}\| = d(1 + o_p(1)) = d + o_p(d).$$

Step 3: Combine. By Slutsky's theorem:

$$\langle \mathbf{X}, \mathbf{Y} \rangle / (\|\mathbf{X}\|\|\mathbf{Y}\|) = O_p(\sqrt{d})/(d(1 + o_p(1))) = O_p(1/\sqrt{d}) \xrightarrow{P} 0.$$

□

Near-orthogonality is exploited in random projections, compressed sensing, and the design of embedding spaces in machine learning.

6 Metric Degeneracy: Distance Loses Discriminative Power

6.1 Intuition

Corollary 4.2 tells us that every distance $D_i = \|\mathbf{X} - \mathbf{X}_i\|$ concentrates around $\sqrt{2d}$ with fluctuations of only $O(1)$. As d grows, all distances collapse to the same value, and the ratio $(\max D_i - \min D_i)/\min D_i \to 0$. There is no meaningful "nearest" neighbour since all points are equidistant.

6.2 Formal Statement

Theorem 6.1 (Relative Distance Collapse)**.** *Let $\mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^d$ be i.i.d. isotropic sub-Gaussian vectors with sub-Gaussian norm at most K, and let $D_i = \|\mathbf{X} - \mathbf{X}_i\|$. Then:*

$$\max_i D_i - \min_i D_i = O_p(\sqrt{\log n})$$

$$(\max_i D_i - \min_i D_i)/\min_i D_i \xrightarrow{P} 0 \quad \textit{as } d \to \infty \textit{ with } n \textit{ fixed.}$$

Proof. From Corollary 4.2 applied to each pair $(\mathbf{X}, \mathbf{X}_i)$:

$$D_i = \sqrt{2d} + \xi_i, \quad \text{where } \xi_i = O_p(1) \text{ for each } i.$$

More precisely, the Bernstein bound from Theorem 4.1 gives, for any $t > 0$:

$$P(|D_i^2 - 2d| \geq t) \leq 2\exp(-ct^2/(K^4 d)).$$

Setting $t = K^2\sqrt{d\log(2n/\delta)}/\sqrt{c}$, a union bound over $i = 1, \ldots, n$ gives that simultaneously for all i, $D_i^2 = 2d \pm O(K^2\sqrt{d\log n})$ with probability at least $1 - \delta$. Via the delta method:

$$D_i = \sqrt{2d} \pm O(K^2\sqrt{\log n}/\sqrt{2d}) \cdot \sqrt{2d} = \sqrt{2d} \pm O(K^2\sqrt{\log n}).$$

Therefore:

$$\max_i D_i - \min_i D_i \leq O(K^2\sqrt{\log n}) \quad \text{with high probability,}$$

while $\min_i D_i \geq \sqrt{2d} - O(\sqrt{\log n}) \approx \sqrt{2d}$ for large d. Dividing:

$$(\max D_i - \min D_i)/\min D_i \leq O(\sqrt{\log n}/\sqrt{d}) \to 0.$$

□

7 Unified Theorem: The Geometry of High-Dimensional Spaces

Theorem 7.1 (Metric Degeneracy of High-Dimensional Representations)**.** *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^d$ be i.i.d. isotropic sub-Gaussian vectors. As $d \to \infty$ with n fixed, the following hold simultaneously with probability at least $1 - 2n^2 \exp(-cd/K^4)$:*

(i) ***Norm Concentration:*** $\|\mathbf{X}_i\| = \sqrt{d} \cdot (1 \pm O(1/\sqrt{d}))$

(ii) ***Distance Concentration:*** $\|\mathbf{X}_i - \mathbf{X}_j\| = \sqrt{2d} \cdot (1 \pm O(1/\sqrt{d}))$

(iii) ***Asymptotic Orthogonality:*** $\langle \mathbf{X}_i, \mathbf{X}_j \rangle = O_p(\sqrt{d})$ *while* $\|\mathbf{X}_i\| \|\mathbf{X}_j\| \approx d$

(iv) ***Metric Collapse:*** $\max_{j \neq i} \|\mathbf{X}_i - \mathbf{X}_j\| / \min_{j \neq i} \|\mathbf{X}_i - \mathbf{X}_j\| \xrightarrow{P} 1$

Consequence: *The Euclidean metric loses discriminative power. For all i, j, k with $j \neq k$:*

$$\|\mathbf{X}_i - \mathbf{X}_j\| \approx \|\mathbf{X}_i - \mathbf{X}_k\| \approx \sqrt{2d},$$

and no distance-based algorithm can reliably distinguish a nearest neighbour from a farthest one.

8 Implications for Machine Learning

8.1 Failure of Nearest-Neighbour Methods

Theorem 6.1 shows that in high dimensions, the notion of "near" collapses. For $d = 1000$, the relative contrast is of order $1/\sqrt{1000} \approx 0.03$. All 1000-dimensional data points are within 3% of the same distance from any query, a regime in which nearest-neighbour search is unreliable.

8.2 Kernel Method Degeneration

For RBF kernels $K(x, y) = \exp(-\gamma \|x - y\|^2)$, Corollary 4.2 implies $\|x - y\|^2 \approx 2d$. Thus $K(x, y) \approx \exp(-2\gamma d)$. If γ is large, the kernel is essentially zero; if γ is small, it is essentially one. Only an exponentially fine-tuned $\gamma \approx 1/(2d)$ keeps the values discriminative.

8.3 Temperature Scaling in Contrastive Learning

In InfoNCE loss, Theorem 5.1 implies cosine similarities are $O(1/\sqrt{d})$ — providing almost no gradient signal. Dividing by a temperature parameter $\tau = O(1/\sqrt{d})$ rescales the range and restores useful signal.

8.4 The Hubness Phenomenon

Certain points — **hubs** — appear as nearest neighbours of many points, while others — **anti-hubs** — appear as neighbours of almost none. This skewness arises because a point slightly closer to the dataset mean will systematically be "closer" to all other points when relative contrast is low.

9 Discussion: Design Principles

1. **Normalise representations.** Projection onto the unit sphere (ℓ_2 normalisation) removes the trivially concentrated norm degree of freedom.
2. **Use cosine similarity with care.** Theorem 5.1 guides the choice of temperature $\tau \propto 1/\sqrt{d}$.
3. **Reduce dimensionality.** PCA or Random Projections can compress data to a space where distance-based methods regain power.
4. **Relative ranking losses.** Design loss functions that use relative margins rather than absolute distances.
5. **Scale RBF bandwidth.** Bandwidth γ must scale as $O(1/d)$.

10 Conclusion

High-dimensional spaces are governed by the concentration of measure, which imposes a rigid geometric structure: all vectors have approximately the same norm, all pairs are approximately equidistant, and all pairs are approximately orthogonal. Understanding metric degeneracy is a prerequisite for designing systems that scale reliably to high-dimensional representation spaces.

References

[1] Vershynin, R. (2018). *High-Dimensional Probability*. Cambridge University Press.

[2] Beyer, K., et al. (1999). *When is "nearest neighbor" meaningful?* ICDT.

[3] Radovanović, M., et al. (2010). *Hubs in space*. Journal of Machine Learning Research.

[4] Blum, A., et al. (2020). *Foundations of Data Science*. Cambridge University Press.

[5] Chen, T., et al. (2020). *A simple framework for contrastive learning (SimCLR)*. ICML.

[6] Wainwright, M. J. (2019). *High-Dimensional Statistics*. Cambridge University Press.