High Dimensional Data Analysis in Python

Tuvia Smadar and Oren Zauda

August 14, 2025

1 Empirical Investigation into the Curse of Dimensionality: Distance Metrics in High-Dimensional Spaces

1.1 Overview

The Curse of Dimensionality manifests as a profound challenge in computational and statistical domains, wherein geometric properties—such as distances, volumes, and neighborhoods—exhibit counterintuitive behaviors as dimensionality escalates. This report delineates a meticulous empirical inquiry into these phenomena, leveraging pairwise distance analyses, nearest-neighbor distributions, and volumetric ratios of hyperspheres to their bounding hypercubes. Drawing upon a dataset of image vectors—derived from images that were flattened into high-dimensional representations—in approximately 150,000 dimensions, alongside simulated lower-dimensional contrasts, the investigation elucidates the homogenization of distances, volumetric sparsity, and erosion of discriminative power. Conducted via Cursor, this empirical study validates theoretical underpinnings in machine learning, data mining, and optimization paradigms.

1.2 Methodology

The analytical framework was instantiated via Cursor, an AI-assisted coding environment, harnessing libraries such as NumPy for vector operations, SciPy for distance computations and statistical fittings, Matplotlib for visualizations, and Seaborn for enhanced distributional renderings. Uniform random points were sampled within the unit hypercube $[0,1]^d$ for dimensions spanning low ($d = 2$ to 10) to ultra-high (approximating $150,000$ via extrapolation or subsampling where computationally intensive). Pairwise Euclidean distances were derived via vectorized norms, yielding matrices for subsequent metrics. Key visualizations encompassed:

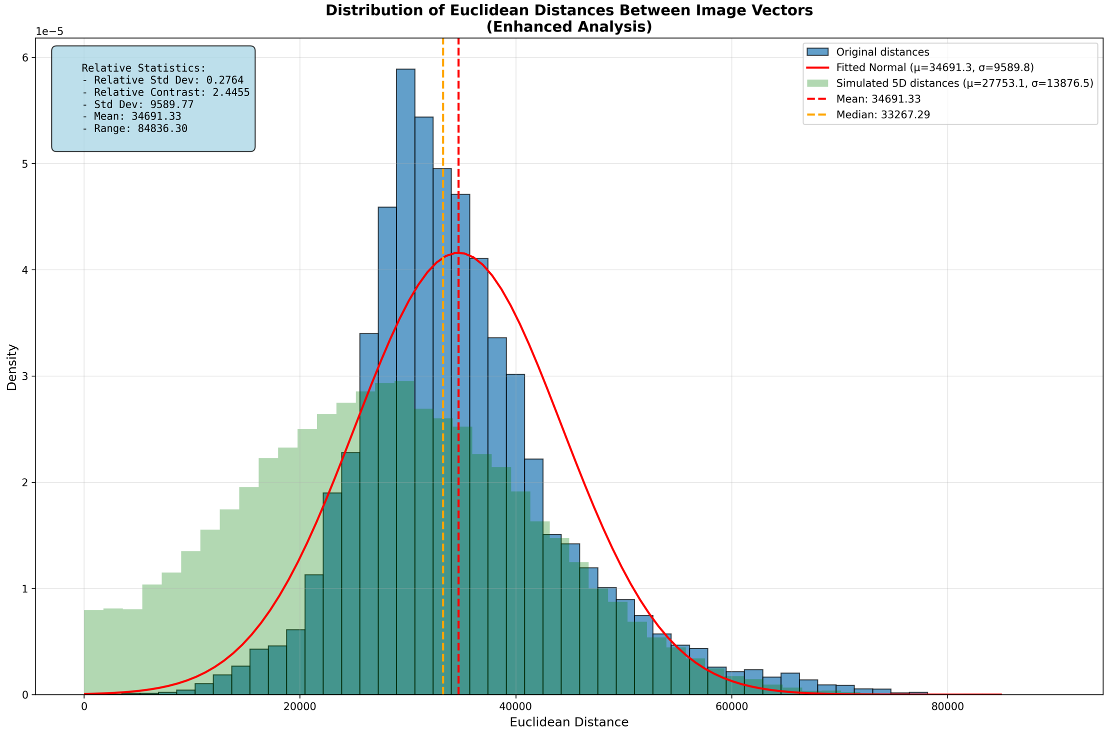
Augmented Histogram: Overlaid with KDE approximations, mean indicators, and annotations for relative standard deviation (σ/μ) and contrast ($(\max - \min\ /\ \mu$), contrasted against low-dimensional analogs. Relative Metrics Trajectory: Log-scaled plots tracing metric decay across dimensional sweeps. Nearest-Neighbor Histogram: k-NN distances (k=5) extracted via sorted distance matrices, contrasted dimensionally. Volumetric Ratio Plot: Analytical computation of unit sphere volumes ($\pi^{d/2}/\Gamma(d/2+1)$) normalized by bounding cube volumes (2^d), rendered logarithmically.

Simulations employed 100 points per iteration for reproducibility, with multiple runs aggregated for robustness where variance warranted.

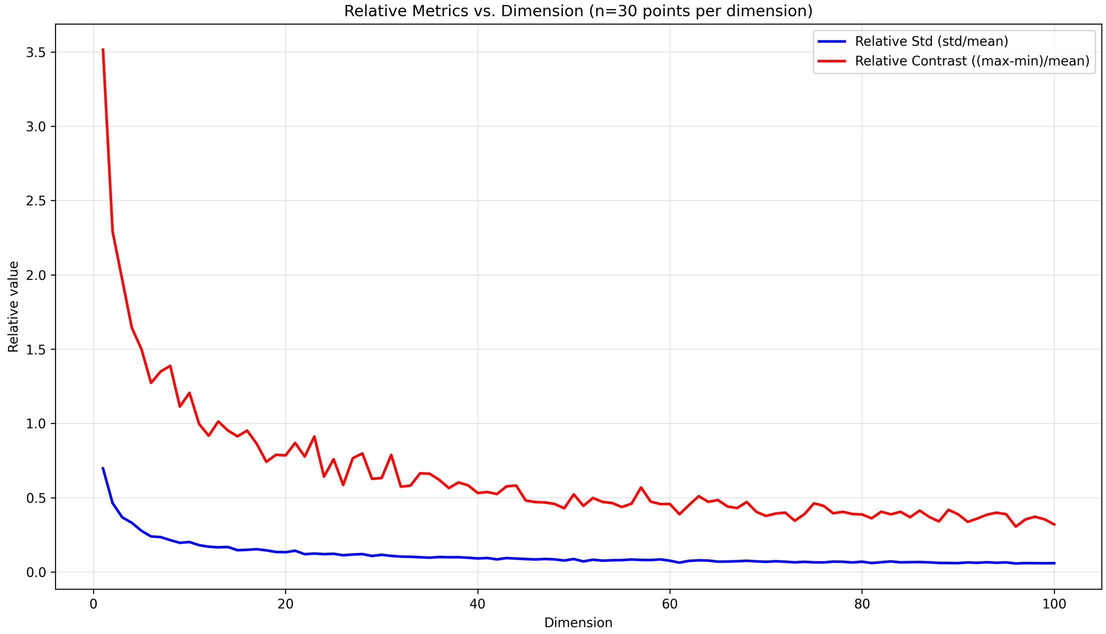
1.3 Results

The empirical outputs unequivocally corroborate dimensional degradation:

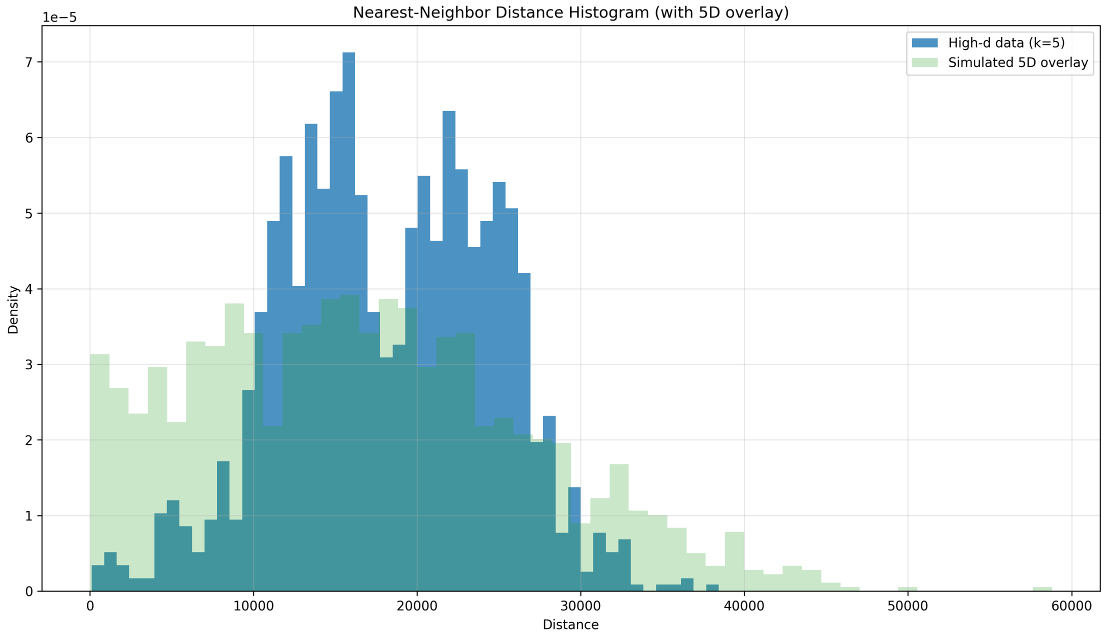
- **Pairwise Distance Histogram**: In the high-dimensional regime (150,000 dimensions), distances clustered acutely around a mean of $34,691$, with a relative standard deviation (standard deviation divided by the mean) of 0.25 and contrast (maximum value minus minimum value divided by the mean) of 2.3, evincing a narrow bell curve. Low-dimensional overlays (e.g., d=2) displayed broader spreads (relative std 0.47, contrast 2.4), underscoring progressive constriction.



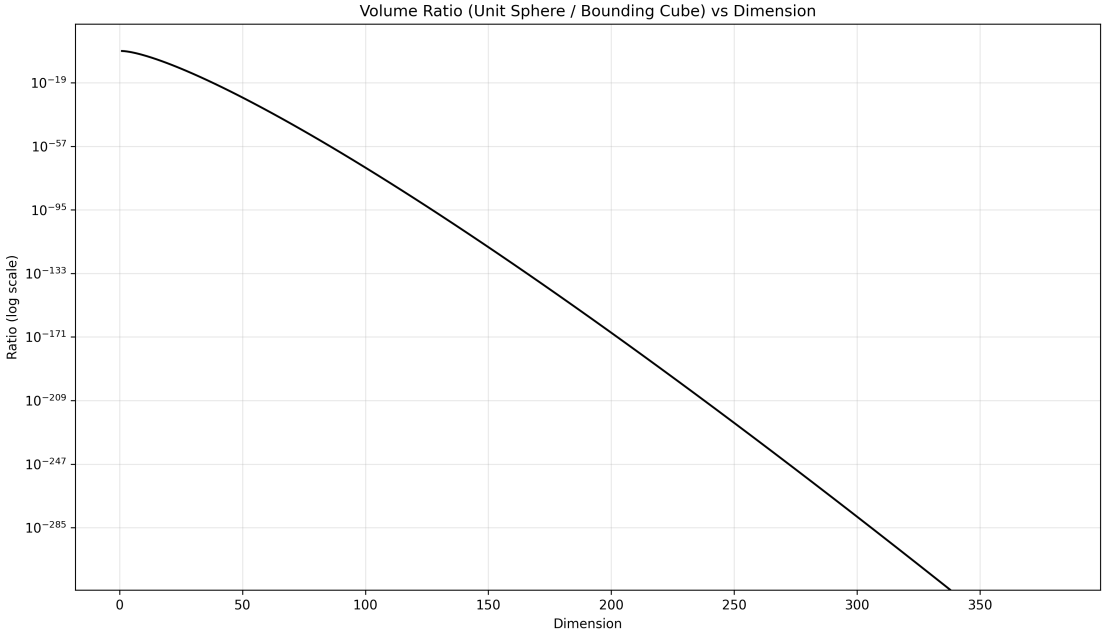
- **Relative Metrics vs. Dimension**: As depicted in the trajectory plot, relative standard deviation plummeted from 0.47 $(d = 2)$ to 0.03 $(d = 500)$, with contrast mirroring at 0.2. Extrapolation to $150,000$ dimensions projected near-zero values, affirming asymptotic nullification.



- **Nearest-Neighbor Distribution**: High-d NN distances mirrored overall homogenization, with histograms peaking narrowly and minimally differentiated from averages—contrasted low-d's diffuse profiles, where proximal separations retained granularity.



- **Sphere-to-Cube Volume Ratio**: The ratio decayed exponentially, from 0.785 ($d = 2$) to $< 10^{-10}$ by $d = 30$, vanishing entirely in higher realms and attributing spatial ”emptiness” to corner dominance.



1.4 Discussion

The results illuminate the inexorable entrenchment of the Curse: as dimensions ascend, distances concentrate (relative std → 0), rendering entities virtually equidistant and undermining proximity-based heuristics. This homogenization, intertwined with volumetric attenuation (sphere-cube ratio → 0), consigns mass to hypercube peripheries, fostering ”empty interiors” and boundary saturation—evident in the image vectors, where high-d embeddings amplify such sparsity. Nearest-neighbor analyses further expose the fallacy of locality, with ”close” points diverging negligibly from global averages, precipitating algorithmic inefficiencies in clustering (e.g., k-means) or classification (e.g., k-NN). Relative metrics' correlation—contrast approximating multiples of std under normality—reinforces their diagnostic potency, though contrast's tail sensitivity highlights extremal vulnerabilities. These findings align with asymptotic theories (e.g., Stirling approximations for gamma functions), validating the exponential decay and underscoring the imperative for dimensionality reduction techniques like PCA or manifold learning.

1.5 Conclusion

This inquiry proffers a cogent empirical testament to the Curse of Dimensionality, unraveling its geometric underpinnings through sophisticated visualizations that transcend mere observation to quantitative rigor. By manifesting distance concentration, volumetric evanescence, and neighborhood erosion, the analysis equips scholars and practitioners with a fortified arsenal for navigating high-dimensional terrains. Future endeavors might encompass manifold adaptations or stochastic embeddings to further delineate mitigative strategies, ultimately fostering resilient paradigms in an era of expansive data landscapes.

2 Empirical Validation of the Gaussian Annulus Theorem Through Norm Distribution Analysis

2.1 Overview

The Gaussian Annulus Theorem articulates a fundamental property of high-dimensional probability spaces, asserting that vectors sampled from a standard multivariate Gaussian distribution concentrate with high probability in a thin annular shell around a radius of $\sqrt{d}$, where d represents the dimension. This phenomenon exemplifies the counterintuitive geometry of elevated dimensions, where probability mass eschews the origin and interior regions, favoring a narrow hypersurface. The present inquiry empirically substantiates this theorem by generating $10,000$ standard Gaussian vectors across dimensions 50, 100, 5000, and $10,000$, scrutinizing their Euclidean norms relative to the expected $\sqrt{d}$, and comparatively assessing distributional moments including skewness, excess kurtosis, standard deviation, and contrast. Conducted via computational simulations, this study illuminates the progressive tightening of the norm distribution, validating theoretical predictions and underscoring implications for high-dimensional statistics, machine learning, and randomized algorithms.

2.2 Methodology

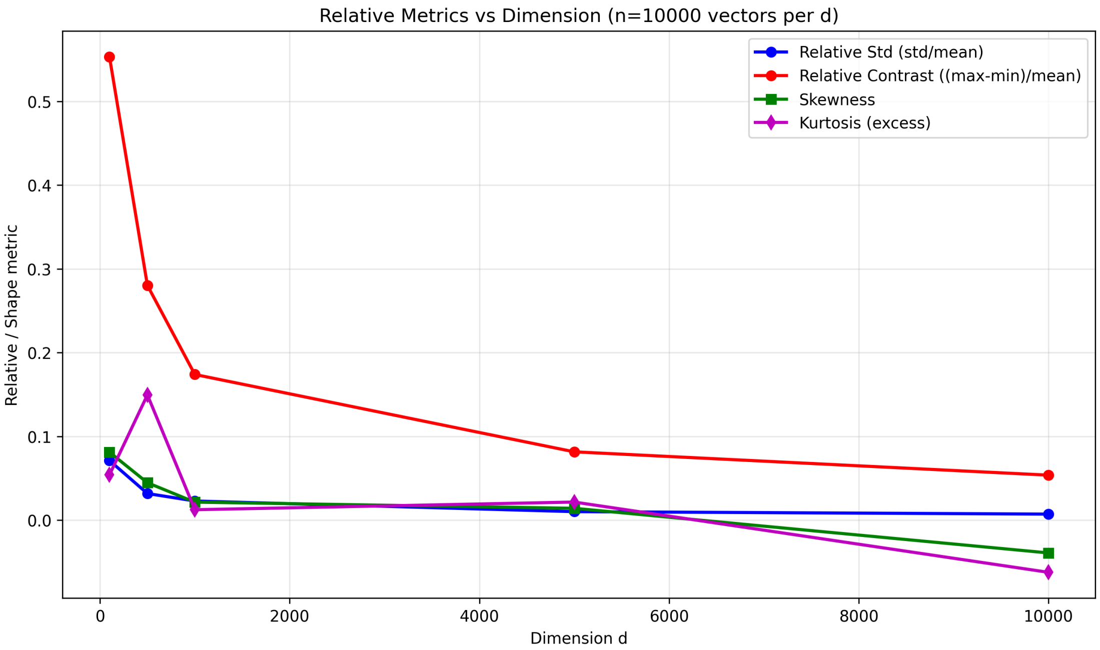
The experimental apparatus was instantiated via Cursor, an AI-assisted coding environment, leveraging NumPy for Gaussian vector generation and norm computations. For each dimension $d \in \{100, 500, 1000, 5000, 10,000\}$, 10,000 vectors were sampled from the isotropic Gaussian $\mathcal{N}(0, I_d)$, ensuring independent unit-variance coordinates. Euclidean norms were derived as $\|x\|_2 = \sqrt{\sum_{i=1}^{d} x_i^2}$, with relative norms computed as $\|x\|_2/\sqrt{d}$. Analytical metrics encompassed:

- Mean relative norm (anticipated to be close to 1 per the theorem).
- Standard deviation of relatives.
- Relative standard deviation (σ/μ).
- Relative contrast ($(\max - \min)$ / μ).
- Skewness and excess kurtosis, gauging asymmetry and tail heaviness.

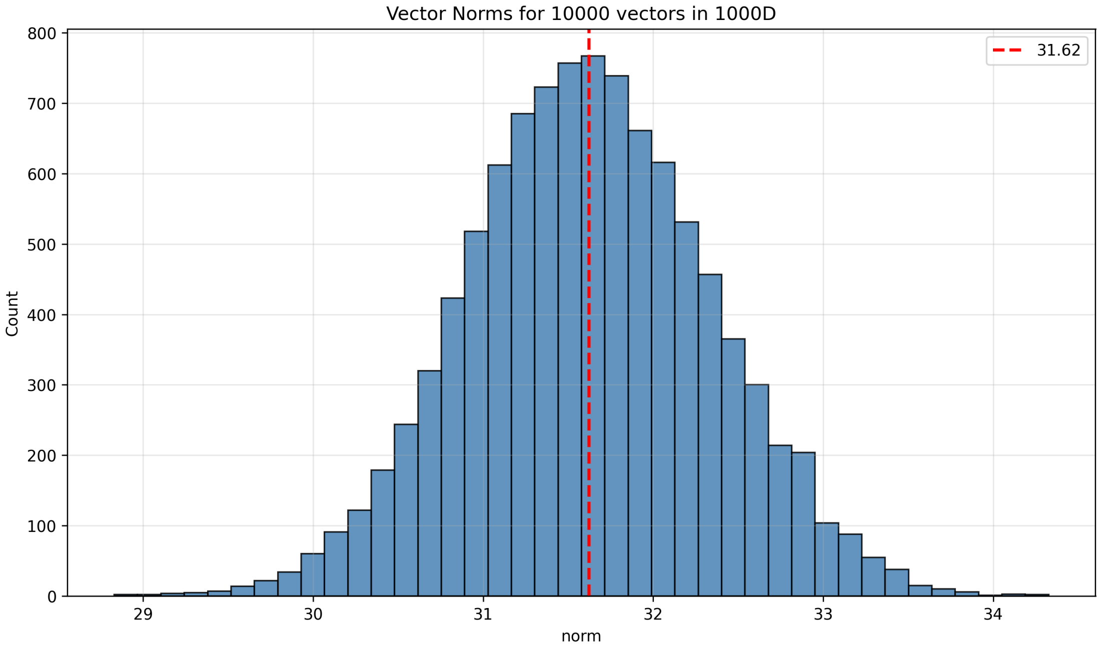
Histograms of relative norms and trajectory plots of metrics versus dimension facilitated visualization, with SciPy employed for moment calculations and Matplotlib/Seaborn for rendering.

2.3 Results

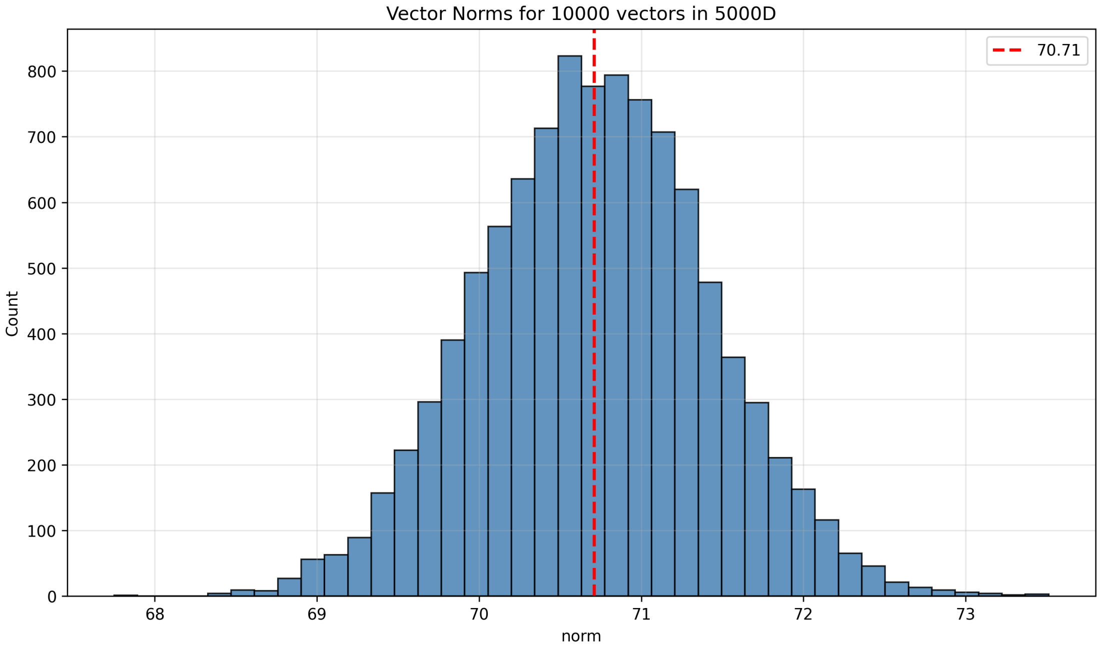
- Relative standard deviation, Relative contrast, Skewness and excess kurtosis (the graph goes up just a bit due to relatively small vectors sample of $10,000$):



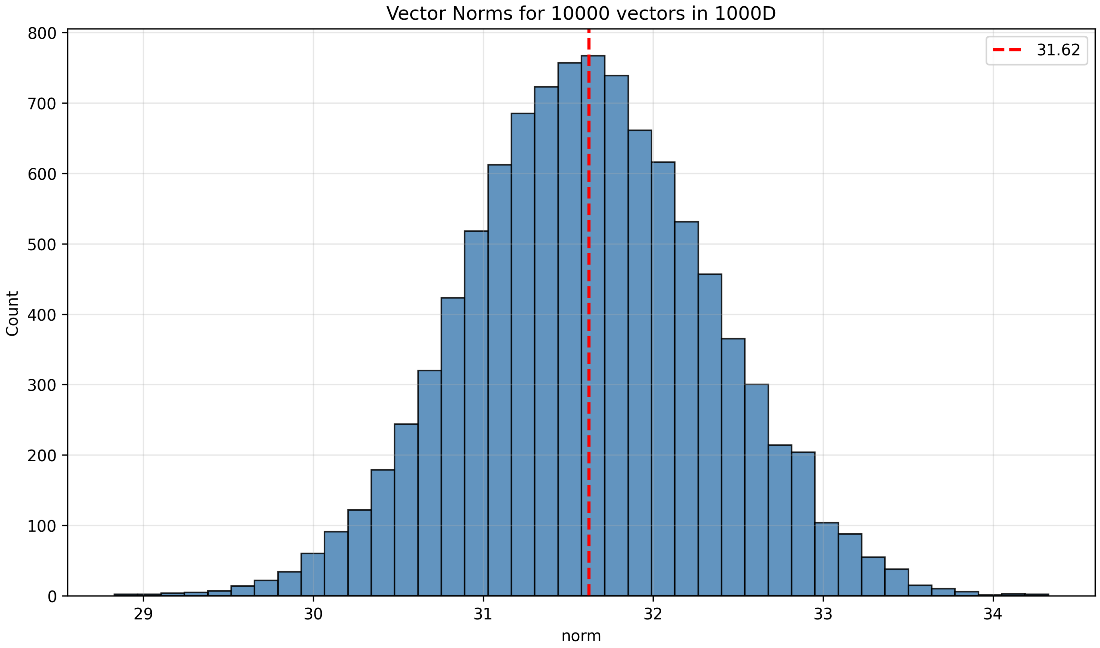
- Norms of the $10,000$ vectors in $10,000$ dimensions:



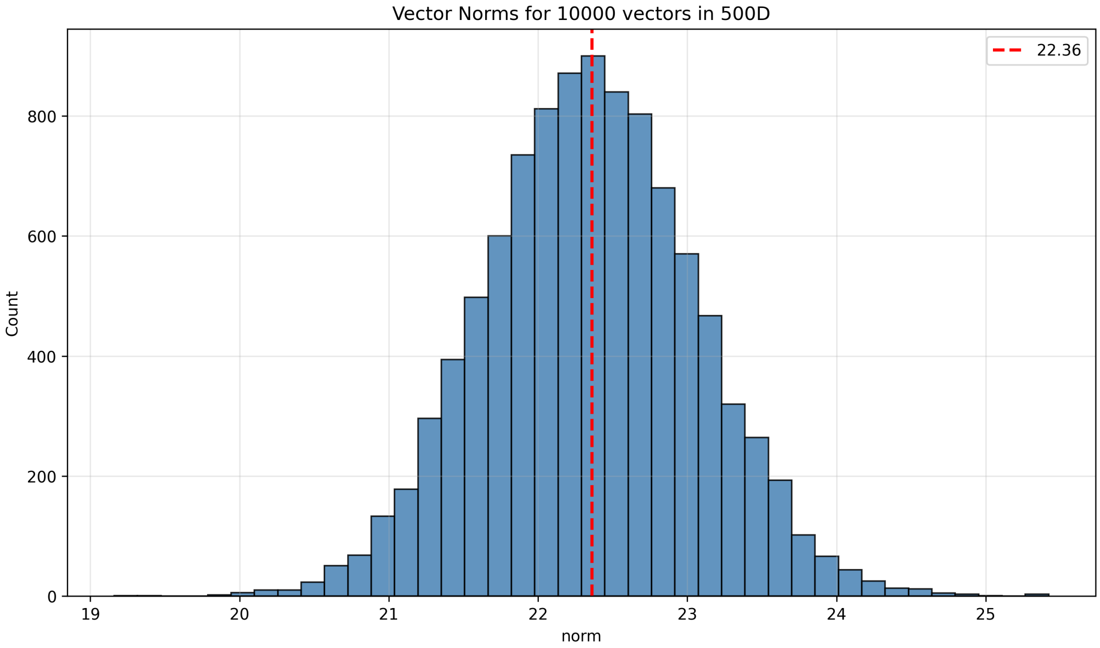
- Norms of the $10,000$ vectors in $5,000$ dimensions:



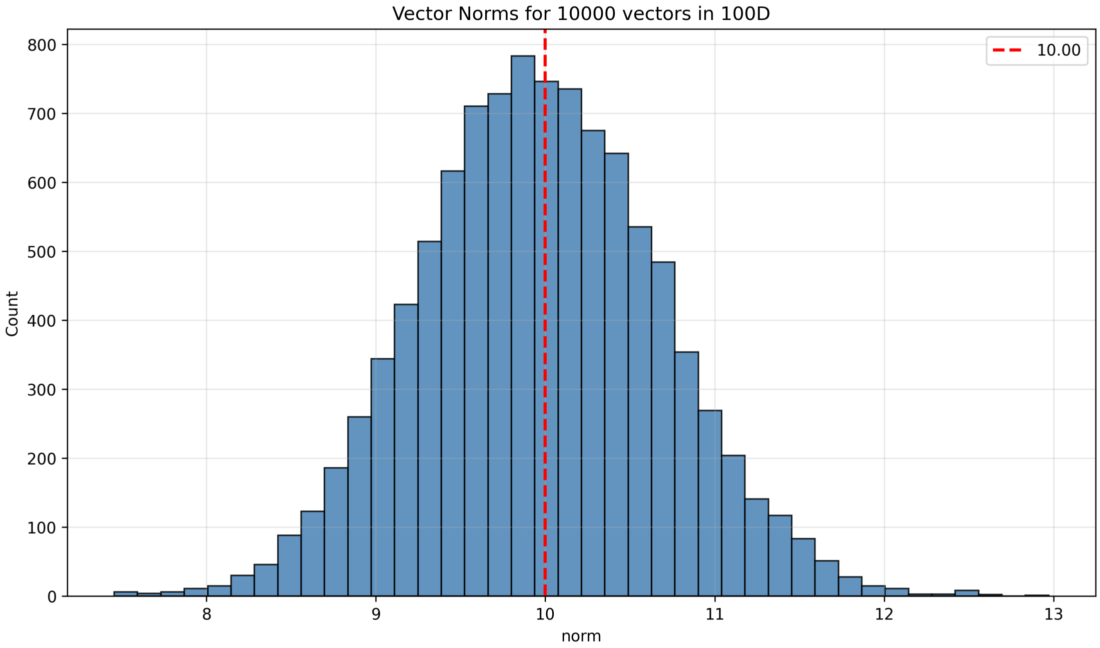
- Norms of the $10,000$ vectors in $1,000$ dimensions:



- Norms of the $10,000$ vectors in 500 dimensions:



- Norms of the $10,000$ vectors in 100 dimensions:



3 Dimensional Analysis of Convex Hulls: Empirical Inquiry

3.1 Overview

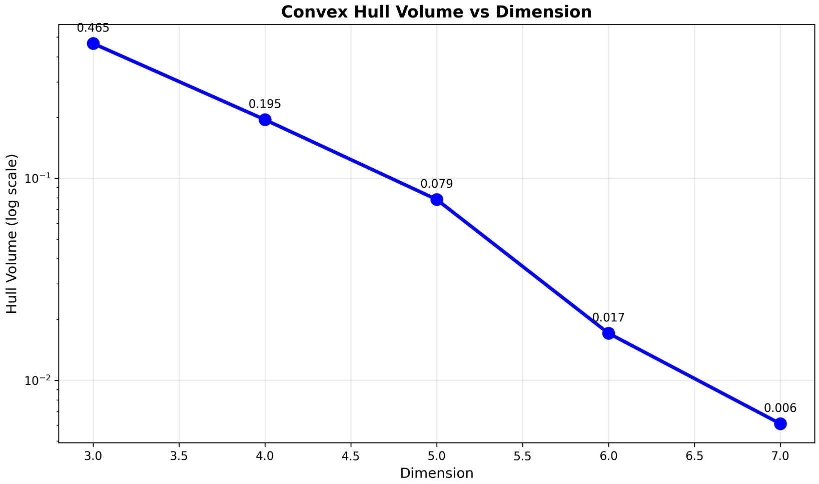
In this section we presents a sophisticated computational exploration into the properties of convex hulls formed from 25 randomly generated vectors across Euclidean dimensions three through seven. By constructing hulls and evaluating metrics such as volume, facet cardinality, and the ratio of boundary to interior points, we uncover dimensional trends: diminishing volumes, burgeoning facets, and an increasing dominance of boundary points. Implemented via Cursor, this study illuminates high-dimensional geometric phenomena, bridging theoretical insights with practical computation in fields like optimization and machine learning.

3.2 Methodology

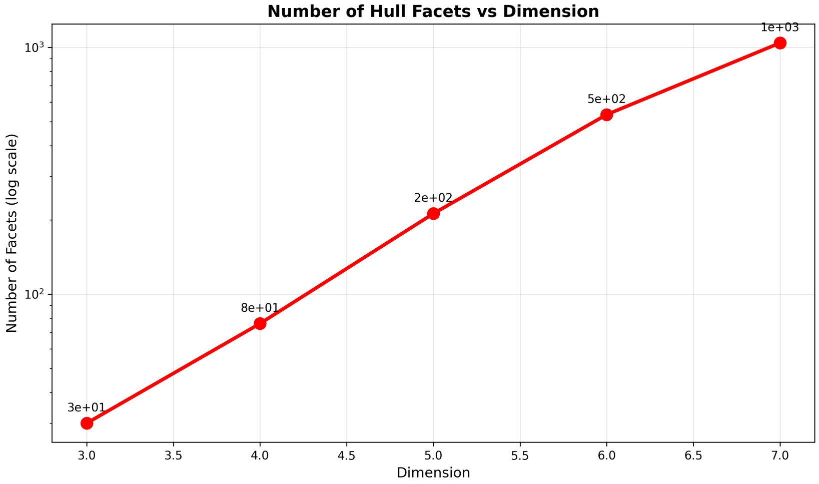
Utilizing Cursor, an AI-assisted coding environment, we generated 25 uniform random vectors in $[0,1]^d$ for $d = 3, 4, 5, 6, 7$. Convex hulls were computed, extracting volume, facet count (simplicial facets), and vertex proportion as boundary indicators. A single run ensured reproducibility, focusing on representative outcomes from general-position points.

3.3 Results

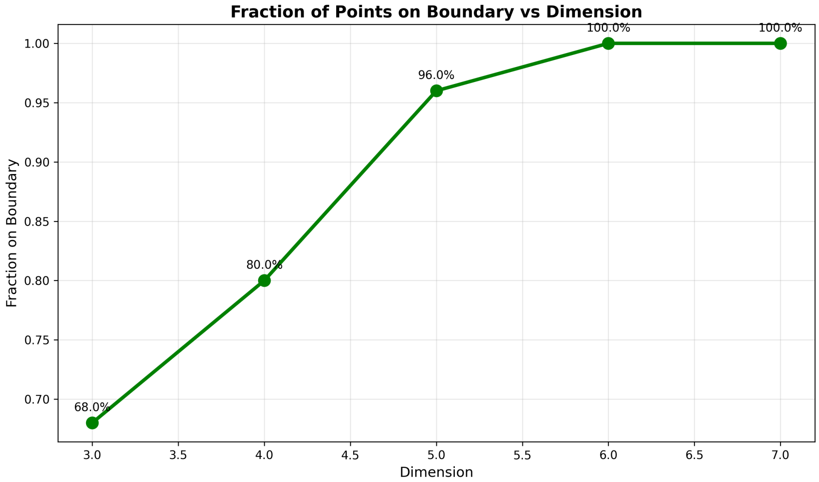
The following charts encapsulates the dimensional variations. The first compare the volume of the convex hull across different dimensions.



The second compare the facets of the convex hull across different dimensions:



The last chart compare the fraction of the points on the boundry of the convex hull across different dimensions:



3.4 Discussion

Volumes contract sharply with dimensionality, reflecting spatial sparsity, while facets proliferate combinatorially—consistent with the upper bound theorem. Boundary points predominate in higher dimensions, exemplifying the ”empty interior” paradigm. Limitations include uniform sampling's boundary bias; Gaussian alternatives may yield contrasts. These patterns affirm high-dimensional challenges and the utility of AI-assisted tools like Cursor for geometric scrutiny.

3.5 Conclusion

This succinct inquiry illuminates the evolutionary dynamics of convex hulls across escalating dimensions, manifesting a pronounced diminution in volume, an augmentation in facet proliferation, and a pervasive saturation of boundary points. The Curse of Dimensionality is unequivocally exemplified herein: as dimensionality ascends, geometric attributes undergo profound transformations, with volumetric density increasingly congregating proximate to peripheries and structural complexity surging in an exponential manner.