On the Explainability of Models with a Large Number of Features

August 20, 2025

Abstract

The challenge of explainability in machine learning models operating on high-dimensional data is profound. This paper argues that the root of this challenge lies in the counter-intuitive geometry of high-dimensional spaces, a phenomenon often termed the ”curse of dimensionality.” We contend that as the number of features (dimensions) grows, the volume of influence around each data point diminishes to a negligible amount. This leads to a critical conclusion: virtually every new data point presented to a model is a significant distance from any point in the training set, making any prediction an act of pure extrapolation, not interpolation.

We first provide a rigorous mathematical foundation by deriving the formula for the volume of an n-dimensional hypersphere (an n-ball), showing that its volume relative to a bounding hypercube tends to zero as the dimension increases. Using this, we formalize the concept of a vanishing ”volume of influence.” Second, we employ set theory to argue that the collective interpolation region formed by the training data is an infinitesimally small subset of the feature space, meaning new samples almost surely fall into the extrapolation domain. Finally, we discuss a topological perspective, considering the idea that high-dimensional spaces might ”collapse” or contain low-dimensional manifolds where interpolation could be conceptually recovered. This topological view offers a potential, albeit challenging, path to understanding how models can generalize at all. We conclude that traditional local, interpolative methods of explainability are foundationally flawed in high-dimensional settings and that new paradigms are required.

1 Introduction

The proliferation of machine learning has ushered in models of remarkable predictive power. However, this power often comes at the cost of transparency. For models trained on high-dimensional data, such as in genomics, image recognition, or natural language processing, the decision-making process is typically opaque, leading to the ”black box” problem. Explainable AI (XAI) seeks to pry open this box, yet many of its methods rely on a flawed assumption: that the model's behavior can be understood by analyzing its response to small perturbations of an input. This is fundamentally an act of interpolation, exploring a local neighborhood.

This paper asserts that in high-dimensional space, the concept of a ”local neighborhood” is almost always empty. The curse of dimensionality dictates that data points in high dimensions are almost always far apart and isolated. The space between them is a vast, uncharted desert. Therefore, when a model makes a prediction on a new data point, it is not interpolating between known examples but extrapolating into this void. This makes local explanations based on perturbations potentially misleading, as they explore a region where no data ever existed.

We will explore this problem from three angles:

1. **Geometric:** By deriving and analyzing the volume of a hypersphere.
2. **Set-Theoretic:** By formalizing the concept of an interpolation region.
3. **Topological:** By considering alternative structures of high-dimensional space.

Our goal is to demonstrate that the difficulty in explaining high-dimensional models is an inherent mathematical property of the spaces they operate in, not just a limitation of our current algorithms.

2 The Geometry of High-Dimensional Space

The foundation of our argument rests on the bizarre and counter-intuitive geometric properties of spaces with many dimensions. Here, we derive the essential formula for the volume of a hypersphere and discuss its consequences.

2.1 Derivation of the Hypersphere Volume

A hypersphere (or n-ball) is the set of points in $\mathbb{R}^n$ at a distance of at most R from a central point. Its volume, $V_n(R)$, is a fundamental quantity. We can derive a general formula for $V_n(R)$ using a powerful method involving the Gaussian integral.

Consider the n-dimensional Gaussian integral:

$$I_n = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-(x_1^2+x_2^2+\cdots+x_n^2)} dx_1 \cdots dx_n$$

This integral can be separated into a product of n identical one-dimensional integrals:

$$I_n = \left(\int_{-\infty}^{\infty} e^{-x^2} dx \right)^n$$

The value of the standard Gaussian integral is $\sqrt{\pi}$. Thus,

$$I_n = (\sqrt{\pi})^n = \pi^{n/2}$$

Now, we can solve the same integral I_n using hyperspherical coordinates. The volume element $dx_1 \cdots dx_n$ becomes $S_{n-1}(r)dr$, where $S_{n-1}(r)$ is the surface area of a hypersphere of radius r. The integral becomes:

$$I_n = \int_0^{\infty} e^{-r^2} S_{n-1}(r) dr$$

The surface area $S_{n-1}(r)$ is proportional to r^{n-1}, so we can write $S_{n-1}(r) = S_{n-1}(1) r^{n-1}$. Let $S_{n-1} = S_{n-1}(1)$ be the surface area of a unit hypersphere.

$$I_n = S_{n-1} \int_0^{\infty} r^{n-1} e^{-r^2} dr$$

We use the substitution $u = r^2$, so $du = 2rdr$, which gives $dr = \frac{du}{2\sqrt{u}}$.

$$\int_0^\infty r^{n-1} e^{-r^2} dr = \int_0^\infty u^{(n-1)/2} e^{-u} \frac{du}{2\sqrt{u}} = \frac{1}{2} \int_0^\infty u^{n/2-1} e^{-u} du$$

This integral defines the Gamma function, $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$. Therefore, the integral evaluates to $\frac{1}{2}\Gamma(\frac{n}{2})$. Substituting this back, we get:

$$I_n = S_{n-1} \cdot \frac{1}{2} \Gamma\left(\frac{n}{2}\right)$$

Equating our two expressions for I_n:

$$\pi^{n/2} = S_{n-1} \cdot \frac{1}{2} \Gamma\left(\frac{n}{2}\right) \implies S_{n-1} = \frac{2\pi^{n/2}}{\Gamma(\frac{n}{2})}$$

The volume of the hypersphere $V_n(R)$ can be found by integrating the surface area from 0 to R:

$$V_n(R) = \int_0^R S_{n-1}(r) dr = \int_0^R S_{n-1} r^{n-1} dr = S_{n-1} \frac{R^n}{n}$$

Substituting our result for S_{n-1}:

$$V_n(R) = \frac{2\pi^{n/2}}{n\Gamma(\frac{n}{2})} R^n$$

Using the identity $z\Gamma(z) = \Gamma(z+1)$, we have $n\Gamma(\frac{n}{2}) = 2\Gamma(\frac{n}{2}+1)$. This gives the final, standard form for the volume of an n-ball:

$$V_n(R) = \frac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)} R^n$$

2.2 The Vanishing Volume of Influence

The formula for the volume of an n-ball, $\mathcal{B}_n(R) = \{\mathbf{x} \in \mathbb{R}^n : ||\mathbf{x}||_2 \leq R\}$, has startling consequences that dismantle our low-dimensional intuition. Let us denote the volume, corresponding to the n-dimensional Lebesgue measure λ_n, as $V_n(R) = \lambda_n(\mathcal{B}_n(R))$. The two most critical phenomena are the asymptotic hollowness of space and the concentration of measure on an infinitesimally thin annulus.

2.2.1 Asymptotic Hollowness of Euclidean Space

First, consider a hypercube $\mathcal{C}_n = [-R, R]^n$ with volume $\lambda_n(\mathcal{C}_n) = (2R)^n$. The ratio of the volume of the inscribed n-ball $\mathcal{B}_n(R)$ to that of the hypercube is:

$$\mathcal{R}_n = \frac{V_n(R)}{V_{\text{cube}}(2R)} = \frac{\pi^{n/2} R^n}{(2R)^n \Gamma(\frac{n}{2}+1)} = \frac{\pi^{n/2}}{2^n \Gamma(\frac{n}{2}+1)}$$

To understand the asymptotic behavior as $n \to \infty$, we employ Stirling's approximation for the Gamma function, where $\Gamma(z+1) \approx \sqrt{2\pi z}(\frac{z}{e})^z$. Let $k = n/2$:

$$\Gamma\left(\frac{n}{2}+1\right) = \Gamma(k+1) \approx \sqrt{2\pi k} \left(\frac{k}{e}\right)^k = \sqrt{\pi n} \left(\frac{n}{2e}\right)^{n/2}$$

Substituting this into our ratio $\mathcal{R}_n$:

$$\mathcal{R}_n \approx \frac{\pi^{n/2}}{2^n\sqrt{\pi n}(\frac{n}{2e})^{n/2}} = \frac{\pi^{n/2}}{2^n\sqrt{\pi n}}\frac{(2e)^{n/2}}{n^{n/2}} = \frac{(\pi 2e)^{n/2}}{2^n\sqrt{\pi n}\ n^{n/2}} = \frac{1}{\sqrt{\pi n}}\left(\frac{2\pi e}{4n}\right)^{n/2} = \frac{1}{\sqrt{\pi n}}\left(\frac{\pi e}{2n}\right)^{n/2}$$

As $n \to \infty$, the term $\left(\frac{\pi e}{2n}\right)^{n/2}$ decays super-exponentially to zero, as $\pi e/2$ is a constant while n grows. Thus, we formally see that:

$$\lim_{n\to\infty} \mathcal{R}_n = 0$$

This implies that for large n, nearly all the volume of a hypercube resides in its "corners," outside the largest possible inscribed hypersphere. The space becomes fundamentally hollow.

2.2.2 Concentration on the Annulus

Second, and more profoundly, the internal structure of the n-ball itself becomes hollow. Let us define an outer annulus (or shell) of the n-ball $\mathcal{B}_n(R)$ as $\mathcal{A}_n(R,\epsilon) = \mathcal{B}_n(R) \setminus \mathcal{B}_n(R(1-\epsilon))$ for some small $\epsilon \in (0,1)$. The fraction of the n-ball's volume contained within this thin shell is:

$$\frac{\lambda_n(\mathcal{A}_n(R,\epsilon))}{\lambda_n(\mathcal{B}_n(R))} = \frac{V_n(R) - V_n(R(1-\epsilon))}{V_n(R)} = \frac{\frac{\pi^{n/2}R^n}{\Gamma(\frac{n}{2}+1)} - \frac{\pi^{n/2}(R(1-\epsilon))^n}{\Gamma(\frac{n}{2}+1)}}{\frac{\pi^{n/2}R^n}{\Gamma(\frac{n}{2}+1)}} = \frac{R^n - R^n(1-\epsilon)^n}{R^n} = 1-(1-\epsilon)^n$$

For any fixed $\epsilon > 0$, no matter how small, the limit is unambiguous:

$$\lim_{n\to\infty} [1-(1-\epsilon)^n] = 1$$

For example, in $n = 1000$ dimensions, the fraction of volume in the outermost shell of thickness $\epsilon = 0.01$ (i.e., the outer 1% of the radius) is $1-(0.99)^{1000} \approx 1 - 4.3\times 10^{-5}$, or over 99.995%.

To formalize this, consider a random vector $\mathbf{X}$ uniformly distributed in $\mathcal{B}_n(R)$. Let its radial coordinate be the random variable $\rho = ||\mathbf{X}||_2$. The cumulative distribution function (CDF) of ρ is given by:

$$F_\rho(r) \equiv P(\rho \le r) = \frac{V_n(r)}{V_n(R)} = \left(\frac{r}{R}\right)^n, \quad \text{for } 0 \le r \le R$$

The probability density function (PDF) is therefore:

$$f_\rho(r) = \frac{dF_\rho}{dr} = \frac{nr^{n-1}}{R^n}$$

The mode of this distribution, which represents the most probable radius, is found by maximizing $f_\rho(r)$, which clearly occurs at $r = R$. The median radius, r_m, is the radius such that $F_\rho(r_m) = 0.5$:

$$\left(\frac{r_m}{R}\right)^n = \frac{1}{2} \implies r_m = R\cdot(0.5)^{1/n} = R\cdot e^{-\frac{\ln 2}{n}}$$

Using the Taylor expansion $e^{-x} \approx 1 - x$ for small x, we see that for large n:

$$r_m \approx R\left(1 - \frac{\ln 2}{n}\right)$$

This demonstrates that the median radius approaches the maximum radius R as n increases. Half of the n-ball's mass lies in a shell of thickness merely $R\frac{\ln 2}{n}$.

Thus, any "volume of influence" we might define around a training point $\mathbf{x}_i$, such as a small n-ball $\mathcal{B}_n(\mathbf{x}_i, r)$, is a paradoxical object. Not only is its total volume $\lambda_n(\mathcal{B}_n(\mathbf{x}_i, r))$ vanishingly small relative to the surrounding space, but its own volume is concentrated on its boundary $\partial\mathcal{B}_n$. This means that even if a new point were to fall within this supposed volume of influence, it is almost certain to be near its boundary, at a distance close to r from $\mathbf{x}_i$. The notion of a cozy, dense, truly local neighborhood completely evaporates.

3 A Set-Theoretic Perspective on Extrapolation

The geometric phenomena described in the previous section have a profound consequence that can be formalized using the language of measure theory and set theory. We can define the domain of "interpolation" and demonstrate that, in high dimensions, this domain constitutes a set of measure zero within the broader feature space.

Definition 1 (Training Set and Sample Space)**.** *Let the training dataset be a finite set of M points $\mathcal{X}_{train} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M\}$ embedded in an n-dimensional Euclidean space, $\mathcal{X}_{train} \subset \mathbb{R}^n$. We consider new data points $\mathbf{x}_{new}$ to be drawn from a probability distribution $\mathbb{P}$ over a bounded sample space $\Omega \subset \mathbb{R}^n$ that contains $\mathcal{X}_{train}$. For simplicity, we assume a uniform distribution, where the probability of drawing a point from any subset $A \subseteq \Omega$ is proportional to its n-dimensional Lebesgue measure, $\lambda_n(A)$.*

We can define the region of interpolation in two primary ways: a conservative definition based on local neighborhoods and a more generous one based on the convex hull of the training data.

Definition 2 (Neighborhood Interpolation Region)**.** *Let the **volume of influence** for each training point $\mathbf{x}_i$ be an n-ball of a small fixed radius $r > 0$, denoted $\mathcal{B}_n(\mathbf{x}_i, r)$. The **neighborhood interpolation region**, $\mathcal{I}_r$, is the union of these volumes:*

$$\mathcal{I}_r = \bigcup_{i=1}^{M} \mathcal{B}_n(\mathbf{x}_i, r)$$

Any point $\mathbf{x}_{new} \in \mathcal{I}_r$ is considered an interpolation by proximity.

Theorem 1. *As the dimension $n \to \infty$, the probability that a new point $\mathbf{x}_{new}$ falls within the neighborhood interpolation region $\mathcal{I}_r$ tends to zero.*

$$\lim_{n\to\infty} \mathbb{P}(\mathbf{x}_{new} \in \mathcal{I}_r) = 0$$

Proof. The n-dimensional volume of the region $\mathcal{I}_r$ can be bounded using Boole's inequality (the union bound for measures):

$$\lambda_n(\mathcal{I}_r) = \lambda_n\left(\bigcup_{i=1}^{M} \mathcal{B}_n(\mathbf{x}_i, r)\right) \leq \sum_{i=1}^{M} \lambda_n(\mathcal{B}_n(\mathbf{x}_i, r))$$

Since $\lambda_n(\mathcal{B}_n(\mathbf{x}_i, r)) = V_n(r)$ for all i, this simplifies to:

$$\lambda_n(\mathcal{I}_r) \leq M \cdot V_n(r) = M \cdot \frac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)} r^n$$

As established in Section 2.2, the volume of an n-ball of fixed radius r vanishes as the dimension grows: $\lim_{n\to\infty} V_n(r) = 0$. Consequently, the total volume of the interpolation region also vanishes:

$$\lim_{n\to\infty} \lambda_n(\mathcal{I}_r) \leq \lim_{n\to\infty} M \cdot V_n(r) = 0$$

The probability of a new point, drawn uniformly from a containing volume Ω with $\lambda_n(\Omega) > 0$, falling into $\mathcal{I}_r$ is:

$$\mathbb{P}(\mathbf{x}_{\text{new}} \in \mathcal{I}_r) = \frac{\lambda_n(\mathcal{I}_r)}{\lambda_n(\Omega)}$$

As the numerator tends to zero while the denominator remains positive, the probability itself must tend to zero. Therefore, almost every new point is an extrapolation, i.e., $\mathbf{x}_{\text{new}} \in \Omega \setminus \mathcal{I}_r$. □

One might argue this definition is too strict. A more generous definition of interpolation considers any point that can be expressed as a convex combination of the training points.

Definition 3 (Convex Hull Interpolation Region)**.** *The **convex hull** of the training set, denoted $\mathcal{I}_{conv}$, is the set of all convex combinations of its points:*

$$\mathcal{I}_{conv} = conv(\mathcal{X}_{train}) = \left\{ \sum_{i=1}^{M} \alpha_i \mathbf{x}_i \mid (\forall i, \alpha_i \geq 0) \wedge \sum_{i=1}^{M} \alpha_i = 1 \right\}$$

This region represents the tightest convex set enclosing all training data. Yet, even under this much broader definition, the conclusion remains the same. The volume of a "flat" polytope defined by a small number of vertices M in a high-dimensional space $n \gg M$ is vanishingly small. Research in high-dimensional probability and geometry confirms this.

Theorem 2 (Balestriero et al., 2021, informal)**.** *For a fixed number of training points M, as the dimension $n \to \infty$, the probability that a new point $\mathbf{x}_{new}$ falls within the convex hull of the training data tends to zero.*

$$\lim_{n\to\infty} \mathbb{P}(\mathbf{x}_{new} \in \mathcal{I}_{conv}) = 0$$

This demonstrates that the problem is fundamental. It is not an artifact of defining interpolation regions via small radii. Any reasonable definition of a region "between" the data points results in a set whose volume, and therefore its probability measure, collapses to zero as the dimension grows. The feature space Ω becomes almost entirely an **extrapolation domain**. For any new sample, we are virtually guaranteed that $\mathbf{x}_{\text{new}} \notin \text{conv}(\mathcal{X}_{\text{train}})$, forcing the model to make predictions in a region for which it has no direct evidence from a convex combination of its training data.

4 Topological Structure and Intrinsic Dimensionality

The preceding analysis presents a profound paradox. The geometry of high-dimensional space dictates that our datasets are infinitesimally sparse, making every new prediction an act of pure extrapolation into a void (Section 2). From a set-theoretic view, the collective region of interpolation is a set of measure zero (Section 3). This unforgiving picture would logically imply that generalization is impossible and machine learning should fail. Yet, it demonstrably does not.

The resolution to this paradox lies in a fundamental shift in perspective: from the properties of the ambient space $\mathbb{R}^n$ to the *intrinsic structure* of the data itself. The implicit assumption that data points are scattered randomly throughout their high-dimensional container is flawed. This section argues that the success of machine learning is predicated on two related phenomena: the statistical stability afforded by measure concentration and, most critically, the fact that real-world data resides on or near a low-dimensional manifold.

4.1 The Statistical Collapse

While Euclidean distances explode in high dimensions, they do so in a surprisingly predictable way. This is due to the phenomenon of **measure concentration**. For a broad class of high-dimensional probability measures (e.g., Gaussian, or uniform on a sphere), Lipschitz functions of a random variable are sharply concentrated around their mean.

Let $\mathbf{X}$ be a random vector in $\mathbb{R}^n$ with independent sub-gaussian components. A key result, often termed Lévy's Lemma, states that for any 1-Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$, there exist constants $c_1, c_2 > 0$ such that:

$$\mathbb{P}(|f(\mathbf{X}) - \mathbb{E}[f(\mathbf{X})]| \geq t) \leq c_1 e^{-c_2 t^2} \quad \forall t \geq 0$$

The Euclidean distance function is Lipschitz. Consider two independent random points $\mathbf{X}, \mathbf{Y}$ drawn from such a distribution. Their squared distance $d_E^2(\mathbf{X}, \mathbf{Y}) = ||\mathbf{X} - \mathbf{Y}||_2^2$ is a sum of n i.i.d. random variables, and by the law of large numbers, it concentrates sharply around its mean. The distribution of pairwise distances becomes a narrow Gaussian-like peak.

This has a crucial implication: while all points are far from each other in an absolute sense, the *relative* distance structure is highly stable. The variance of pairwise distances is small compared to its mean. This statistical "collapse" provides a stable, non-local signal that algorithms can exploit. Models may not learn from "local neighborhoods" in the traditional sense, but from this highly structured global geometry of pairwise distances.

4.2 The Manifold Hypothesis

A far more powerful explanation for generalization is the **Manifold Hypothesis**. It posits that high-dimensional data observed in the real world is not arbitrarily distributed in $\mathbb{R}^n$, but lies on or near a low-dimensional topological manifold $\mathcal{M}$ embedded within $\mathbb{R}^n$.

Definition 4 (Embedded Manifold). *An m-dimensional manifold $\mathcal{M}$ is a topological space that is locally homeomorphic to the Euclidean space $\mathbb{R}^m$. It is embedded in the* *ambient space $\mathbb{R}^n$ via a mapping $\psi : \mathcal{M} \to \mathbb{R}^n$, where the intrinsic dimension m is typically much smaller than the ambient dimension n (i.e., $m \ll n$).*

For example, the set of all 1024×1024 pixel images forms an ambient space $\mathbb{R}^{1,048,576}$. However, the subset of images corresponding to human faces that can be described by parameters like age, expression, head pose, and lighting lies on a manifold with an intrinsic dimension of perhaps only $m \approx 50$.

The distinction between distance in the ambient space versus distance on the manifold is critical.

- **Euclidean Distance** $d_E(\mathbf{x}, \mathbf{y})$: The straight-line distance $||\mathbf{x} - \mathbf{y}||_2$ in $\mathbb{R}^n$. This is the "chord distance" that ignores the structure of the manifold.
- **Geodesic Distance** $d_{\mathcal{M}}(\mathbf{x}, \mathbf{y})$: The length of the shortest path between points $\mathbf{x}$ and $\mathbf{y}$ while remaining entirely on the surface of $\mathcal{M}$.

It is entirely possible for $d_E(\mathbf{x}, \mathbf{y})$ to be large while $d_{\mathcal{M}}(\mathbf{x}, \mathbf{y})$ is small. This single fact resolves the "island" paradox. Data points are not isolated; they are connected by paths along the manifold. The vast "empty space" is irrelevant because it is "off-manifold" and does not support the data-generating distribution.

A successful machine learning model, therefore, is not learning a function on $\mathbb{R}^n$. It is implicitly performing a two-stage process:

1. **Manifold Approximation**: It learns a mapping $\Phi : \mathbb{R}^n \to \mathcal{Z}$, where $\mathcal{Z} \subset \mathbb{R}^m$ is a low-dimensional latent space that effectively parameterizes the manifold $\mathcal{M}$. This is the role of encoders in autoencoders or the latent space in generative models.
2. **Function Interpolation on the Manifold**: It then learns a much simpler function $f : \mathcal{Z} \to Y$ from this well-behaved latent space to the output. In this low-dimensional space, neighborhoods are dense, volumes are not vanishing, and interpolation is once again a valid and meaningful concept.

The "collapse" is not a property of $\mathbb{R}^n$ but a learned projection of the data onto its intrinsic, low-dimensional manifold representation $\mathcal{Z}$. Techniques from Topological Data Analysis (TDA), such as using persistent homology to identify the Betti numbers (counts of connected components, loops, voids, etc.) of a point cloud, provide concrete tools to empirically validate the shape and dimensionality of $\mathcal{M}$, lending strong support to this hypothesis.

5 From Ambient Perturbations to Manifold Traversal

The reconciliation of the extrapolation paradox through the manifold hypothesis forces a complete re-evaluation of what constitutes a valid explanation for a model's prediction. The geometric and topological realities render explanations based on ambient space perturbations fundamentally unsound, while simultaneously illuminating a new, more robust path forward.

5.1 The Fundamental Invalidity of Ambient-Space Explanations

Local explanation methods like LIME and SHAP operate by sampling a neighborhood $\mathcal{N}(\mathbf{x}_0) = \{\mathbf{x}_0 + \boldsymbol{\delta} : ||\boldsymbol{\delta}||_2 \leq \epsilon\}$ around a point of interest $\mathbf{x}_0$ and fitting a simple, interpretable model to the predictor's outputs in this region. The feature importance scores ϕ_i are derived from this local approximation.

The profound flaw in this approach is the near-certainty of a **domain mismatch**. As established, the data-generating distribution is supported on a low-dimensional manifold $\mathcal{M}$. The perturbation sampling, however, is performed in the ambient space $\mathbb{R}^n$. The intersection of the perturbation ball with the data manifold, $\mathcal{N}(\mathbf{x}_0) \cap \mathcal{M}$, is a set of measure zero within the ball itself. As such, these methods build their explanations using exclusively out-of-distribution (OOD) samples that are off-manifold.

Formally, the explanation is conditioned on an event of measure zero with respect to the true data-generating probability measure $\mathbb{P}_{\mathcal{M}}$:

$$\mathbb{P}_{\mathcal{M}}(\text{sample} \in \mathcal{N}(\mathbf{x}_0) \setminus \mathcal{M}) \approx 1$$

The resulting feature attributions ϕ_i do not explain the model's behavior on the data distribution it was trained on. Instead, they explain the model's behavior in the meaningless void between manifold sheets. This is not just a quantitative error; it is a fundamental misrepresentation of the model's reasoning.

5.2 Manifold-Based Explainability

A valid explanation must respect the intrinsic geometry of the data. This requires shifting the goal of XAI from attributing features at an isolated point to understanding the model's behavior along paths and within regions *on the learned manifold.*

Geodesic Paths over Linear Interpolation. Instead of asking "how does the prediction change along the straight line from $\mathbf{x}_A$ to $\mathbf{x}_B$?", which immediately leaves the manifold, we should ask "how does the prediction change along the geodesic path from $\mathbf{x}_A$ to $\mathbf{x}_B$ on $\mathcal{M}$?". For models with an explicit latent space $\mathcal{Z}$ (e.g., VAEs, GANs), this can be approximated by first encoding the endpoints to get $\mathbf{z}_A = \Phi(\mathbf{x}_A)$ and $\mathbf{z}_B = \Phi(\mathbf{x}_B)$, then performing linear interpolation in the latent space $\mathbf{z}(t) = (1-t)\mathbf{z}_A + t\mathbf{z}_B$, and finally decoding the path back to the manifold: $\mathbf{x}(t) = \text{Decoder}(\mathbf{z}(t))$. This traces a path of high data density.

On-Manifold Counterfactuals. A meaningful counterfactual explanation seeks the "minimal change" to an input to alter its prediction. The geometric perspective shows that minimizing the Euclidean norm $||\Delta\mathbf{x}||_2$ is flawed, as it may require moving to a point far away geodesically. A robust counterfactual is a point $\mathbf{x}_{cf}$ that minimizes the geodesic distance on the learned manifold:

$$\mathbf{x}_{cf} = \arg\min_{\mathbf{x}' \in \mathcal{M}} d_{\mathcal{M}}(\mathbf{x}, \mathbf{x}') \quad \text{subject to} \quad \text{Model}(\mathbf{x}') \neq \text{Model}(\mathbf{x})$$

This finds the "closest" example on the manifold that belongs to a different class, providing a truly meaningful and plausible explanation.

Latent Space Gradients. The final linchpin of this new paradigm is to analyze model behavior in the coordinate system of the manifold itself—the latent space. The gradient of the prediction function F with respect to the ambient input, $\nabla_{\mathbf{x}} F(\mathbf{x})$, is a high-dimensional vector that is difficult to interpret and is confounded by the embedding's geometry. In contrast, the gradient with respect to the latent variables, $\nabla_{\mathbf{z}} f(\mathbf{z})$, where $F(\mathbf{x}) = f(\Phi(\mathbf{x}))$, reveals the model's sensitivity to changes along the principal axes of data variation. This low-dimensional vector tells us which semantic features (as captured by the latent dimensions) are most influential for a given prediction. This is the ultimate goal of explainability: to understand the model's logic in the language of the data's core concepts.

6 Conclusion

The challenge of explaining models with a large number of features is a direct and unavoidable consequence of the geometry of high-dimensional space. The "curse of dimensionality" ensures that the volume of influence around any data point is negligible, making the space between points a vast desert. As we have shown through geometric derivation and set-theoretic reasoning, any prediction for a new point is an act of pure extrapolation.

This fact fundamentally invalidates the premise of many current explainability methods that rely on local interpolation. However, the success of high-dimensional models suggests that they are not merely making wild guesses. The topological perspective, particularly the manifold hypothesis, offers a compelling resolution: models may be implicitly discovering low-dimensional structures where the data is concentrated and where the notion of interpolation can be restored.

Therefore, the future of explainability for high-dimensional models does not lie in refining local perturbation methods. It lies in developing new techniques capable of extracting, visualizing, and interpreting the underlying manifolds that models learn. Only by understanding the landscape of this learned space can we hope to provide meaningful explanations for why a model makes a particular decision.

References

[1] R. Balestriero, J. Pesenti, and Y. LeCun, "Learning in High Dimension Always Amounts to Extrapolation," *arXiv preprint arXiv:2110.09485*, 2021.

[2] R. E. Bellman, *Adaptive Control Processes: A Guided Tour.* Princeton University Press, 1961.

[3] W. B. Johnson and J. Lindenstrauss, "Extensions of Lipschitz mappings into a Hilbert space," *Contemporary Mathematics*, vol. 26, pp. 189–206, 1984.

[4] G. Singh, F. Mémoli, and G. Carlsson, "Topological methods for the analysis of high dimensional data sets and 3d object recognition," in *SPBG*, 2007.