Token Embeddings Violate the Manifold Hypothesis

Michael Robinson[1], Sourya Dey[2], Tony Chiang[3]

[1]Mathematics and Statistics, American University, Washington, DC, USA, michaelr@american.edu
[2]Galois, Inc., Arlington, VA, USA, sourya@galois.com
[3]Department of Mathematics, University of Washington, Seattle, WA, chiang@math.washington.edu

Abstract

To fully understand the behavior of a large language model (LLM) requires our understanding of its input space. If this input space differs from our assumption, our understanding of and conclusions about the LLM is likely flawed, regardless of its architecture. Here, we elucidate the structure of the token embeddings, the input domain for LLMs, both empirically and theoretically. We present a generalized and statistically testable model where the neighborhood of each token splits into well-defined signal and noise dimensions. This model is based on a generalization of a manifold called a *fiber bundle*, so we denote our hypothesis test as the "fiber bundle null." Failing to reject the null is uninformative, but rejecting it at a specific token indicates that token has a statistically significant local structure, and so is of interest to us. By running our test over several open-source LLMs, each with unique token embeddings, we find that the null is frequently rejected, and so the token subspace is provably not a fiber bundle and hence also not a manifold. As a consequence of our findings, when an LLM is presented with two semantically equivalent prompts, and if one prompt contains a token implicated by our test, that prompt will likely exhibit more output variability proportional to the local signal dimension of the token.

1 Introduction

Large language models (LLMs) produce a response to a given query, by using a deep neural network to predict the next token given a window of previous tokens. How interchangeable are these tokens? From a linguistic perspective, those tokens that can be exchanged without impacting the meaning of a statement should be considered *synonyms*. Some tokens have more synonyms, whereas others have fewer. Those with fewer synonyms tend to be syntactically essential:

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited).

if you swap such a token for another, the resulting sentence is not likely to occur. Conversely, tokens with many synonyms are likely to be viewed as being interchangeable.

Logically prior to understanding the syntax learned by an LLM is the understanding of its *token subspace*, the internal representation of *individual* tokens (not sequences of tokens in context). Numerous papers have pointed to unexpected behaviors exhibited by LLMs that hinge on subtle changes in wording and text layout between apparently similar prompts, suggesting that certain—apparently semantically similar—tokens have dramatically different neighborhoods in the token subspace (for instance, see [1]). These differences in neighborhoods correspond to places where the token subspace is not a manifold; it is *singular* at such a token. Linguistically, singularities may correspond to *polysemy* or *homonyms*—tokens with multiple distinct meanings [2].

If the token subspace is singular, then these singularities can persist into the output of the LLM, perhaps unavoidably and regardless of its architecture. Not accounting for singularities in the token subspace may thereby impede the understanding of the LLM's behavior. Suppose the LLM is presented with two similar prompts, but one prompt has a token that is near the singularity. The prompt with a token near the singularity will likely exhibit more variability if both prompts are changed in the same way, depending on how well the transformer can resolve the singularity.

We present a test that determines whether the neighborhood of a given token contains a singularity. The test works by identifying changes in subspace dimension that are inconsistent with the token subspace being a *fiber bundle*, which is a strict generalization of a manifold. When our model finds a singularity at a token, this implies that the token has far fewer synonyms than its neighbors. In a context where such a token is used, its use in that role is syntactically essential, indicating that it plays an outsized role in the LLM.

We applied our test to four open source LLMs' token subspaces (GPT2 [3], Llemma7B [4], Mistral7B [5], and Pythia6.9B [6]). In each LLM we tested, we found that *the token subspace is not a manifold, because it is also not a fiber bundle*. Moreover, we observe highly statistically significant differences in the singular tokens *between* LLMs—even for those with identical sets of tokens overall—which indicates that their respective training methodologies have a strong impact on the token subspace. Under this situation, none of these LLMs should be expected to have similar responses to a prompt involving any of these singular tokens [7].

1.1 Background

At an abstract but precise level, an LLM consists of several interacting processes, as outlined in Figure 1. An LLM implements a transformation of a sequence of tokens (the query) into a new sequence of tokens (the response). Formally, if each input token is an element of a metric space T, then the LLM is a transformation $T^n \to T^m$, where n is the number of tokens in the query and m is the number of tokens in the response. This transformation is typically

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited).

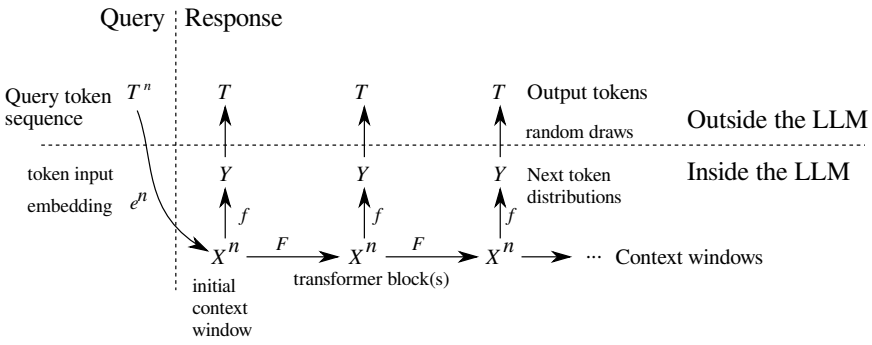

Figure 1: Data flow in a typical LLM. A sequence of tokens forming the query is converted via the token input embedding e^n into the initial context window, as a point in the latent space X^n. Each of these windows in the latent space are converted, token-by-token, into probability distributions via f into the single token latent space X. From these, each token presented in the output (in the set Y) is obtained via a random draw. These output tokens are then used for subsequent windows.

not a function because it is stochastic—it involves random draws.

To operate upon tokens using numerical models, such as could be implemented using neural networks, we must transform the finite set of tokens T into numerical data. This is typically done by way of a pair of *latent spaces* $X = \mathbb{R}^d$ and $Y = \mathbb{R}^q$. The dimension q of Y is chosen to be equal to the number of elements in T, so that elements of Y have the interpretation of being (unnormalized) probability distributions over T.

The transformation $T^n \to T^m$ is constructed in several stages.

Input tokenization : Each token is embedded individually via the *token input embedding* function $e : T \to X$. As a whole, X^n is called a *latent window*.

Transformer blocks : The probability distribution for the next token is constructed by a continuous function $f : X^n \to Y$. This is usually implemented by one or more *transformer blocks*.

Output tokenization : Given the output of one of the transformer blocks f, one can obtain an output token in T by a random draw. Specifically, if $(x_1, x_2, \ldots, x_n)$ is the current window in X^n, then the next token t is drawn from the distribution given by $f(x_1, x_2, \ldots, x_n)$.

Next window prediction : Given that token t was drawn from the distribution, the next latent window itself is constructed by a transformation $F : X^n \to X^n$, which advances the window as follows:

$$F(x_1, x_2, \ldots, x_n) := (x_2, \ldots, x_n, t).$$

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited).

The focus of this paper is specifically upon the structure of the *token input embedding* $e : T \to X = \mathbb{R}^d$. Since the token set T is finite, e can be stored as a matrix. In this matrix, each column corresponds to an element of T, and thereby ascribes a vector of numerical coordinates to each token. By replacing the last layer of the deep neural network $f : X^n \to Y$, a vector of probabilities for the next token is obtained from the activations of the last layer. One can therefore interpret the probabilities as specifying a *token output embedding*. Both the tokenization and the transformer stages are learned during training, and many strategies for this learning process are discussed extensively in the literature. Although their training is usually performed separately, these two stages interact when they produce the LLM output, so it is important to understand the lineage of a given tokenization as being from a particular LLM. We emphasize that *only the input tokenization is discussed in this article*.

The token input embedding matrix itself is interesting, as it defines "where" the tokens are located. It is reasonable to consider the tokens as being sampled from a larger latent subspace within the space of all possible activations. Such a space is quite unconstrained. There is no *a priori* reason to suspect it is a manifold, for instance. It has already been shown that local neighborhoods of each token have salient topological structure [8]. One of the most basic parameters is the *dimension* near any given token in this space. Higher dimension at a token means that the token has more near-neighbors—more synonyms—while lower dimensional tokens are less interchangeable [2].

Dimension is a manifestly local property. However, for manifolds, dimension is locally constant, hence global. It is for this reason that manifold learning is popular. If one computes PCA locally for a random sampling of a manifold embedded within Euclidean space, most of the variance in the data is captured within a few principal directions, namely those tangent to the manifold. In essence, these represent the signal within the data. The number of these directions is the dimension of the manifold, and this is a constant over (each connected component of) the manifold. The remaining directions, which are not tangent to the manifold, represent noise. The basic assumption is that transformers act on the entire input space, and that (clearly) is a manifold, because it is Euclidean space. But the truth is that a transformer *in the context of an LLM* really only acts on the *token subspace*, the image of the token input embedding $e : T \to X$, which is a subspace of that Euclidean space. That the token subspace is not a linear subspace is widely acknowledged, but more problematic is that *it is not a manifold* [9].

Assuming that word embeddings yield manifolds, some researchers have used global dimension estimators on token input embeddings and word embeddings [10, 11]. *A priori* one should not suspect that a set of tokens (or other samples) lies on a manifold. Although there are rigorous statistical tests for manifolds [12], they are arduous to apply in practice.

By using a local (not global) dimension estimator, [9] presented the first (to our knowledge) direct test of whether the token subspace is a manifold for the token input embeddings for several LLMs. A strongly negative result was obtained: the subspace of tokens is apparently never a manifold, so global

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited).

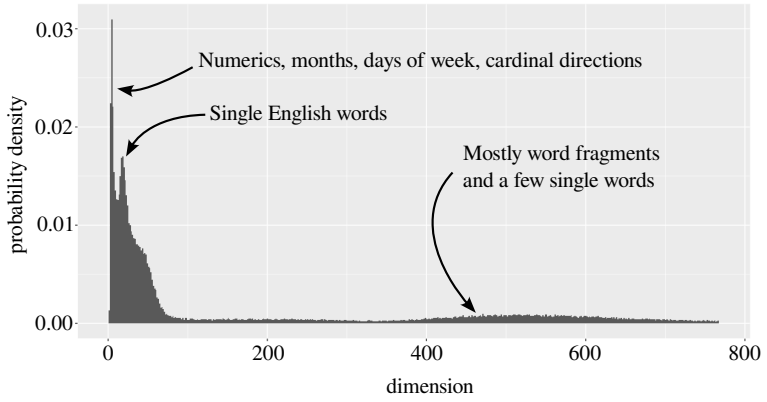

Figure 2: The distribution of local dimensions estimated near tokens in GPT2, from [9].

dimension estimators are not reliable. Figure 2 shows the distribution of dimensions they obtained for GPT2 (March 11, 2024 version) [3]. Recall that dimension correlates with the number of free parameters one can perturb a point and still stay within the space, and that for manifolds this number is a constant. The highly multi-modal nature of the distribution is a reflection of the inherent non-manifold structure of the token subspace.

There are several clusters of low dimensional tokens, which accord with the low dimensions obtained by others using global estimators [10, 11]. However, the high dimensional mode indicates that there are many tokens that can be perturbed more substantially. Intuitively, the token subspace is dimensionally "thicker" near these tokens with higher dimensional neighborhoods. This yields a striking interpretation: the high dimensional modes correspond to tokens with a much higher variance, while the lower dimensional modes have a lower variance. Therefore, an immediate consequence of Figure 2 is that the noise near a token is strongly and unavoidably dependent upon that token.

1.2 Contributions

The dependence of variability near a token upon that token is a form of *heteroscedasticity*. In order to construct a manifold hypothesis testing framework, we formalize the notion of heteroscedasticity by making a very general model of non-heteroscedastic noise: it is a probability distribution supported on a *fiber bundle*. Roughly speaking, instead of having one local dimension (as a manifold does), a fiber bundle has *two* local dimensions. These two dimensions correspond to a clean split between "signal" and "noise" dimensions. The fiber bundle hypothesis asserts that the noise dimension is valid near a given point, while the signal dimension is valid further away from that point. While this hypothesis may not be true, if it is true then the noise model is quite benign.

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited).

According to Theorem 1, which we call the "fiber bundle hypothesis," proven in the supplementary material, it is easy to test for a fiber bundle using the volume-versus-radius plots of [9] by finding places where the slope is discontinuous and increases at this discontinuity. The present paper explains that the token subspaces for LLMs mostly, but not entirely, look like fiber bundles. The places where the token subspace has singularities (violates the fiber bundle hypothesis) are likely to be at interesting tokens.

In Section 2, we explain how to test and interpret the fiber bundle hypothesis. As a benefit, Theorem 1 yields two new dimension estimators that aid in performing the test. Rejection of the fiber bundle hypothesis therefore implies a very strong heteroscedasticity. We rebuilt the dimension estimator in [9] to automatically find the stratification boundaries.

In Section 3, we exhibit results from our new estimator on GPT2 [3], Llemma7B [4], Mistral7B [5], and Pythia6.9B [6]. Tokens near violations of the fiber bundle hypothesis are near places where the noise distribution is guaranteed to change abruptly. Furthermore, the two dimension estimators from Theorem 1 also identify sets of tokens with interesting structure.

2 Methods

Our method assumes that the set of tokens T is a random sample of a probability distribution m on a topological space (not necessarily a manifold) E that represents all possible tokens (including those that have not been seen before). We can safely assume that the token input embedding $e : T \to X = \mathbb{R}^d$ is a continuous function.

Supposing that t is a token of interest, our method estimates the probability distribution m in the neighborhood of $x = e(t)$, and uses this estimate to infer properties about the structure of E. Since we assumed that T was randomly sampled from m, the number of tokens within radius r of x,

$$N_x(r) := \{y \in T : \|x - y\|_2 < r\},$$

will converge in expectation to

$$\mathbb{E}(N_x(r)) = m(e^{-1}(B_r(x)))\#T, \tag{1}$$

as the number of tokens grows large, provided e is continuous.

Theorem 1, provides an asymptotic estimate of Equation (1) under additional assumptions about E. Specifically, Theorem 1 asserts that if E is a manifold and e is smooth, then $\log \mathbb{E}(N_x(r))$ depends linearly upon $\log r$, in which the slope of this linear relationship is the dimension of E. Moreover, Theorem 1 asserts that if E is a *fibered manifold*, a generalization of a manifold that is a type of *fiber bundle* (as discussed in Section 2.1), then the relationship is piecewise linear, and the slopes *must decrease* as the radius increases.

If the conclusion of Theorem 1 does not hold, namely the relationship between $\log \mathbb{E}(N_x(r))$ and $\log r$ is not piecewise linear or the slopes do not decrease

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited).

with increasing r, then we *reject* the fiber bundle hypothesis. In particular, *rejecting implies that the neighborhood of x is inconsistent with a fiber bundle, and as a consequence, it is also inconsistent with a manifold.*

2.1 Interpretation as signal and noise

It is usual to describe measurements as exhibiting the combined effect of signal and noise. If we were to know both of these quantities, we could express each measurement as being an ordered pair (signal, noise). Therefore, if the space of all possible signals is B and the space of all noise values is V, we could represent the space of all possible measurements as the cartesian product $E = B \times V$. In what follows, we will call B the *base space* and V the *fiber space*. The product $E = B \times V$ describes the situation when the set of possible noise values does not depend on the signal value, and is called a *homoscedastic* noise model. In contrast, in a *heteroscedastic* noise model, the set of possible noise values depends on the signal value.

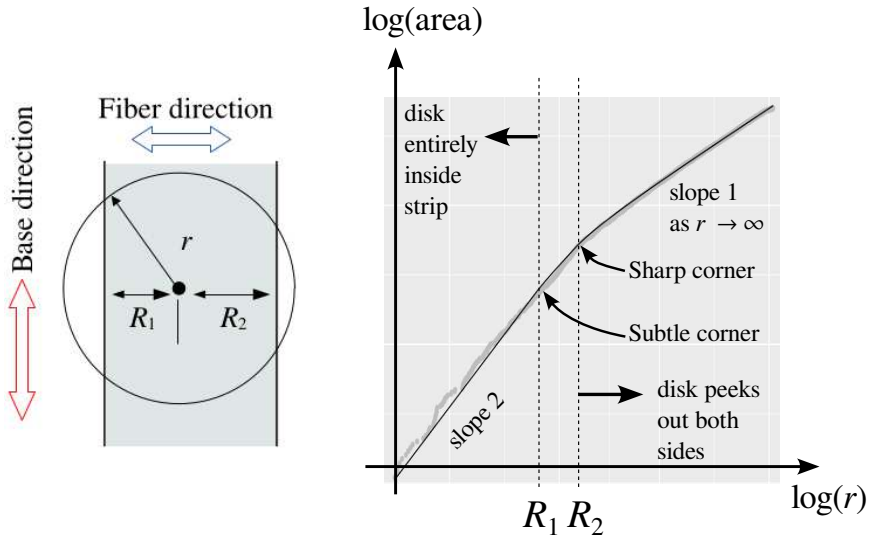

Figure 3: Our method applied to a fiber bundle in $\mathbb{R}^2$. The vertical direction is the base space (signal), while the horizontal direction represents the fibers space (noise). Gray points on the right frame show estimates from a random sampling of points in the strip; the solid line shows the theoretical area versus radius curve.

Figure 3 shows an example of this situation. It consists of a 1-dimensional base space (the signal) and 1-dimensional fibers (the noise), which in this case forms a narrow strip in the plane. Volumes (areas, in this case) of balls of small radius scale quadratically (slope 2 in a log-log plot), but scale asymptotically linearly (slope 1 in a log-log plot) for large radii. The transition between these

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited).

two behaviors is detectable by way of a corner in the plot. This situation is easily and robustly estimated from the data; the gray points in Figure 3 (right) are derived from a random sampling of points drawn from the strip.

To test for heteroscedastic noise, we propose a substantial *nonlinear* generalization of homoscedastic noise. While the *strength* of noise can depend on the signal value, the number of dimensions necessary to describe it does not. This situation is modeled mathematically by a *fiber bundle*. In a fiber bundle, the signal is still modeled by a space B, but the possible measurements are modeled by a function $p : E \to B$. The idea is that *fibers* $p^{-1}(b)$ are still cartesian products: pairs of signal and noise, and these are all identical up to *diffeomorphism*. Our method relies upon a particular geometric property of fiber bundles: we can identify if the fibers are not all identical according to when the conclusion of Theorem 1 is violated.

Figure 4 shows a situation that is not a fiber bundle, since there is a change in the dimension of the fiber. In the upper portion of the figure, the fiber dimension is 0 while in the lower portion the fiber dimension is 1. This is detectable by looking at the volume versus radius plots for two samples. While both samples show corners in their volume versus radius plots, Theorem 1 establishes that *the slopes always decrease with increasing radius* for a fiber bundle. This is violated for the sample marked (a), so we conclude that the space is not a fiber bundle. On the other hand, because the sample marked (b) does not exhibit this violation, it is important to note that if a sample yields data consistent with Theorem 1, *we cannot* conclude that the space is a fiber bundle.

The statement of Theorem 1 is rather technical, but can be summarized in a simple way. Consider a token x, and count *as a function of radius* r, the number of tokens within radius r of the token x. If we plot this function on a log-log scale, it will be roughly linear for small radii anywhere where the space has the local structure of a manifold near the token x. The manifold hypothesis prohibits discontinuities in the derivative of this function for small radii, but according to Theorem 1, fiber bundles permit the slope to *decrease* through a discontinuity. Therefore, anywhere the slope *increases* through a discontinuity will cause us to conclude that the vicinity of that token cannot be a fiber bundle. Rejecting the fiber bundle hypothesis implies that the token x has far fewer synonyms than its neighbors, and might be a token corresponding to multiple distinct meanings.

2.2 Testing framework for the fiber bundle hypothesis

Our method is summarized by Figure 5. The first three blocks of Figure 5 compute $N_x(r)$ directly, while the last three blocks perform the test to see whether the conclusion of Theorem 1 holds.

Note that the test itself—the final block—is rather straightforward. Theorem 1 asserts that $\log N_x(r)$ as a function of $\log r$ is a piecewise linear function, in which the slopes decrease as r increases. If there is a statistically significant increase in the slope estimates, then we reject.

The most subtle of the blocks in Figure 5 is the fourth block, labeled "detect slope changes". This block consists of estimating the slope by using the standard

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited).

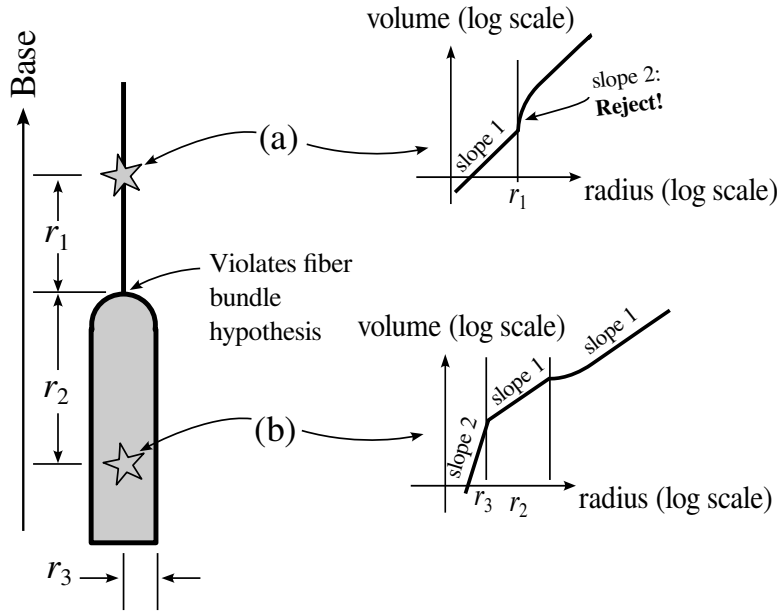

Figure 4: Rejecting fiber bundle model; another example in $\mathbb{R}^2$. Two samples are marked as (a) and (b) along with their volume versus radius plots.

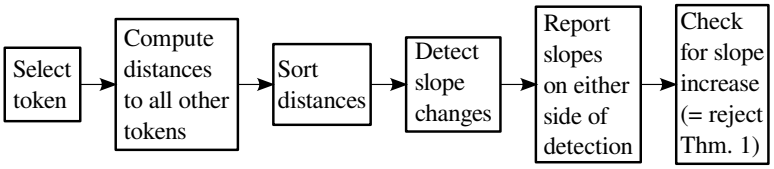

Figure 5: Flow chart of the proposed method.

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited).

three-point centered differences method, and then uses a constant false alarm rate detector to identify changes in these slope estimates as a function of radius. It is worth noting that the false alarm rate is the significance level for our test. For the results shown in Section 3, the significance level was set at 10^{-3}. Nevertheless, we found that our results in Section 3 were insensitive to the false alarm rate, which means that the rejections were highly significant.

3 Results

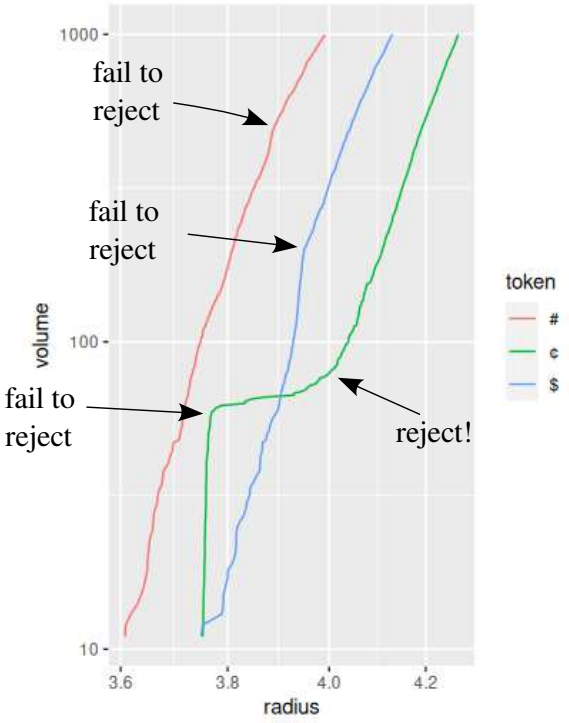

Figure 6: Log-log plots of volume (token count) versus radius for three tokens in GPT2 with significant slope changes marked.

Figure 6 shows the volume versus radius curves for three tokens used by GPT2. Of these, most of the slope changes shown are not inconsistent with the fiber bundle hypothesis posited by Theorem 1. While this does not allow one to conclude that the vicinity of `$` and `#` are fiber bundles, if this were to be the case, we could use Theorem 1 to estimate the base and fiber dimension from the slopes on either side of the marked points.

Notice that the curve for `¢` exhibits two slope changes. One slope change

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited).

in Figure 6 represents a violation of the fiber bundle hypothesis posited by Theorem 1, which implies that the vicinity of ¢ does not split cleanly into signal versus noise. The rejection for ¢ is interesting: there are some sentences in which the presence of ¢ is essential. (Note that ¢ was chosen for illustrative purposes. The p-value for rejecting the fiber bundle hypothesis at ¢ is larger than $\alpha = 10^{-3}$, so ¢ does not appear in Table 2.)

Given that each token subspace consists of multiple tokens, and we perform the testing methodology in Section 2 for each token, it is important to distinguish between two variants of the manifold and fiber bundle hypotheses: "is the token subspace a manifold (or fiber bundle) overall?" and "is the token subspace a manifold (or fiber bundle) near a given token?" The methodology in Section 2 performs the latter directly. Each token consists of a statistical test, the collection of which is aggregated over the entire token space. Therefore, we applied the Holm-Bonferroni multiple test correction to the p-values of each token's test. Rejections were reported using a significance level of $\alpha = 10^{-3}$. To address the former question, the number of rejections for the two slope changes (if they occur) are shown as two separate columns in Table 1.

Table 1: Dimensional data for and number of tokens rejecting the manifold and fiber bundle hypotheses

Model	Manifold	Base		Fiber	
	rejects	dim.	rejects	dim.	rejects
GPT2	68	14	7	389	12
$n = 50257$	$p < 3 \times 10^{-8}$		$p < 3 \times 10^{-8}$		$p < 9 \times 10^{-6}$
Llemma7B	33	11	1	$> 10^6$	0
$n = 32016$	$p < 5 \times 10^{-9}$		$p < 3 \times 10^{-4}$		N/A
Mistral7B	40	6	2	48	1
$n = 32016$	$p < 3 \times 10^{-7}$		$p < 8 \times 10^{-5}$		$p < 8 \times 10^{-4}$
Pythia6.9B	54	2	0	135	0
$n = 50254$	$p < 2 \times 10^{-7}$		N/A		N/A

Table 1 shows the results for the four models we analyzed. It is clear that the models have quite different token input embeddings, and all of them exhibit highly significant rejections of the manifold hypothesis. GPT2, Llemma7B and Mistral7B also reject the fiber bundle hypothesis. The rejections of the fiber bundle hypothesis are more frequent in the base space than the fiber space, which is consistent with the polysemy interpretation of [2]. Table 2 shows each of the fiber bundle violations for each model that are listed in Table 1.

While most of the tokens are not shared between the LLMs, Llemma7B and Mistral7B do have identical token sets. The fact that Table 1 shows significant differences between these two models indicates that the structure of the singularities for these these two models is quite different. This implies that their response to the same prompt is expected to be markedly different, even without considering their respective transformer stages.

There are many more rejections of the manifold hypothesis than can be

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited).

Table 2: Violations of the fiber bundle hypothesis

Model	Token	Base/fiber	p-value	Comment
GPT2	`Xan`	Base	3×10^{-8}	Must start a word
GPT2	`aunder`	Base	2×10^{-4}	
GPT2	`Dri`	Base	2×10^{-4}	
GPT2	`ney`	Base	3×10^{-4}	
GPT2	`rodu`	Base	3×10^{-4}	
GPT2	`Insert`	Base	4×10^{-4}	
GPT2	`Ying`	Base	4×10^{-4}	Must start a word
GPT2	`laughable`	Fiber	9×10^{-6}	Must start a word
GPT2	`nuance`	Fiber	2×10^{-4}	Must start a word
GPT2	`dt`	Fiber	2×10^{-4}	
GPT2	`Mesh`	Fiber	2×10^{-4}	
GPT2	`affect`	Fiber	3×10^{-4}	Must start a word
GPT2	`Thankfully`	Fiber	3×10^{-4}	
GPT2	`swat`	Fiber	6×10^{-4}	Must start a word
GPT2	`Malaysian`	Fiber	6×10^{-4}	Must start a word
GPT2	`Palestinian`	Fiber	7×10^{-4}	Must start a word
GPT2	`wins`	Fiber	8×10^{-4}	Must start a word
GPT2	`hedon`	Fiber	9×10^{-4}	
GPT2	`donor`	Fiber	9×10^{-4}	Must start a word
Llemma7B	`pax`	Base	3×10^{-4}	
Mistral7B	`H0`	Base	5×10^{-4}	
Mistral7B	`monitor`	Base	8×10^{-5}	Must start a word
Mistral7B	`änge`	Fiber	8×10^{-4}	

conveniently listed in a table. Therefore, we list some general trends of which tokens cause the manifold hypothesis to be rejected.

- The GPT2 tokens at singularities are tokens that can only appear at the beginning of words.
- The Pythia6.9B tokens at singularities are nearly all word fragments or short sequences of text that are quite meaningless on their own.
- The Llemma7B and Mistral7B tokens at singularities are a combination of the previous two: either they can only appear at the beginning of words or they are word fragments.

Figures 7–10 show representations of each of the models we analyzed. The visualizations were created by first reducing the latent space dimension from its original value to 50 via principal components analysis, then further reducing to 2 dimensions via t-SNE. The fiber space is clearly stratified in each of the models, but the kinds of stratifications are rather different.

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited).

In Llemma7B, Pythia6.9B, and GPT2 (Figures 7, 8, and 10 respectively), there are isolated regions of tokens with very small dimensional neighborhoods. This suggests that these particular low-dimensional tokens may exhibit semantic polysemy as anticipated in [2]. In Pythia6.9B, the "pinch point" shown in Figure 8 consists mostly of long strings of non-printing and whitespace characters.

In Lemma7B and Mistral7B (Figures 7 and 9, respectively), there are stratification boundaries: on one side of the boundary the dimension of tokens is much higher than on the other side. While the interpretation of this kind of stratification is unclear, it suggests that there may be variability in the training data support for the implicated tokens. Given the significant difference in the structure of the spaces shown in Figures 7 and 9, we can conclude that the token subspaces for these two models are quite different, even though both of these LLMs use the same tokens.

The fiber space of GPT2 (Figure 10) also exhibits a feature not seen in the other models, namely a large cluster of low-dimensional tokens isolated from the others. This cluster was identified in [9], and investigation of the cluster revealed that it mostly contains numeric tokens and date-related tokens. Clustered numeric tokens likely means it is hard for GPT2 to distinguish different numbers. This could cause GPT2 to fail to distinguish between prompts involving dates from those involving mathematical operations.

The base space is not visibly stratified in Llemma7B, Pythia6.9B, and GPT2 (Figures 7, 8, and 10 respectively), but is visibly stratified in Mistral7B (Figure 9).

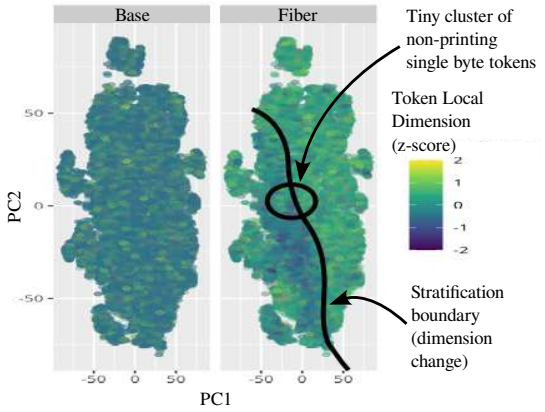

Figure 7: Scatterplot of Llemma7B tokens colored by local base and fiber dimension, projected to 2d via principal components analysis. Because the distribution of dimensions is very different for base and fiber, the colors are normalized via z-scores independently for base and fiber.

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited).

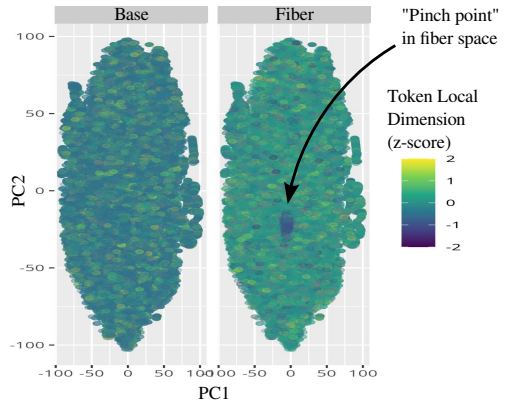

Figure 8: Scatterplot of Pythia6.9B tokens colored by local base and fiber dimension, projected to 2d via principal components analysis. Because the distribution of dimensions is very different for base and fiber, the colors are normalized via z-scores independently for base and fiber. The pinch point shown in the fiber space consists mostly of strings of non-printing and whitespace characters.

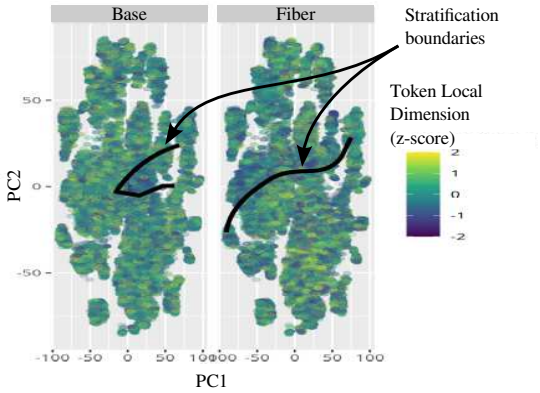

Figure 9: Scatterplot of Mistral7B tokens colored by local base and fiber dimension, projected to 2d via principal components analysis. Because the distribution of dimensions is very different for base and fiber, the colors are normalized via z-scores independently for base and fiber.

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited).

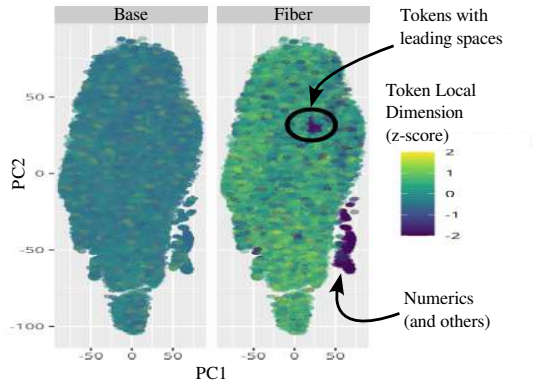

Figure 10: Scatterplot of GPT2 tokens colored by local base and fiber dimension, projected to 2d via principal components analysis. Because the distribution of dimensions is very different for base and fiber, the colors are normalized via z-scores independently for base and fiber.

4 Discussion

None of the four LLMs we studied have token subspaces that are manifolds, and three of the four are also not fiber bundles. Singularities—tokens that cause rejections of the manifold hypothesis—occur in different ways across all four LLMs. Additionally, singularities correspond to violations of the fiber bundle hypothesis are tokens whose neighborhoods exhibit a dependency between the large- and small-scale variability.

Singularities may arise either as artifacts of the training process or from features of the languages being represented. Consistent with the idea that polysemy may yield singularities [2], several of the tokens in Table 2 are clear homonyms. For instance, both "affect" and "monitor" can be used either nouns or verbs, and their meanings are different in these two roles.

Because tokens are fragments of text, a token may correspond to homonyms after the addition of a prefix or suffix. A token like "aunder" can be prefixed to yield the word "launder", which is a *contranym*—a word with multiple meanings of opposite sense. Specifically, one can "launder" clothing (which has a positive connotation) or "launder" money (which has a negative connotation). Several other tokens in Table 2 form words with substantially different meanings or grammatical roles upon adding a prefix or suffice. For instance, "wins" can appear as a noun, a verb, and is also part of the adjective "winsome".

The grammatical roles of tokens is likely a root cause for some of sensitivity of LLMs to their prompts that has been observed in the literature, and may explain why "explaining LLM behavior" remains difficult. Most methods for explaining LLM behavior in terms of dynamical systems, for instance, derive their inferential power from assuming that the token subspace is a manifold.

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited).

Our results show that these theoretical methods simply do not apply to actual LLMs.

The fact that the LLMs are not manifolds means that the geodesic distance between tokens can be very unstable. As a result, while the distance along geodesics can be defined, it may not correlate with any sense of semantic distance between tokens. Furthermore, as [9] indicated, in most of the models, there are tokens with dimension 0 neighborhoods. These tokens are therefore *isolated*, which implies that the token subspace is disconnected. The geodesic distance between an isolated token and any other token is therefore infinite.

The differences in how the manifold and fiber bundle hypotheses are rejected across different LLMs suggest that the training methodology for each model leaves an indelible fingerprint. Making general assertions about LLMs without consideration of the details of their training is likely fraught. Even between Llemma7B and Mistral7B, which have identical tokens, *prompts likely cannot be "ported" from one LLM to another without significant change if they contain tokens near singularities.*

A few clear patterns among tokens near singularities are nevertheless noticeable. Tokens that begin a word or are a word fragment are often located at a singularity. Additionally, in Llemma7B (but not Mistral7B) and Pythia6.9B the tokens with unusually low fiber dimension often contain non-printing or whitespace characters. This suggests that these models are quite sensitive to text layout, perhaps to the exclusion of more semantically salient features in the text. Given our findings, future experiments can be run to explore the impact of singular tokens on the variability of responses produced by different LLMs.

Acknowledgments

The authors would like to thank Anand Sarwate and Andrew Lauziere for helpful suggestions on a draft of this manuscript. This article is based upon work partially supported by the Defense Advanced Research Projects Agency (DARPA). Any opinions, findings and conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

References

[1] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. *ArXiv*, abs/2310.11324, 2023.

[2] Alexander Jakubowski, Milica Gasic, and Marcus Zibrowius. Topology of word embeddings: Singularities reflect polysemy. In Iryna Gurevych, Marianna Apidianaki, and Manaal Faruqui, editors, *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 103–113,

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited).

Barcelona, Spain (Online), December 2020. Association for Computational Linguistics.

[3] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[4] Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. Llemma: An open language model for mathematics, 2024.

[5] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral7b, 2023.

[6] Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling, 2023.

[7] Max Vargas, Reilly Cannon, Andrew Engel, Anand D Sarwate, and Tony Chiang. Understanding generative ai content with embedding models, 2024.

[8] Archit Rathore, Yichu Zhou, Vivek Srikumar, and Bei Wang. Topobert: Exploring the topology of fine-tuned word representations. *Information Visualization*, 22(3):186–208, 2023.

[9] Michael Robinson, Sourya Dey, and Shauna Sweet. The structure of the token space for large language models, 2024.

[10] Vasilii A. Gromov, Nikita S. Borodin, and Asel S. Yerbolova. A language and its dimensions: Intrinsic dimensions of language fractal structures. *Complexity*, 2024.

[11] Eduard Tulchinskii, Kristian Kuznetsov, Laida Kushnareva, Daniil Cherniavskii, Serguei Barannikov, Irina Piontkovskaya, Sergey Nikolenko, and Evgeny Burnaev. Intrinsic dimension estimation for robust detection of AI-generated texts, 2023.

[12] Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

[13] J. Lee. *Smooth Manifolds*. Springer, 2003.

[14] Alfred Gray. The volume of a small geodesic ball of a Riemannian manifold. *Michigan Mathematical Journal*, 20(4):329 – 344, 1974.

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited).

Supplementary

This section contains mathematical justification for the fiber bundle model proposed earlier in the paper and the proof of Theorem 1. The central idea is the use of a special kind of fiber bundle, namely a *fibered manifold*. This is done by placing a specific structure on a manifold E that describes the data, by relating it to another, lower dimensional, manifold B, called the *base space*, via a smooth map $p : E \to B$.

Definition 1. A *fibered manifold* is a surjective function $p : E \to B$ such that the Jacobian matrix $d_x p$ at every point $x \in E$ has rank equal to the dimension of B.

By the submersion theorem [13], if the Jacobian matrix of p at every point has rank equal to the dimension of B, then the preimages $p^{-1}(x) \subseteq E$ of each point $x \in E$ are all diffeomorphic to each other. These preimages form the *fibers* discussed in the earlier sections of the paper.

As a consequence, each point y in the base space B has an open neighborhood U where the preimage $p^{-1}(U)$ is diffeomorphic to the product $U \times p^{-1}(y)$, which is precisely the base-fiber split discussed in Section 2.1. Specifically, the base dimension is simply the dimension of B, whereas the fiber dimension is the $(\dim E - \dim B)$.

The notion of a fibered manifold $p : E \to B$ forms the intrinsic model of the data, which is only implicit in an LLM. The tokens present in a given LLM can be thought of as a sample from a probability distribution m on E, which can be taken to be the Riemannian volume form on E normalized so that $m(E) = 1$.

Definition 2. If $f : E \to \mathbb{R}^d$ is a smooth map and m is a volume form on E, then the *pushforward* is defined by

$$(f_* m)(V) := m\left(f^{-1}(V)\right)$$

for each measurable set V.

It is a standard fact that if f is a fibered manifold or an embedding, then $f_* m$ is also a volume form.

The explicit representation of the token subspace arises by embedding the tokens within a Euclidean space $\mathbb{R}^d$. On the hypothesis that the tokens lie on a fibered manifold—recall that they may not—the token input embedding consists of a smooth embedding $e : E \to \mathbb{R}^d$. If this is the correct representation of the tokens, then the probability distribution m on E will impact the distribution of tokens within $\mathbb{R}^d$. Theorem 1 characterizes the resulting probability distribution using parameters (the exponents in Equation (2)) that can be estimated from the token input embedding, as described by the earlier sections of this paper. These parameters are bounded by the dimensions of the base and fiber spaces.

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited).

Theorem 1. *Suppose that E is a compact, finite-dimensional Riemannian manifold with boundary*[1]*, with a volume form m satisfying $m(E) < \infty$, and let $p : E \to B$ be a fibered manifold.*

If $e : E \to \mathbb{R}^d$ is a smooth embedding with reach τ, then there is a function $\rho : e(E) \to [0, \tau]$ such that if for $x \in e(E)$,

$$(e_*m)\left(B_r(x)\right) = \begin{cases} O(r^{\dim E}) & \text{if } 0 \le r \le \rho(x), \\ (e_*m)\left(B_{\rho(x)}(x)\right) + O((r-\rho(x))^{\dim B}) & \text{if } \rho(x) \le r, \end{cases} \tag{2}$$

where the asymptotic limits are valid for small r.

As a special case, m may be normalized to yield a probability measure.

Proof. Since e is assumed to be a smooth embedding, the image of e is a manifold of dimension $\dim E$. The pushforward of a volume form is a contravariant functor, so this means that e_*m is the volume form for a Riemannian metric on $e(E)$. Using this Riemannian metric on $e(E)$, then [14, Thm 3.1] implies that for every $x \in e(E)$, if $r \ll \tau$, then

$$(e_*m)\left(B_r(x)\right) = O\left(r^{\dim E}\right). \tag{3}$$

Since E is compact, B is also compact via the surjectivity of p. This implies that there is a maximum radius r_1 for which a ball of this radius centered on a point on $x \in e(E)$ is entirely contained within $e(E)$. Also by compactness of B, there is a minimum radius r_2 such that a ball of radius r_2 centered on a point $x \in e(E)$ contains a point outside of $e(E)$.

Since e is assumed to be an embedding, by the tubular neighborhood theorem [13], it must be that $r_2 < \tau$. Define

$$\rho(x) := \operatorname{argmax}_r \left\{B_r(x) \subseteq e(E)\right\},$$

from which it follows that $0 < r_1 \le \rho(x) \le r_2 < \tau$. As a result, Equation (3) holds for all $r < \rho(x)$, which is also the first case listed in Equation (2).

If r is chosen such that $\rho(x) < r < \tau$, the volume of the ball centered on x of radius r will be less than what is given by Equation (3), namely

$$(e_*m)\left(B_r(x)\right) < O(r^{\dim E}).$$

Since m is a volume form, its pushforward (p_*m) onto B is also a volume form. Moreover, via the surjectivity of p,

$$\begin{aligned} (e_*m)(B_r(x)) &= m(e^{-1}(B_r(x))) \\ &\le m(p^{-1}(p(e^{-1}(B_r(x))))) \\ &\le (p_*m)(p(e^{-1}(B_r(x)))) \\ &\le O(r^{\dim B}). \end{aligned}$$

[1] Every point in a *manifold with boundary* has a neighborhood that is locally homeomorphic to a half-space. As a consequence, manifolds are a special case of manifolds with boundary.

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited).

From this, the second case of Equation (2) follows by recentering the asymptotic series on $\rho(x)$. □

Notice that the second case in Equation (2) may be precluded since while it holds for small r, it may be that $\rho(x)$ may not be sufficiently small. As a consequence, the second case only occurs when both r and $\rho(x)$ are sufficiently small. In the results shown in Section 3, both cases appear to hold frequently.

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Distribution Statement "A" (Approved for Public Release, Distribution Unlimited).