Chamfer Distance Breaks the Curse of Dimensionality: On the Stability of Multi-Vector Retrieval

Anonymous

Abstract

Modern vector databases enable efficient retrieval over high-dimensional neural embeddings, powering applications from web search to retrieval-augmented generation. However, classical theory predicts such tasks should suffer from *the curse of dimensionality*, where distances between points become nearly indistinguishable, thereby crippling efficient nearest-neighbor search. We revisit this paradox through the lens of *stability*, the property that small perturbations to a query do not radically alter its nearest neighbors. Building on foundational results, we extend stability theory to multi-vector search, where we prove that the popular Chamfer distance metric preserves single-vector stability, while average pooling aggregation may destroy it. Across synthetic and real datasets, our experimental results match our theoretical predictions, offering concrete guidance for model and system design to circumvent the curse of dimensionality in multi-vector settings.

Keywords

multi-vector search, late interaction, vector databases, near-neighbor search

The *curse of dimensionality* is a notorious barrier to algorithmic efficiency in computational geometry. Informally, the curse states that points in high-dimensional metric spaces tend to be arbitrarily close in distance. Consequently, this property makes it challenging for techniques such as near-neighbor search and near-neighbor classification to identify the closest points to a query without resorting to a full brute-force scan, which can be prohibitively slow in practice. Nevertheless, modern vector databases, powered by stunning advances in neural representation learning, have achieved remarkable adoption in spite of high vector dimensionalities, powering applications such as retrieval-augmented generation [1], web-search [2], and product search [3].

While a number of prior works have studied how traditional vector search interacts with the curse of dimensionality [4, 5, 6, 7], extending this research to the burgeoning framework of *multi-vector retrieval*, which involves near-neighbor search in the case where both the query and the database items consist of *sets* of vectors, remains an open problem unexplored in the current research literature. This style of search has become extremely popular due to the recent success of "late interaction" models such as ColBERT [8] and continues to achieve state-of-the-art performance in a number of retrieval domains [9].

In this paper, we study the question of *why* sublinear search techniques for multi-vector retrieval [10, 11, 12] succeed inspite of the curse of dimensionality. In particular, we present novel theoretical end experiments results extending the formal notion of *stability* to the multi-vector setting to characterize the conditions under which the curse of dimensionality is overcome. Intuitively, stability captures the notion that a slight perturbation to a query should not dramatically change the list of near neighbors to that query. The landmark paper of Beyer et al. [4] and a subsequent follow-up work of Durrant and Kabán [5] together establish a necessary and sufficient condition for stability via a quantity called *relative variance*. However, to our knowledge, no work has studied extending these prior results to the multi-vector setting despite its increasingly critical role in modern retrieval systems. Thus, our work fills a gap in the literature in understanding what properties of multi-vector search algorithms and systems enable sub-linear vector search in high dimensions. Our specific contributions are as follows:

- We prove that multi-vector search with the commonly-used Chamfer distance is stable if the *induced* single vector search problem is *strongly stable* (along with some technical conditions on the composition of the vector sets).

Woodstock'22: Symposium on the irreproducible science, June 07–11, 2022, Woodstock, NY

- However, not all aggregations are stability preserving; we give a counter-example showing that stability does not hold in general for average pooling. This result provides a theoretical justification for the choice of the Chamfer distance used in the popular ColBERT family of models [8, 10, 9].
- In addition to our theoretical results, we also empirically validate each of these theorems on a combination of synthetic and real embedding datasets[1].

1. Background

In this section, we introduce the necessary mathematical preliminaries for our main results. We begin with a formal definition of nearest-neighbor search.

Definition 1.1 (Near-Neighbor Search Problem)**.** A *near-neighbor search problem* is a 3-tuple (Q, D, δ) for a query set Q and database collection D where $q, d \in \mathbb{R}^m$ for all $q \in Q$ and $d \in D$. For each $q \in Q$, we use the distance metric $\delta : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}_{\geq 0}$ to compute

$$\operatorname*{argmin}_{d \in D} \delta(q, d)$$

In the remainder of this paper, we will assume that all database elements $d \in D$ and all query elements $q \in Q$ are sampled from two respective distributions. Under this probabilistic model, the distance $\delta(q, d)$ for $q \sim Q$ and $d \sim D$ is a random variable. In a seminal paper, Beyer et al. [4] introduced this probabilistic model to provide a precise formalization of the curse of dimensionality in the context of near-neighbor search by defining the notion of instability.

Definition 1.2 (Instability [4])**.** A nearest neighbor problem (Q, D, δ) is *unstable* if for all $\epsilon > 0$

$$\lim_{m \to \infty} \Pr_{q \sim Q} \left[\max_{d \in D} \{\delta(q, d)\} \leq (1 + \epsilon) \min_{d \in D} \{\delta(q, d)\} \right] = 1$$

If this property does not hold, we say the problem is *stable*.

Remark 1.3. For notational convenience, we will not explicitly index the random variable $\delta(q, d)$ by the dimensionality, but it is important to remember that $\delta(q, d)$ is a function of the dimension m.

We can also equivalently characterize instability as the convergence of the ratio of the maximum instance to the minimum distance as we take the dimensionality to infinity.

Corollary 1.4 (Beyer et al. [4])**.** *For a nearest neighbor problem (Q, D, δ), let $q \sim Q$ and define* $\text{DMAX} = \max_d \delta(q, d)$ *and* $\text{DMIN} = \min_d \delta(q, d)$*. A nearest neighbor problem is unstable if and only if $\frac{\text{DMAX}}{\text{DMIN}} \to_p 1$ where the sequence is taken over the dimensionality.*

In other words, a nearest neighbor problem is unstable if the distance between the closest and farthest points across all queries becomes nearly indistinguishable. We can also interpret stability as the property that small perturbations to a query do not dramatically alter its list of near neighbors. This is a mathematical formalization of the curse of dimensionality, which states that high-dimensional spaces lack the contrast needed for meaningful near-neighbor search. In this setting, sublinear approximate search algorithms fail, and a brute force linear scan becomes necessary.

Example 1.5 (Example of Stability). One concrete example of a stable near-neighbor search problem, originally proposed by [4], arises from constructing query and document vectors such that each component in every vector is equal, and this value is iid sampled from some nontrivial distribution.

Example 1.6 (Example of Instability). An example of an unstable search problem, also originally identified by [4], is one where the query and document distributions have finite moments and each component of all vectors is iid sampled from these respective distributions.

[1] Link to code redacted to preserve author anonymity

To prove that a near-neighbor problem is stable or unstable, we will also stand on the shoulders of two prior theorems from Beyer et al. [4] and Durrant and Kabán [5], respectively.

Theorem 1.7 (Beyer et al. [4]). *If* $\lim_{m\to\infty} \frac{\mathrm{Var}[\delta(q,d)]}{\mathbb{E}[\delta(q,d)]^2} = 0$*, then the nearest neighbor problem is unstable.*

Theorem 1.8 (Durrant and Kabán [5]). *For a sufficiently large database size* $|D|$*, if* $\lim_{m\to\infty} \frac{\mathrm{Var}[\delta(q,d)]}{\mathbb{E}[\delta(q,d)]^2} \neq 0$*, then the nearest neighbor problem is stable.*

Following the convention established in [4], we refer to the quantity $\frac{\mathrm{Var}[\delta(q,d)]}{\mathbb{E}[\delta(q,d)]^2}$ as the *relative variance* or RelVar. In this paper, we will utilize both the convergence in probability characterization and the relative variance criterion to prove stability.

2. Related Work

2.1. Conditions for Stability

The paper of Beyer et al. [4] was a foundational work that inspired significant follow-up research in understanding the conditions for stable vector search. Aggarwal et al. [6] study stability under common ℓ_p norm distance metrics, where they find that ℓ_1 and fractional distance metrics are theoretically and practically favorable to ℓ_2 distances. Furthermore, Houle et al. [7] prove that the so-called *shared-neighbor* distance function, which measures the number of near-neighbors in common between two points, is a more stable measure than absolute distance since the latter loses discriminatory power in high dimensions. In a different vein, Hui and Gao [13] study stability in the context of sequential data, such as genomes and sensor readings, and show that, under uniformly distributed sequences, the RelVar under edit distance goes to zero in the limit. Moreover, Pestov [14] extends the results of Beyer et al. [4] from k-nearest neighbor *search* to k-nearest neighbor *classification* where they find that instability of vector search also implies barriers to the learnability of near-neighbor classification.

2.2. Intrinsic Dimensionality

A closely related concept to stability is *intrinsic dimensionality* (ID) which measures the dimensionality of the subspace that a collection of vectors lie within. Intuitively, if a database of high-dimensional vectors lies within a much lower-dimensional manifold, then a near-neighbor search over these points is insulated from the curse of dimensionality effect. Thus, much like stability, ID provides a foundational theoretical tool for explaining why modern vector retrieval is capable of circumventing the curse of dimensionality. Consequently, ID has been well-studied with multiple proposed formal definitions. Perhaps the most influential definition of ID is that of [15] via a geomeric property called the *expansion rate*. Another popular definition heavily utilized in near-neighbor search algorithm design is the notion of *doubling dimension* proposed by [16]. We refer interested readers to the recent book of [17] for the precise definitions. More recently, several works have measured the intrinsic dimensionality of modern neural embeddings and have reported that the ID is often much less than the ambient dimensionality. For instance, [18] study the intrinsic dimensionality of embeddings from transformer models [19] and find that the ID typically lies between 20 and 60 over the course of pretraining. In addition, [20] compute the intrinsic dimensionality of popular image model embeddings, finding that the ID in practice typically ranges from 12 to 19, well below the ambient dimensions of 128 to 512. Moreover, [21] go beyond embeddings and measure the ID of convolutional neural network (CNN) layers where they find the ID is an order of magnitude smaller than the ambient dimension.

Given that neural models consistently exhibit low intrinsic dimensionality, a natural question to ask is whether we still need to study why modern vector databases succeed in spite of the curse of dimensionality. On the contrary, we argue that intrinsic dimensionality alone is insufficient to explain

the phenomena we see in practice. First and foremost, as we empirically measure in Section **??**, a dimensionality as low as 32, which is within the empirical ID measurements of real embeddings, is still high enough for the curse of dimensionality to diminish the search quality of approximate near neighbor search algorithms. This observation motivates the need to study stability explicitly. Secondly, modern embeddings produced by large language models and used in practice continue to grow in dimension into the thousands and tens of thousands of components. It is not clear if the intrinsic dimensionality will remain low as these embedding models scale and yet we observe they remain amenable to sub-linear vector search techniques. Moreover, stability analysis provides actionable guidance that ID alone cannot. For instance, our multi-vector stability theorem explains why ColBERT's choice of Chamfer distance succeeds where average pooling might fail, a distinction invisible to ID measurements which only characterize the embedding space geometry, not the aggregation function.

2.3. Multi-Vector Search

Our focus in this work is multi-vector search, where each query and document are represented as *sets* of vectors under a chosen set-similarity function. This style of vector retrieval first gained traction with the celebrated ColBERT paper [8]. Since the introduction of ColBERT, multiple works have studied how to accelerate of multi-vector search algorithms, including PLAID [10], DESSERT [11], and MUVERA [12]. In addition, [22] demonstrate that multi-vector search under the commonly-used Chamfer distance [23] can be integrated into the existing graph-based DiskANN index [24] originally conceived for single-vector search. Furthermore, Faysse et al. [9] extend the multi-vector modeling paradigm pioneered by ColBERT beyond text to multimodal vision-language retrieval. However, despite the tremendous research interest in multi-vector search, we note that no prior work has studied how the use of vector sets relates to the curse of dimensionality. We address this gap in the literature.

3. Stability of Multi-vector Search

We now turn our attention to the stability of multi-vector search. We begin by introducing a formal mathematical model of multi-vector search including a new, analogous definition of stability in this setting. We then proceed to our main theorem of this section identifying a sufficient condition for such stability to exist.

3.1. Theoretical Model

Definition 3.1 (Multi-Vector Search Problem)**.** A multi-vector search problem is specified by a 4-tuple $(\mathcal{Q}, \mathcal{D}, \delta, \text{Agg})$, where:

- $\mathcal{Q}$ is a collection of *query sets*. Each query $Q \in \mathcal{Q}$ is a finite set of vectors.
- $\mathcal{D}$ is a database of *document sets*. Each document $D \in \mathcal{D}$ is a finite set of vectors.
- $\delta : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}_{\geq 0}$ is a *primitive distance metric* between individual vectors.
- Agg is a *set aggregation function* that computes the distance between a query set Q and a document set D, based on the primitive distances between their constituent vectors.

For a given query $Q \in \mathcal{Q}$, the goal is to find:

$$\arg\min_{D \in \mathcal{D}} \text{Agg}(Q, D)$$

Below, we define two concrete aggregation functions that we consider in this work.

Definition 3.2. For two finite vector sets A and B where $A, B \subset \mathbb{R}^m$, and a vector distance function δ, the *Chamfer Distance* [23] is defined as

$$\text{Chamfer}(A, B) = \sum_{a \in A} \min_{b \in B} \delta(a, b)$$

Definition 3.3. For two finite vector sets A and B where $A, B \subset \mathbb{R}^m$, and a vector distance function δ, *Average Pooling* aggregation is defined as

$$\text{AvgPool}(A, B) = \frac{1}{|A| \cdot |B|} \sum_{a \in A} \sum_{b \in B} \delta(a, b)$$

Now, we will define a new notion of stability in the multi-vector search setting.

Definition 3.4 (Multi-Vector Search Stability)**.** A multi-vector search problem is unstable if for all $\epsilon > 0$

$$\lim_{m \to \infty} \Pr_{Q \sim \mathcal{Q}} \left[\max_{D \in \mathcal{D}} \{\text{Agg}(Q, D\} \leq (1 + \epsilon) \min_{D \in \mathcal{D}} \{\text{Agg}(Q, D)\} \right] = 1$$

We can now state our first theorem of this section generalizing the results of [4] and [5] to the multi-vector setting.

Theorem 3.5. *For sufficiently large databases, a multi-vector search instance is stable if and only if*

$$\lim_{m \to \infty} \frac{\text{Var}\left[\text{Agg}(Q, D)\right]}{\mathbb{E}\left[\text{Agg}(Q, D)\right]^2} \neq 0$$

Proof. This result follows directly from the prior proofs of [4] and [5] since the original arguments require only that the distance metric is non-negative. We provide a full, self-contained argument here for completeness.

As a matter of notation, let $\text{Agg}(Q, D)$ be the random variable for the aggregated distance in an m-dimensional space. Let its expected value be $\mathbb{E}[\text{Agg}(Q, D)]$ and its variance be $\text{Var}[\text{Agg}(\text{Q}, \text{D})]$. The database produces n samples of this random variable. Let $\text{DMIN} = \min_{D \in \mathcal{D}} \text{Agg}(Q, D)$ and $\text{DMAX} = \max_{D \in \mathcal{D}} \text{Agg}(Q, D)$.

$\implies$ This direction follows the argument from [4]. We will show that if $\lim_{m \to \infty} \frac{\text{Var}[\text{Agg}(Q,D)]}{\mathbb{E}[\text{Agg}(Q,D)]^2} = 0$, then the multi-vector search problem is unstable. Define a normalized random variable $Y = \frac{\text{Agg}(Q,D)}{\mathbb{E}[\text{Agg}(Q,D)]}$. Its expectation is $\mathbb{E}[Y] = 1$ and its variance is $\text{Var}[Y] = \frac{\text{Var}[\text{Agg}(Q,D)]}{\mathbb{E}[\text{Agg}(Q,D)]^2}$ The condition is thus equivalent to $\lim_{m \to \infty} \text{Var}[Y] = 0$. Since $\lim_{m \to \infty} \mathbb{E}[Y] = 1$ and $\lim_{m \to \infty} \text{Var}[Y] = 0$, the random variable Y converges in probability to 1. Let $\{Y_1, \ldots, Y_n\}$ be the normalized distances for the n items in the database. Since min() and max() are continuous functions, Slutsky's Theorem implies that $\min Y_i \xrightarrow{p} 1$ and $\max Y_i \xrightarrow{p} 1$. Since both the numerator and denominator converge in probability to 1, their ratio also converges in probability to 1. By definition, $\frac{\text{DMAX}}{\text{DMIN}} \xrightarrow{p} 1$ means that for any $\epsilon > 0$, we have $\lim_{m \to \infty} \Pr\left[\left|\frac{\text{DMAX}}{\text{DMIN}} - 1\right| \leq \epsilon\right] = 1$. Since $\text{DMAX} \geq \text{DMIN}$, this is equivalent to $\lim_{m \to \infty} \Pr\left[\text{DMAX} \leq (1 + \epsilon)\text{DMIN}\right] = 1$. This is precisely the definition of an unstable search problem as stated in Definition 3.4.

$\impliedby$ The converse result follows the proof presented by [5]. We proceed by proving the contrapositive: If the multi-vector search problem is unstable, then the relative variance must be zero in the limit. Assume the problem is unstable. Thus, for any $\epsilon > 0$, $\lim_{m \to \infty} \Pr\left[\text{DMAX} \leq (1 + \epsilon)\text{DMIN}\right] = 1$. This directly implies that the ratio of distances converges in probability to 1: $\frac{\text{DMAX}}{\text{DMIN}} \xrightarrow{p} 1$. For a sufficiently large database, we can assume that the expectation $\mathbb{E}[\text{Agg}(Q, D)]$ lies between the sample minimum and maximum. This allows us to bound the normalized random variable $Y = \frac{\text{Agg}(Q,D)}{\mathbb{E}[\text{Agg}(Q,D)]}$ for any sample $D_i \in \mathcal{D}$ as follows:

$$\frac{\text{DMIN}}{\text{DMAX}} \leq \frac{\text{Agg}(Q, D_i)}{\mathbb{E}[\text{Agg}(Q, D_i)]} \leq \frac{\text{DMAX}}{\text{DMIN}}$$

From the instability assumption, we know both the lower bound $\frac{\text{DMIN}}{\text{DMAX}}$ and the upper bound $\frac{\text{DMAX}}{\text{DMIN}}$ converge in probability to 1 By the Squeeze Theorem, the random variable Y must also converge in

probability to 1. A random variable that converges in probability to a constant must have a variance that converges to zero. Therefore:

$$\lim_{m\to\infty} \mathrm{Var}[Y] = \lim_{m\to\infty} \frac{\mathrm{Var}[\mathrm{Agg}(Q,D)]}{\mathbb{E}[\mathrm{Agg}(Q,D)]^2} = 0$$

This completes the proof of the contrapositive. □

With this extension of stability to the multi-vector setting in hand, we will now define a procedure for constructing a single-vector search instance from a multi-vector problem.

Definition 3.6 (Induced Single-Vector Search Problem)**.** Given a multi-vector search instance $(\mathcal{Q}, \mathcal{D}, \delta, \mathrm{Agg})$, we define the *induced* single-vector search problem (Q', D', δ) as follows

- $Q' = \bigcup_{Q_i \in \mathcal{Q}} Q_i$ is the set of unique query vectors, across all multi-vector query sets.
- $D' = \bigcup_{D_i \in \mathcal{D}} D_i$ is the set of unique document components, across all documents.
- The distance metric δ is the primitive distance function of Agg.

3.2. Theoretical Results

From the previous definitions, we can now establish conditions for the stability of multi-vector search under Chamfer distance via the stability of the induced single vector search instance. The key idea behind our argument is to leverage the structure of the Chamfer distance, which, in essence, solves multiple scaled-down near neighbor search problems.

Before stating our main theorem of this section, we will need to define one additional property we call *strong stability*.

Definition 3.7 (Strong Stability)**.** A near-neighbor search problem (Q, D, δ) is c-strongly stable for some constant $c > 1$ if $\frac{\mathrm{DMAX}}{\mathrm{DMIN}} > c + \Omega(1)$ almost surely.

Intuitively, strong stability asserts that a near neighbor problem is not only stable but also stable with a non-decreasing gap. We will need this stronger property when analyzing the stability of multi-vector search because, even if the induced single-vector search problem is stable, the organization of document vectors into sets might contract the maximum distance gap seen in the single vector search problem. Strong stability ensures that, as long as the vector sets obey a certain mild non-degeneracy condition, we retain stability even if the stability gap contracts.

We will now proceed to our main results on the stability of multi-vector search. We will begin by proving a lemma establishing stability in the case where the query sets are singletons (each containing only one vector) while the document sets may be larger in size. Then, we will use this result to prove stability in the general case of both query and document sets with multiple elements.

Lemma 3.8 (Chamfer Stability with Singleton Query Sets)**.** *Let $(\mathcal{Q}, \mathcal{D}, \delta, \mathrm{Chamfer})$ be a multi-vector search problem where $|Q_i| = 1$ for each query set $Q_i \in \mathcal{Q}$. Let (Q', D', δ) denote the corresponding induced single-vector instance . Let $q \sim Q'$ and let $\{D_1, \ldots, D_n\}$ be the document sets in $\mathcal{D}$. Assume the induced single-vector search problem is c-strongly stable.*

If the non-degeneracy condition

$$c \cdot \max_k \min_{d \in D_k} \delta(q,d) \geq \max_k \max_{d \in D_k} \delta(q,d)$$

holds, meaning that there exists at least one document set whose vectors are all sufficiently far from a given query vector, then the multi-vector search instance is stable.

Proof. Let

$$\text{DMIN} = \min_{d \in D'} \delta(q, d), \quad \text{DMAX} = \max_{d \in D'} \delta(q, d).$$

Also define

$$\text{DMIN}' = \min_{k} \min_{d \in D_k} \delta(q, d) = \text{DMIN}, \quad \text{DMAX}' = \max_{k} \min_{d \in D_k} \delta(q, d).$$

We thus have

$$\frac{\text{DMAX}'}{\text{DMIN}'} = \frac{\max_k \min_{d \in D_k} \delta(q, d)}{\text{DMIN}} \geq \frac{1}{c} \frac{\text{DMAX}}{\text{DMIN}} > \frac{1}{c}(c + \epsilon) > 1.$$

for some constant $\epsilon > 0$ by c-strong stability. Thus, it follows that $\frac{\text{DMAX}'}{\text{DMIN}'} \not\to_p 1$, completing the proof. □

Before proceeding to our main theorem, we state one additional helper lemma.

Lemma 3.9 (Dan's Favorite Inequality [25])**.** *Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be positive numbers. Then*

$$\min_i \frac{a_i}{b_i} \leq \frac{\sum_i a_i}{\sum_i b_i} \leq \max_i \frac{a_i}{b_i}$$

Theorem 3.10 (Multi-Vector Stability with Chamfer Distance)**.** *Let $(\mathcal{Q}, \mathcal{D}, \delta, \text{Chamfer})$ be a multi-vector search problem where $|Q| = k$ for some constant $k \geq 1$ for all $Q \in \mathcal{Q}$. Let $A_i = \min_{d \in D} \delta(q_i, d)$ be the nearest-neighbor distance for an individual query vector $q_i \in Q$ for some $Q \sim \mathcal{Q}$. The problem is stable if:*

(a) The induced single-vector search instance is c-strongly stable.

(b) The document sets satisfy non-degeneracy as defined in Lemma 3.8.

(c) $\sum_{i<j} \text{Cov}(A_i, A_j) \geq 0$.

Proof. We will show that

$$\lim_{m \to \infty} \frac{\text{Var}[\text{Chamfer}(Q, D)]}{\mathbb{E}[\text{Chamfer}(Q, D)]^2} > 0$$

which implies the desired result.
To begin, we observe that

$$\lim_{m \to \infty} \frac{\text{Var}[\text{Chamfer}(Q, D)]}{\mathbb{E}[\text{Chamfer}(Q, D)]^2} = \lim_{m \to \infty} \frac{\text{Var}\Big(\sum_{i=1}^k A_i\Big)}{\Big(\mathbb{E}\Big[\sum_{i=1}^k A_i\Big]\Big)^2} = \lim_{m \to \infty} \frac{\text{Var}\Big(\sum_{i=1}^k A_i\Big)}{\Big(\sum_{i=1}^k \mathbb{E}[A_i]\Big)^2}$$

where the second equality follows from the linearity of expectation.
We note that we can lower bound the numerator of the above expression as follows:

$$\text{Var}\Big(\sum_{i=1}^k A_i\Big) = \sum_{i=1}^k \text{Var}[A_i] + 2 \sum_{i<j} \text{Cov}(A_i, A_j) \geq \sum_{i=1}^k \text{Var}[A_i]$$

where the inequality follows from the third assumption in the theorem statement.
Furthermore, we can upper bound the term in the denominator via the Cauchy-Schwarz inequality

$$\Big(\sum_{i=1}^k \mathbb{E}[A_i]\Big)^2 \leq k \sum_{i=1}^k \mathbb{E}[A_i]^2.$$

Putting these bounds together, we have

$$\lim_{m \to \infty} \frac{\text{Var}[\sum_{i=1}^k A_i]}{\big(\sum_{i=1}^k \mathbb{E}[A_i]\big)^2} \geq \lim_{m \to \infty} \frac{\sum_{i=1}^k \text{Var}[A_i]}{k \sum_{i=1}^k \mathbb{E}[A_i]^2} \geq \frac{1}{k} \lim_{m \to \infty} \min_i \frac{\text{Var}[A_i]}{\mathbb{E}[A_i]^2} > 0$$

where the last two inequalities follow from applying Lemma 3.9 and Lemma 3.8 respectively. This completes the proof.

□

Dataset	Stability Constant (c)	Non-Degeneracy Pass Rate (%)	Cov. Bound Pass Rate (%)
MS MARCO[2]	2.14	100.0	100.0
Natural Questions	2.36	100.0	100.0
HotpotQA	2.12	100.0	100.0
TREC-COVID	2.59	100.0	100.0
NFCorpus	1.91	100.0	100.0
Requirement	> 1	$= 100$	$= 100$

Table 1
Validation of Theorem 3.10's assumptions. All conditions are satisfied across all datasets, demonstrating the practical relevance of our theoretical results.

Remark 3.11. Stability follows from applying Dan's Favorite Inequality (Lemma 3.9) to reduce the general multi-vector search problem to multi-vector search over singleton query sets, namely the setting in Lemma 3.8, at the cost of an $\frac{1}{k}$ overhead factor. The $1/k$ factor suggests that stability weakens as the query set size grows, which is also consistent with empirical observations.

3.3. Empirical Validation of Multi-Vector Theorem Assumptions

Theorem 3.10 relies on three assumptions to prove the stability of multi-vector search under Chamfer distance: c-strong stability, non-degenerate document sets, and a non-negative bound on the sum of the covariances of the individual Chamfer distance terms. In this section, we experimentally demonstrate that all three of these assumptions are satisfied in practice, which underscores the practical utility of our result. Specifically, we use ColBERT embeddings [8] to construct query and document vector sets for five standard English-based information retrieval datasets from the BEIR benchmark [26]. For each dataset, we empirically estimate the strong stability constant c as the smallest value of the ratio between the maximum distance and the minimum distance across all queries in the induced single-vector search instance. We then use this value of c to check the proportion of document sets that satisfy the non-degeneracy condition. Finally, we then compute the sum of the covariances of the near-neighbor distances of individual query vectors in a given query set and verify that the sum is non-negative. This non-negative covariance assumption is natural in practice since the tokens in a query set tend to be topically related. We summarize our empirical measurements in Table 1 where we see that all three theorem conditions are comfortably satisfied in practice.

3.4. Stability of ColBERT Embeddings

Following our empirical validation of Theorem 3.10's conditions in the previous section, we also compute the empirical stability ratio (maximum distance to minimum distance per query) and relative variance on a subset of the MSMarco dataset with ColBERT embeddings. We note that this experiment does not, strictly speaking, relate to the theoretical result since ColBERT embeddings maintain a fixed embedding size while stability is an asymptotic property. Nevertheless, we observe in Figure 1 that the asymptotic behavior holds in practice at the ColBERT dimensionality of 768. Under Chamfer distance, the ColBERT vector sets are overwhelmingly more stable than with averaging, which aligns with the fact that we can rigorously prove stability under Chamfer distance but not with averaging.

3.5. Stability of Synthetic Embeddings

To experimentally validate Theorem 3.10, we generate a synthetic family of multi-vector search instances where we can take the dimensionality to be arbitrarily large. In Figure 2 we report the results of our experiment with 1000 document sets and 100 query sets each with 4 vectors per set. Our goal in

[2]For MS MARCO, we use the Passage Ranking dataset: https://github.com/microsoft/MSMARCO-Passage-Ranking.

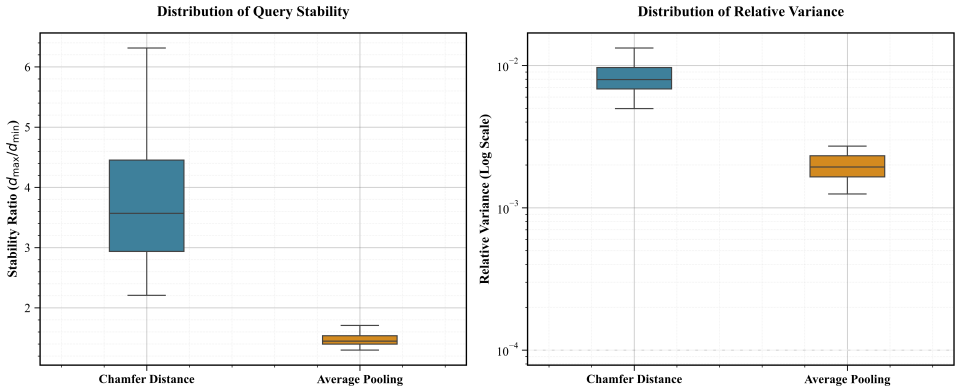

Figure 1: Box plots of query stability and relative variance on the MSMarco dataset. We observe that Chamfer distance is more stable than average pooling, which supports our theoretical findings.

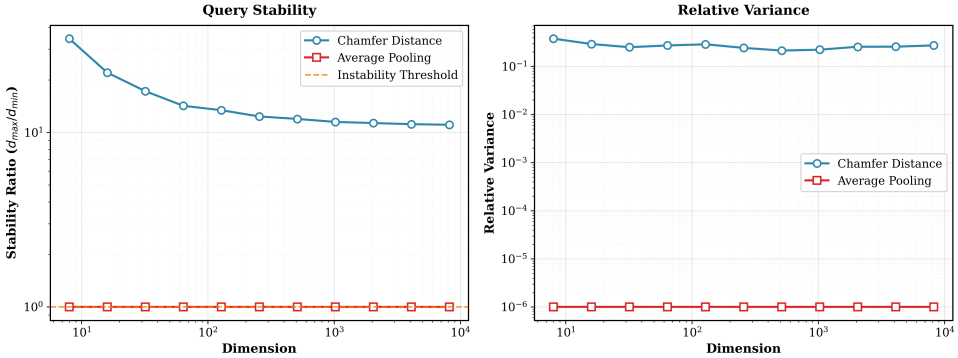

Figure 2: We observe that multi-vector search with Chamfer distance maintains a non-decaying stability ratio and relative variance at high dimensions while average pooling, under the same vector set construction, exhibits instability.

this construction is to generate a synthetic family of vectors sets that satisfy the strong stability, non-degeneracy condition, and non-negative covariances of the theorem while showcasing a case where average pooling exhibits instability while Chamfer aggregation does not. To that end, we generate query and document vectors sampled from a standard multivariate Gaussian distribution and add additional Gaussian noise to each vector. Crucially, for each vector v we add to a document set, we also add the antipodal vector $-v$ to the set. This construction forces the average pooling aggregation with cosine distance to essentially "cancel out" the signal and forcing all set distances to be essentially equivalent. However, Chamfer distance does not suffer from this issue since it will simply select the nearest neighbor document vector for each query vector. Thus, in Figure 2, we present a counterexample where, with all other theorem conditions being satisfied, Chamfer distance exhibits overwhelming stability while averaging results in instability.

4. Conclusion

Modern vector search, driven by advances in neural representation learning, has achieved remarkable success in enabling information retrieval across a variety of modalities. However, the high dimensions of these embeddings suggest that vector retrieval over neural representations should suffer from the curse of dimensionality, thereby rendering approximate search algorithms as ineffective. This paper investigates why sublinear vector retrieval algorithms for multi-vector search succeed in spite of the curse of dimensionality by building off foundational results on near-neighbor search *stability*. We introduce a formal model for multi-vector search and identify sufficient conditions for which search remains stable in this settings. Crucially, we note that the mathematical techniques we present in this

work are general in nature and can very likely be applied to other vector search settings as well. We believe this may constitute an exciting direction for future work.

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in neural information processing systems 33 (2020) 9459–9474.

[2] J.-T. Huang, A. Sharma, S. Sun, L. Xia, D. Zhang, P. Pronin, J. Padmanabhan, G. Ottaviano, L. Yang, Embedding-based retrieval in facebook search, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 2553–2561.

[3] P. Nigam, Y. Song, V. Mohan, V. Lakshman, W. Ding, A. Shingavi, C. H. Teo, H. Gu, B. Yin, Semantic product search, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 2876–2885.

[4] K. Beyer, J. Goldstein, R. Ramakrishnan, U. Shaft, When is "nearest neighbor" meaningful?, in: International conference on database theory, Springer, 1999, pp. 217–235.

[5] R. J. Durrant, A. Kabán, When is 'nearest neighbour'meaningful: A converse theorem and implications, Journal of Complexity 25 (2009) 385–397.

[6] C. C. Aggarwal, A. Hinneburg, D. A. Keim, On the surprising behavior of distance metrics in high dimensional space, in: International conference on database theory, Springer, 2001, pp. 420–434.

[7] M. E. Houle, H.-P. Kriegel, P. Kröger, E. Schubert, A. Zimek, Can shared-neighbor distances defeat the curse of dimensionality?, in: International conference on scientific and statistical database management, Springer, 2010, pp. 482–500.

[8] O. Khattab, M. Zaharia, Colbert: Efficient and effective passage search via contextualized late interaction over bert, in: Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, 2020, pp. 39–48.

[9] M. Faysse, H. Sibille, T. Wu, B. Omrani, G. Viaud, C. Hudelot, P. Colombo, Colpali: Efficient document retrieval with vision language models, in: The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025, 2025.

[10] K. Santhanam, O. Khattab, C. Potts, M. Zaharia, Plaid: an efficient engine for late interaction retrieval, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, 2022, pp. 1747–1756.

[11] J. Engels, B. Coleman, V. Lakshman, A. Shrivastava, Dessert: an efficient algorithm for vector set search with vector set queries, Advances in Neural Information Processing Systems 36 (2023) 67972–67992.

[12] R. Jayaram, L. Dhulipala, M. Hadian, J. D. Lee, V. Mirrokni, Muvera: Multi-vector retrieval via fixed dimensional encoding, Advances in Neural Information Processing Systems 37 (2024) 101042–101073.

[13] A. Hui, B. J. Gao, When is nearest neighbor meaningful: Sequential data, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021, pp. 3103–3106.

[14] V. Pestov, Is the k-nn classifier in high dimensions affected by the curse of dimensionality?, Computers & Mathematics with Applications 65 (2013) 1427–1437.

[15] D. R. Karger, M. Ruhl, Finding nearest neighbors in growth-restricted metrics, in: Proceedings of the thiry-fourth annual ACM symposium on Theory of computing, 2002, pp. 741–750.

[16] A. Gupta, R. Krauthgamer, J. R. Lee, Bounded geometries, fractals, and low-distortion embeddings, in: 44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings., IEEE, 2003, pp. 534–543.

[17] S. Bruch, Foundations of Vector Retrieval, volume 1, Springer, 2024.

[18] A. Razzhigaev, M. Mikhalchuk, E. Goncharova, I. Oseledets, D. Dimitrov, A. Kuznetsov, The shape

of learning: Anisotropy and intrinsic dimensions in transformer-based models, in: Findings of the Association for Computational Linguistics: EACL 2024, 2024, pp. 868–874.
[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
[20] S. Gong, V. N. Boddeti, A. K. Jain, On the intrinsic dimensionality of image representations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3987–3996.
[21] A. Ansuini, A. Laio, J. H. Macke, D. Zoccolan, Intrinsic dimension of data representations in deep neural networks, Advances in Neural Information Processing Systems 32 (2019).
[22] S. Gollapudi, R. Krishnaswamy, B. Landrum, N. Rao, K. Shiragur, S. Silwal, H. Wardhan, α-reachable graphs for multi-vector nearest neighbor search, in: The 1st Workshop on Vector Databases, 2025.
[23] H. G. Barrow, J. M. Tenenbaum, R. C. Bolles, H. C. Wolf, Parametric correspondence and chamfer matching: Two new techniques for image matching, Technical Report, 1977.
[24] S. Jayaram Subramanya, F. Devvrit, H. V. Simhadri, R. Krishnawamy, R. Kadekodi, Diskann: Fast accurate billion-point nearest neighbor search on a single node, Advances in neural information processing Systems 32 (2019).
[25] D. Spielman, Spectral graph theory lecture 3.5: Dan's favorite inequality, https://www.cs.yale.edu/homes/spielman/561/lect03b-18.pdf, 2018. Accessed: 2025-11-24.
[26] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, I. Gurevych, Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models, in: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), ????