# WHEN IS MULTI-VECTOR, FILTERED, AND SPARSE NEAREST NEIGHBOR MEANINGFUL? ON THE STABILITY OF MODERN VECTOR RETRIEVAL

**Anonymous authors**
Paper under double-blind review

## 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 that for many high-dimensional workloads, nearest neighbor search should be inefficient as distances between points become nearly indistinguishable. Prior works have studied the notion of stability in traditional nearest-neighbor search, and identify stable workloads, which are amenable to efficient search. However, no existing research has studied stability in the context of modern formulations of similarity search that have gained traction in practice. Building on foundational results, we extend stability theory to three key retrieval settings: (i) multi-vector search, where we prove that the popular Chamfer distance metric preserves single-vector stability, while average pooling aggregation may destroy it; (ii) filtered vector search, where we show that sufficiently large penalties for mismatched filters can preserve stability; and (iii) sparse vector search, where we formalize and prove novel sufficient stability conditions via the concepts of concentration of importance and overlap of importance. Together, our results extend the classical theory of nearest neighbor meaningfulness to new retrieval settings prevalent in modern practice, offering an explanation for several of the empirically successful design choices behind state-of-the-art vector search systems.

## 1 INTRODUCTION

Nearest-neighbor search over high-dimensional vector embeddings has become a foundational primitive in modern information systems. Driven by advances in neural representation learning, vector databases now routinely index hundreds of millions or even billions of embeddings, consisting of hundreds to thousands of dimensions, and nevertheless serve sub-linear-time approximate nearest-neighbor (ANN) queries at scale. This empirical success is, on its face, surprising: classical results on the *curse of dimensionality* suggest that, in high-dimensional spaces, distances between points should concentrate so tightly that the notion of a "nearest" neighbor loses its meaning, and any sub-linear search algorithm should degrade to a brute-force scan. The seminal work of (Beyer et al., 1999) made this intuition rigorous, introducing a probabilistic model of nearest-neighbor search and showing that, under broad conditions, the ratio of the maximum to the minimum query-to-database distance converges to one as the dimension grows. They termed this phenomenon *instability* and used it to delineate when nearest-neighbor queries are not "meaningful".

A decade later, Durrant and Kabán (2009) established the converse to complete the picture. Whenever the *relative variance* of the distance distribution does not vanish in the limit, the search problem remains *stable*, in the sense that nearest neighbors stay well-separated from the bulk of the database. Beyond this generic criterion, (Durrant and Kabán, 2009) identified concrete sufficient conditions under which stability holds for Euclidean distance; most notably, for the broad family of linear latent variable models, where they showed that distances do not concentrate provided the cumulative contribution of the systematic latent factors grows no slower than the data dimensionality (Theorem C.1). Together, these two results provide a clean necessary-and-sufficient characterization of stability for the classical, single-vector formulation of nearest-neighbor search via the relative variance criterion, and they help explain why neural embeddings, which exhibit precisely this kind



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

of structured, low-noise correlation across dimensions (Table 1), remain searchable via sublinear algorithms despite their high dimensionality.

However, the nearest neighbor search landscape has shifted substantially since these foundational results were established. Practical vector retrieval applications now increasingly go beyond the simple form of a single query vector against a database of single document vectors: (1) queries and documents are now often *sets* of vectors compared under a set-similarity function (Multi-Vector Search); (2) results are often subject to attribute *filters* (Filtered Search); and (3) embeddings themselves are increasingly *sparse* and ultra-high-dimensional (Sparse Search). Whether stability survives the move to these modern settings, and under what conditions, has not, to our knowledge, been studied. In this paper, we address this gap by extending the stability framework of (Beyer et al., 1999) and (Durrant and Kabán, 2009) to these three retrieval settings widely used in practice.

In summary, we make the following contributions in this paper:

- **Multi-vector search:** We prove that multi-vector search with the widely used Chamfer distance is stable whenever the induced single-vector problem is *strongly stable* (stable with a non-vanishing gap between the nearest and farthest neighbors) along with mild non-degeneracy and covariance conditions that we verify empirically on real ColBERT embeddings. In contrast, we construct an explicit counterexample in which average pooling is unstable while Chamfer distance, on the same vector sets, remains stable. Together, these results provide a theoretical justification for the Chamfer aggregation adopted by the popular ColBERT family of models (Khattab and Zaharia, 2020; Santhanam et al., 2022; Faysse et al., 2025).
- **Filtered search:** We analyze filtered vector search in what we call the *distance penalty model*, in which the base distance incurs an additive penalty for documents that violate the query filter. We prove that if the base (unfiltered) problem is stable and the penalty is sufficiently large relative to the maximum base distance and the filter mismatch probability, then the filtered problem is stable, and we verify the theorem's assumption on a popular filtered-search benchmark. We also construct an explicit counterexample showing that a base nearest neighbor problem that is stable can become unstable in the filtered setting with a penalty that is insufficiently large, which motivates the need for our theorem proving an explicit lower bound on the penalty size to guarantee stability.
- **Sparse search:** While the prior two results establish stability preservation, in the sparse setting we establish sufficient conditions for stability directly, since we show that the existing explanation for the stability of dense embeddings does not carry over to sparse vectors (Table 2). We give a new formal definition of *concentration of importance*, a property described intuitively in the seminal work of (Bruch et al., 2024), and introduce a complementary new property we call *overlap of importance*. We prove that the two properties jointly imply stability (under a minor non-degeneracy condition), and we show empirically that neither property suffices on its own. We also verify that both conditions hold on real SPLADE (Formal et al., 2021) embeddings.

**Goals and Non-Goals.** Our goal is to develop a theoretical framework that explains several design decisions in modern vector retrieval that the community arrived at largely through trial and error. Building on the classical understanding of when single-vector search is stable, we ask what happens when retrieval moves beyond this setting: which aggregation functions and filtering schemes preserve stability, and which properties of sparse representations give rise to it? It is not our goal to design new search algorithms, to predict which algorithm will perform best, or to characterize the hardness of a given dataset. Our results concern retrieval formulations rather than particular indexes or datasets, and characterize when meaningful search is possible *in principle*.

### 1.1 Additional Results in Appendix

We defer several results and discussion points to the appendix. In Appendix A, we discuss related work on stability and related concepts in nearest neighbor search theory, namely intrinsic dimensionality and relative contrast. In Appendix B, we demonstrate empirically that stability is not merely a theoretical curiosity but rather has substantial bearing on the quality of vector databases in practice. In Appendix C, we show that the prior theory of Durrant and Kabán (2009) provides an explanation

for why modern dense neural embeddings exhibit stability, but we also demonstrate that this result does not extend to the sparse setting, which motivates our new results. In the remaining Appendices (D, E, and F), we provide full proofs for all of the new mathematical results introduced in this paper, and we empirically verify that the conditions of our theorems hold on real embedding datasets.

## 2 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 2.1** (Nearest-Neighbor Search Problem)**.** A *nearest-neighbor search problem* is a 3-tuple $(Q, D, \delta)$ 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 $\delta \colon \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}_{\geq 0}$ to compute

$$\arg\min_{d \in D} \delta(q, d)$$

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., 1999) introduced this probabilistic model to define the notion of instability.

**Definition 2.2** (Instability (Beyer et al., 1999))**.** A nearest neighbor problem $(Q, D, \delta)$ 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* 2.3. 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 2.4** ((Beyer et al., 1999))**.** *For a nearest neighbor problem $(Q, D, \delta)$, let $q \sim Q$ and define $\mathrm{DMAX} = \max_d \delta(q, d)$ and $\mathrm{DMIN} = \min_d \delta(q, d)$. A nearest neighbor problem is unstable if and only if $\frac{\mathrm{DMAX}}{\mathrm{DMIN}} \to_p 1$ 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 nearest neighbors.

To prove that a nearest-neighbor problem is stable or unstable, we will also stand on the shoulders of two prior theorems from (Beyer et al., 1999) and (Durrant and Kabán, 2009), respectively.

**Theorem 2.5** (Beyer et al. (1999))**.** *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 2.6** (Durrant and Kabán (2009))**.** *Let $(\mathcal{Q}, \mathcal{D}, \delta)$ be a nearest-neighbor search problem with database $\{d_1, \ldots, d_n\} \subset \mathcal{D}$, and for $q \sim \mathcal{Q}$ let $D_{\min} = \min_j \delta(q, d_j)$ and $D_{\max} = \max_j \delta(q, d_j)$. Assume that*

$$D_{\min} \;\leq\; \mathbb{E}[\delta(q, d)] \;\leq\; D_{\max}$$

*holds for all sufficiently large $m$. If $\frac{\mathrm{Var}[\delta(q,d)]}{\mathbb{E}[\delta(q,d)]^2}$ does not converge to $0$ as $m \to \infty$, then the nearest-neighbor problem is stable.*

*Remark* 2.7. For brevity, in our subsequent results that invoke Theorem 2.6, we will refer to the condition $D_{\min} \;\leq\; \mathbb{E}[\delta(q, d)] \;\leq\; D_{\max}$ that the true expectation lies between the empirical min and max as the *bounded expectation* property.

Following the convention established in (Beyer et al., 1999), 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.

Using Theorem 2.6, Durrant and Kabán (2009) also present a theorem establishing a sufficient condition for the vector search stability based on linear latent variable models. Due to space limitations, we defer the statement of the theorem to Appendix C. We also show empirically that dense neural embeddings satisfy the conditions of the above theorem, which explains why single-vector dense search workloads remain stable. However, we also demonstrate that SPLADE sparse vectors do not satisfy the theorem's conditions, which motivates our main result in Section 5

# 3 Stability of Multi-vector Search

## 3.1 Theoretical Model

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

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

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 $\delta$, the *Chamfer Distance* (Barrow et al., 1977) is defined as

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

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

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

Next we 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 $\epsilon > 0$

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

We can now state our first theorem of this section generalizing the results of (Beyer et al., 1999) and (Durrant and Kabán, 2009) to the multi-vector setting.

**Theorem 3.5.** *Suppose that $\mathbb{E}[\mathrm{Agg}(Q, D)]$ is bounded as defined in Remark 2.7. Then, a multi-vector search instance is stable if and only if*

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

We defer the proof of this theorem to Appendix D. 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 $(Q', D', \delta)$ 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 $\delta$ is the primitive distance function of $\mathrm{Agg}$.

### 3.2 Main Result

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. We begin by introducing two new definitions:

**Definition 3.7** (Strong Stability)**.** Let $c > 1$ be a constant. A nearest-neighbor search problem $(Q, D, \delta)$ is $c$-strongly stable if there exists a constant $\gamma > 0$ such that for every query $q \in Q$ and all sufficiently large $m$, $\mathrm{DMAX}/\mathrm{DMIN} \geq c + \gamma$

Intuitively, strong stability asserts that a nearest 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.

**Definition 3.8** ($c$-Non-Degenerate Document Sets)**.** Let $(\mathcal{Q}, \mathcal{D}, \delta, \mathrm{Chamfer})$ be a multi-vector search problem with induced single-vector instance $(\mathcal{Q}', \mathcal{D}', \delta)$, and let $\{D_1, \ldots, D_n\}$ be the document sets in $\mathcal{D}$. For a constant $c > 1$, we say the document sets are *c-non-degenerate* if for every query vector $q \in \mathcal{Q}'$ and all sufficiently large $m$, we have that $c \cdot \max_k \min_{d \in D_k} \delta(q, d) \geq \max_k \max_{d \in D_k} \delta(q, d)$

*Remark* 3.9. $c$-Non-degeneracy asserts that at least one document set lies entirely outside the ball of radius $D_{\max}/c$ around a query $q$, while placing no restriction on how spread out the vectors within a set may be. This rules out groupings in which every document set places a representative near every query vector, which we observe does not occur in practice.

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 3.10** (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 $(Q', D', \delta)$ 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 and the document sets are c-non degenerate, then the multi-vector search instance is stable.*

**Theorem 3.11** (Multi-Vector Stability with Chamfer Distance)**.** *Let $(\mathcal{Q}, \mathcal{D}, \delta, \mathrm{Chamfer})$ 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 $q_i \in Q$. The problem is stable if: (i) the induced single-vector search instance is c-strongly stable; (ii) the document sets satisfy c-non-degeneracy; and (iii) $\sum_{i<j} \mathrm{Cov}(A_i, A_j) \geq 0$.*

We provide the proofs of Lemma 3.10 and Theorem 3.11 in Appendix D. In Appendix D.2, we also verify that all three theorem assumptions comfortably hold true in practice.

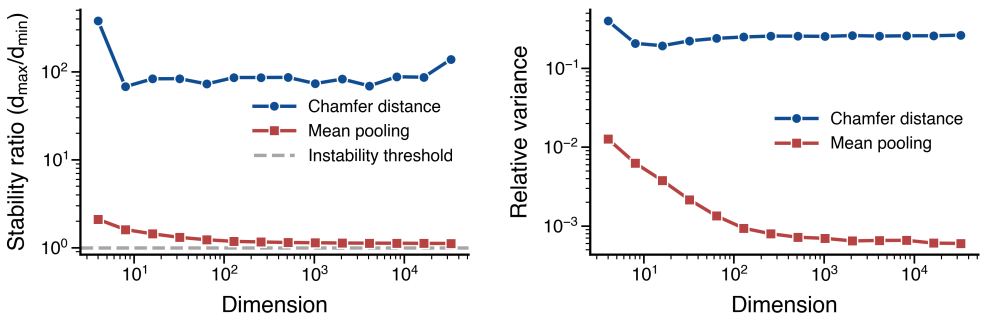



Figure 1: We show that there exists a multi-vector search instance satisfying the assumptions of Theorem 3.11 where Chamfer distance maintains a non-decaying stability ratio and relative variance at high dimensions while average pooling exhibits instability. We provide the details of this construction in Appendix D.3.

# 4 Stability of Filtered Vector Search

## 4.1 Theoretical Model

We begin with a formal definition of the filtered (or hybrid) nearest neighbor search problem which combines traditional vector similarity retrieval with attribute matching.

**Definition 4.1** (Filtered nearest-neighbor Search)**.** A *filtered nearest-neighbor search problem* is a 6-tuple $(\mathcal{Q}, \mathcal{D}, \delta, A_Q, A_D, f)$ where:

- $\mathcal{Q} = \{(q_1, A_{q_1}), (q_2, A_{q_2}), \dots\}$ is a set of query tuples where each $q_i \in \mathbb{R}^m$ and each $A_{q_i} \subseteq A_Q$ is a set of *query attributes*.
- $\mathcal{D} = \{(d_1, A_{d_1}), (d_2, A_{d_2}), \dots\}$ is the set of database tuples where each $d_i \in \mathbb{R}^m$ and each $A_{d_i} \subseteq A_D$ is a set of *document attributes*.
- $\delta : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}_{\geq 0}$ is a standard vector distance metric.
- $f \colon A_Q \times A_D \to \{0, 1\}$ is a *filter function* that outputs 1 if the query and document attributes match and 0 otherwise.

For a given query $(q, A_q)$, we compute

$$\arg\min_{(d, A_d) \in \mathcal{D}} \delta(q, d) \quad \text{subject to } f(A_q, A_d) = 1$$

*Remark* 4.2. This definition of filtered nearest neighbor search places no restriction on how the filter function $f$ operates. For instance, if the query and document attributes are discrete, then a natural definition of $f$ might be to output 1 if $A_q \subseteq A_d$ and 0 otherwise. Furthermore, our definition extends to other filtered search settings. In the *approximate window search* problem Engels et al. (2024), each document is associated with a numeric label, and each query aims to find the nearest documents within arbitrary label ranges. In this case, we can model $A_q \subset \mathbb{R}$ as a continuous range of real numbers and $A_d \in \mathbb{R}$ as a scalar value. The filter function $f$ in this setting would output 1 if $A_d \in A_q$ and 0 otherwise.

In this paper, we will focus our analysis on *penalty-based* filtered nearest-neighbor search algorithms that augment the standard geometric vector distance metric with an additive penalty for points that do not satisfy the query filter. These functions are also referred to as *fusion* distances in the literature Wang et al. (2023). We will define the *penalized* distance function as $\delta'(q, d) = \delta(q, d) + \alpha(1 - f(A_q, A_d))$ so that the distance metric incurs an additive scalar penalty of $\alpha$ if and only if $f(A_q, A_d) = 0$ meaning the query and document attributes do not match.

This formal model of fused distance functions captures nearly all filtered vector search algorithms in the literature. Some approaches, such as NHQ (Wang et al., 2023) and CAPS (Gupta et al., 2023) set the penalty value via heuristics or parameter tuning. On the other hand, many approaches in the literature, such as Acorn (Patel et al., 2024), Filtered DiskANN (Gollapudi et al., 2023), and window search (Engels et al., 2024) never consider points that violate the query filter predicate throughout the entire search process. These algorithms can also be modeled by our framework by setting $\alpha$ to be infinite. However, to our knowledge, no work has theoretically studied how to set this penalty value. In our next theorem, we give an explicit lower bound on the penalty to guarantee the stability of a stable base problem. Furthermore, we show in Figure 2 that there exists a nearest neighbor search problem where an insufficiently large penalty can transform a *stable* base problem into an *unstable* filtered search instance.

### 4.2 Main Result

**Theorem 4.3** (Stability of Filtered Search)**.** *Let $(\mathcal{Q}, \mathcal{D}, \delta', A_Q, A_D, f)$ be a filtered nearest-neighbor search problem with $\delta'(q,d) = \delta(q,d) + \alpha(1 - f(A_q, A_d))$ as the penalized distance function over a base vector distance function $\delta$. Suppose that the base nearest-neighbor search problem $(Q, D, \delta)$ is stable. Furthermore, suppose that the filter mismatch probability for any query-document pair is bounded by $0 < p_{\max} < 1$ for some constant $p_{\max}$ and that the expectation $\mathbb{E}[\delta'(q,d)]$ is bounded. If the penalty value satisfies $\alpha > \frac{2\Delta}{1-p_{\max}}$ where $\Delta = \max_{q,d} \delta(q,d)$, then the filtered search problem is stable under fusion.*

We defer the full proof of this theorem to Appendix E.

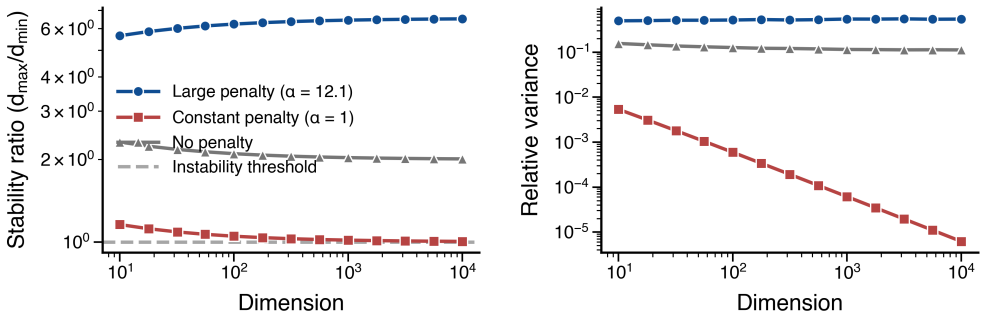



Figure 2: We construct a filtered search instance with a stable base problem such that a sufficiently large mismatch penalty exhibits stability but a small penalty produces instability. This result shows small penalties can actually transform a stable base problem into an unstable filtered search instance, which motivates our result proving a lower bound on the penalty value to guarantee stability. We include the details of our construction in Appendix E.2.

## 5 Stability of Sparse Vector Search

We now turn our attention to the final vector search setting we study in this work: sparse vector retrieval. Unlike the previous two theorems, which give conditions that preserve the stability of an underlying base search problem, here we prove stability conditions for learned sparse retrieval from first principles. As shown in Appendix C, the prior theory that explains why dense neural embeddings exhibit stability (Theorem C.1) do not translate to the sparse setting (Table 2) which motivates the need for new theory, which we develop in this section.

### 5.1 Theoretical Model

In our setup, a sparse vector $d \in \mathbb{R}^m$ is characterized by its support $S_d \coloneqq \{i \in [m] \mid d_i \neq 0\}$. While in many sparse retrieval settings $|S_d| \ll m$ (often $|S_d| = O(\log m)$), for full generality we impose

no growth rate on $|S_d|$ and only rely on the concentration and overlap of importance to establish stability. In other words, we do not restrict ourselves to any particular definition of sparsity.

**Definition 5.1** (Concentration of Importance (CoI))**.** Let $d \in \mathbb{R}^m_{\geq 0}$ and let $d_{(1)} \geq d_{(2)} \geq \cdots \geq d_{(m)}$ be its coordinates sorted by value and denote by $||d||_p$ its $\ell_p$ norm. For a constant $\kappa \ll m$ where $\kappa$ independent of $m$ (i.e., not growing with the dimension), define

$$C_d(\kappa) \coloneqq \frac{\sum_{i=1}^{\kappa} |d_{(i)}|^p}{||d||_p^p} \in [0, 1]$$

For a query set $Q \subset \mathbb{R}^m_{\geq 0}$ and $\alpha \in (0, 1]$, the concentrated queries are $Q_\alpha \coloneqq \{\, q \in Q : C_q(\kappa) \geq \alpha \,\}$.

Let $Q, D \subset \mathbb{R}^m_{\geq 0}$. We say that the pair $(Q, D)$ satisfies concentration of importance with parameters $(\kappa, \alpha, R, \rho)$ if

(a) $Q = Q_\alpha$, i.e., $C_q(\kappa) \geq \alpha$ for all $q \in Q$.

(b) $\Pr_{d\sim D}\left[C_d(R\kappa) \geq \alpha\right] \geq \rho$.

(c) $R\kappa \ll m$, where $R \geq 1$ is fixed and independent of $m$.

*Remark* 5.2. Notice that, by construction, $(Q_\alpha, D)$ satisfies $(a)$ with the same parameters. In practice a small fraction of queries (e.g., long, multi-part queries) may spread their mass over more than $\kappa$ coordinates. We state our results for the concentrated queries $Q_\alpha$ and report the retained fraction $|Q_\alpha|/|Q|$ in our experiments. In addition, restricting to nonnegative vectors entails no loss of generality. We provide a full proof of this claim in Lemma F.1. We state our results for non-negative vectors for notational simplicity.

**Definition 5.3** (Overlap of Importance)**.** Fix $p \geq 1$ and $0 < \kappa \ll m$. For $d \in \mathbb{R}^m_{\geq 0}$, let $T_d$ be the indices of the top-$R\kappa$ coordinates of $d$ by magnitude. We say a $(Q, D, \delta)$ sparse vector search problem exhibits *overlap of importance* if there exist constants $\gamma > 0$ and $\pi \in (0, 1)$ such that, for $q \in Q$ and $d \in D$ with $\|q\|_p = \|d\|_p = 1$,

$$\Pr\left[\sum_{i \in T_q \cap T_d} \min\{\, q_i^p,\, d_i^p \,\} \;\geq\; \gamma\right] \;\geq\; \pi.$$

### 5.2 Main Result

Given the sparse vector search model introduced in the previous section, we now state our main result via concentration and overlap of importance.

**Theorem 5.4.** *Let $(Q, D, ||\cdot||_p)$ be a nearest neighbor search problem where $Q, D \subset \mathbb{R}^m_{\geq 0}$ with $||q||_p = ||d||_p = 1$ for all $q \in Q, d \in D$ and $p \geq 1$. Fix $\alpha > 1/2$, let $Q_\alpha$ be the concentrated queries from definition 5.1, and let $(q, d) \sim P_m$ be a random query-document pair where $P_m$ is the joint distribution.*

*Assume that*

*(a) The query marginal of $P_m$ is supported on $Q_\alpha$, i.e., $C_q(\kappa) \geq \alpha$ for every query.*

*(b) $(Q_\alpha, D)$ satisfies overlap of importance with parameters $(\gamma, \pi)$ as stated in Definition 5.3.*

*(c) Let $B$ be the event defined by $B \coloneqq \{T_q \cap T_d = \emptyset\} \cap \{C_d(R\kappa) \geq \alpha\}$ and assume that there exists $\beta > 0$ such that $\Pr(B) \geq \beta$.*

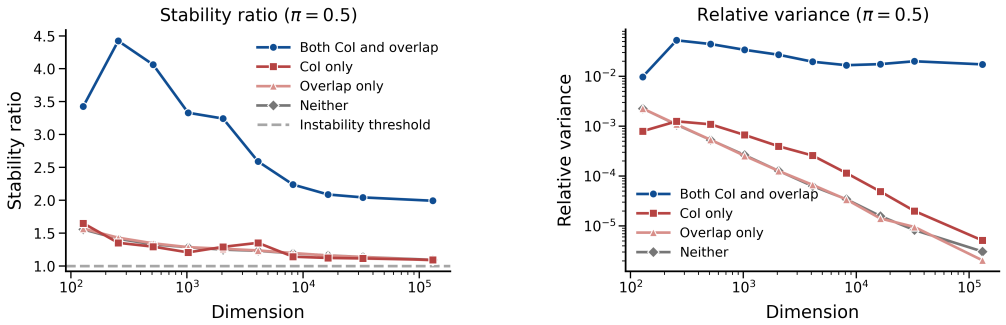
*Under assumptions $(a)$–$(c)$ above, define $X \coloneqq (2 - 2\gamma)^{1/p}$ and $Y \coloneqq 2^{1/p}\left(\alpha^{1/p} - (1-\alpha)^{1/p}\right)$ and assume that $Y > X$. Then, $\liminf_{m\to\infty} \mathrm{RelVar}_m \geq \frac{\pi\beta}{2^{2/p}}(Y - X)^2 > 0$.*

We defer the full proof of this theorem to Appendix F. This result is, to our knowledge, the first to prove sufficient conditions for stability in the context of sparse vector search.

The above theorem demonstrates that sparse vector search is stable whenever $X = (2 - 2\gamma)^{1/p}$ is less than $Y = 2^{1/p}\left(\alpha^{1/p} - (1-\alpha)^{1/p}\right)$. These two expressions represent different geometric

configurations for the vectors $q$ and $d$. The quantity $X$ is an upper bound on the expected distance when $q$ and $d$ overlap substantially in their head coordinates. In this case, the vectors partially align and the distance shrinks. In contrast, $Y$ represents the expected distance in the opposite configuration where $T_q$ and $T_d$ are disjoint. Here the vectors place their largest coordinates in opposite directions, and hence the distance grows.

In Appendix F.2, we empirically verify that the conditions of Theorem 5.4 hold true in practice on real SPLADE embeddings.




(a) Stability ratio

(b) Relative variance

Figure 3: We observe that search over sparse embeddings that violate either concentration or overlap of importance or both exhibits instability. However, when we fulfill both conditions we reach stability. We include the description of this construction in Appendix F.3.

## 6 Conclusion

Vector search has become widely adopted with the advent of neural representation learning, but the question of when nearest-neighbor search is meaningful remains as fundamental as when Beyer et al. (1999) first posed it. In this paper, we extended the stability framework to multi-vector, filtered, and sparse vector search, identifying sufficient conditions under which nearest-neighbor search remains stable, and thus meaningful, in each. We close with three unifying insights for future work on stability theory.

**Stability as an explanatory tool:** Chamfer distance, concentration of importance, and large filter penalties are design choices or embedding properties the community converged upon largely through trial and error; our results show that each is favorable for stability, and thus for efficient search.

**A reusable decomposition for proving stability preservation:** A recurring technique in this paper is to assume the underlying single-vector problem is stable, as prior work shows how to achieve, and identify conditions (Chamfer aggregation and large penalties, respectively) that preserve this stability in a richer setting. Despite their very different formulations, our multi-vector and filtered proofs share a core maneuver: we decompose the relative variance into a fraction of sums and apply a well-known bound on sums of ratios (Lemma D.1) to obtain a simple lower bound. This pattern may serve as a template for proving other "preservation"-style stability results.

**Stability as an analytical tool beyond embedding geometry:** To our knowledge, this paper is the first to demonstrate that the stability framework of (Beyer et al., 1999) and (Durrant and Kabán, 2009) can yield insights into nearest-neighbor search beyond the level of vector geometry. Our filtered search theorem, for example, establishes stability preservation from properties of the fused distance function, not from how the vectors are positioned in space, a distinction invisible to frameworks such as intrinsic dimensionality.

## AI DISCLOSURE

In this work, we used generative AI tools for providing feedback on the correctness of our mathematical proofs and for suggesting proof strategies. We have not used generative AI tools to design or provide feedback on our research methodology or experiments or interpreting results. Additionally, we used generative AI tools for creating and editing our experimental code and to edit the writing in this paper. We have reviewed all AI-assisted work by having all authors review each proof for correctness and having at least one author review the output from AI coding tools. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

## REFERENCES

Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim. On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory*, pages 420–434. Springer, 2001.

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Harry G Barrow, Jay M Tenenbaum, Robert C Bolles, and Helen C Wolf. Parametric correspondence and chamfer matching: Two new techniques for image matching. Technical report, 1977.

Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In *International conference on database theory*, pages 217–235. Springer, 1999.

Sebastian Bruch, Franco Maria Nardini, Cosimo Rulli, and Rossano Venturini. Efficient inverted indexes for approximate retrieval over learned sparse representations. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 152–162. ACM, 2024. doi: 10.1145/3626772.3657769. URL https://doi.org/10.1145/3626772.3657769.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *CoRR*, 2024.

Robert J Durrant and Ata Kabán. When is 'nearest neighbour'meaningful: A converse theorem and implications. *Journal of Complexity*, 25(4):385–397, 2009.

Joshua Engels, Benjamin Landrum, Shangdi Yu, Laxman Dhulipala, and Julian Shun. Approximate nearest neighbor search with window filters. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre 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.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292, 2021.

Siddharth Gollapudi, Neel Karia, Varun Sivashankar, Ravishankar Krishnaswamy, Nikit Begwani, Swapnil Raz, Yiyong Lin, Yin Zhang, Neelam Mahapatro, Premkumar Srinivasan, et al. Filtered-diskann: Graph algorithms for approximate nearest neighbor search with filters. In *Proceedings of the ACM Web Conference 2023*, pages 3406–3416, 2023.

Sixue Gong, Vishnu Naresh Boddeti, and Anil K Jain. On the intrinsic dimensionality of image representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3987–3996, 2019.

Anupam Gupta, Robert Krauthgamer, and James R Lee. Bounded geometries, fractals, and low-distortion embeddings. In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, pages 534–543. IEEE, 2003.

Gaurav Gupta, Jonah Yi, Benjamin Coleman, Chen Luo, Vihan Lakshman, and Anshumali Shrivastava. Caps: A practical partition index for filtered similarity search. *arXiv preprint arXiv:2308.15014*, 2023.

Junfeng He, Sanjiv Kumar, and Shih-Fu Chang. On the difficulty of nearest neighbor search. *arXiv preprint arXiv:1206.6411*, 2012.

Michael E Houle, Hans-Peter Kriegel, Peer Kröger, Erich Schubert, and Arthur Zimek. Can shared-neighbor distances defeat the curse of dimensionality? In *International conference on scientific and statistical database management*, pages 482–500. Springer, 2010.

Aaron Hui and Byron J Gao. When is nearest neighbor meaningful: Sequential data. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3103–3106, 2021.

Patrick Iff, Paul Bruegger, Marcin Chrapek, Maciej Besta, and Torsten Hoefler. Benchmarking filtered approximate nearest neighbor search algorithms on transformer-based embedding vectors, 2025. URL https://arxiv.org/abs/2507.21989.

David R Karger and Matthias Ruhl. Finding nearest neighbors in growth-restricted metrics. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 741–750, 2002.

Omar Khattab and Matei 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*, pages 39–48, 2020.

Carlos Lassance, Hervé Déjean, Thibault Formal, and Stéphane Clinchant. Splade-v3: New baselines for splade. *arXiv preprint arXiv:2403.06789*, 2024.

Yu A Malkov and Dmitry A Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836, 2018.

Liana Patel, Peter Kraft, Carlos Guestrin, and Matei Zaharia. Acorn: Performant and predicate-agnostic search over vector embeddings and structured data. *Proceedings of the ACM on Management of Data*, 2(3):1–27, 2024.

Vladimir Pestov. Is the k-nn classifier in high dimensions affected by the curse of dimensionality? *Computers & Mathematics with Applications*, 65(10):1427–1437, 2013.

Anton Razzhigaev, Matvey Mikhalchuk, Elizaveta Goncharova, Ivan Oseledets, Denis Dimitrov, and Andrey Kuznetsov. The shape of learning: Anisotropy and intrinsic dimensions in transformer-based models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 868–874, 2024.

Keshav Santhanam, Omar Khattab, Christopher Potts, and Matei Zaharia. Plaid: an efficient engine for late interaction retrieval. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1747–1756, 2022.

Daniel Spielman. Spectral graph theory lecture 3.5: Dan's favorite inequality. https://www.cs.yale.edu/homes/spielman/561/lect03b-18.pdf, September 2018. Accessed: 2025-11-24.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna 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)*.

Mengzhao Wang, Lingwei Lv, Xiaoliang Xu, Yuxiang Wang, Qiang Yue, and Jiongkang Ni. An efficient and robust framework for approximate nearest neighbor search with attribute constraint. *Advances in Neural Information Processing Systems*, 36:15738–15751, 2023.

# A Related Work

## A.1 Conditions for Stability

The paper of (Beyer et al., 1999) was a foundational work that inspired significant follow-up research in understanding the conditions for stable vector search. (Aggarwal et al., 2001) study stability under common $\ell_p$ norm distance metrics, where they find that $\ell_1$ and fractional distance metrics are theoretically and practically favorable to $\ell_2$ distances. Furthermore, (Houle et al., 2010) prove that the so-called *shared-neighbor* distance function, which measures the number of nearest-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, 2021) 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, 2013) extends the results of (Beyer et al., 1999) 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 nearest-neighbor classification.

## A.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 nearest-neighbor search over these points is insulated from the curse of dimensionality effect. Consequently, ID has been well-studied with multiple proposed formal definitions (Karger and Ruhl, 2002; Gupta et al., 2003). 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 (Razzhigaev et al., 2024; Gong et al., 2019; Ansuini et al., 2019).

Given that neural models consistently exhibit low intrinsic dimensionality, a natural question to ask is whether stability is still a relevant concept to study, or if ID alone explains the success of vector retrieval in practice. In this paper, we show that 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. Similarly, our filtered search results suggest that practitioners can preserve stability through sufficiently large penalties, even when the underlying embeddings are intrinsically high-dimensional.

## A.3 Relative Contrast

In another vein, (He et al., 2012) study an alternative theoretical framework to stability for finite-dimensional vector search instances called relative contrast (RC). RC derives non-asymptotic guarantees on nearest-neighbor meaningfulness as a function of database size, dimensionality, and sparsity which can be used to measure the hardness of a given vector search dataset. Our stability differs from RC primarily in the fact that our results are asymptotic in nature as opposed to finite-dimensional. Extending our results to such a framework is not straightforward. RC relies on a central-limit-theorem-based Gaussian approximation that assumes independent coordinates and an additive-over-coordinates distance function, neither of which holds in the settings we consider. Set-aggregation distances such as Chamfer break additivity through the inner min operation, the penalty model in filtered search introduces a non-$\ell_p$ distance, and the concentration-of-importance structure in sparse search is precisely a model of correlated, non-i.i.d. coordinates. Developing a non-asymptotic version of our stability conditions for the multi-vector, filtered, and sparse settings in the spirit of RC is an exciting direction for future work.

# B Practical Implications of Stability

In this section, we empirically demonstrate that stability is not merely a theoretical curiosity but also has substantial bearing on the quality of vector retrieval algorithms in practice. This is an empirical demonstration of the effects of instability on popular approximate ANN search algorithms.

In Figure 4, we report the results of running two of the most popular ANN search techniques, namely the hierarchical navigable small worlds (HNSW) graph-based index (Malkov and Yashunin, 2018) and the inverted file (IVF) partitioning-based index (Douze et al., 2024), across stable and unstable datasets.

We fix a standard configuration of hyperparameters for each respective algorithm and compute the top-10 recall by taking the average of 1000 query vectors over a database of 1 million document vectors. To generate the unstable datasets at varying dimensionalities, we sample the value of each component independently from a univariate standard Gaussian distribution following the classic construction of (Beyer et al., 1999). To construct the stable datasets, we generate vectors such that they belong to one of five distinct clusters. For consistency, we fix the search index hyperparameters across all experiments; for HNSW, we set `ef-construction` and `ef-search` to be 200 and `M` $= 16$; for IVF, we set the number of clusters to `nlist` $= \sqrt{n}$, where $n$ is the number of database elements and the number of probes to `nprobes` $= \frac{\sqrt{n}}{4}$.

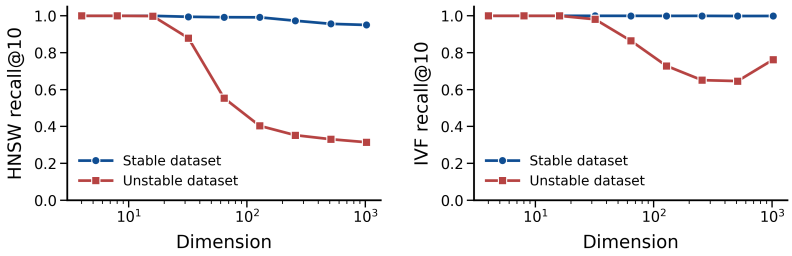



Figure 4: Comparison of stable vs. unstable recall@10 across algorithms and dimensions. We observe the instability causes approximate search algorithms to lose recall in higher dimensions while stability preserves search quality as the dimensionality scales.

From Figure 4, we see that these approximate nearest-neighbor search algorithms severely suffer in recall over unstable databases since the lack of contrast between points renders sublinear algorithms that avoid a brute force scan ineffective. We also see that the impact of instability worsens in higher dimensions. These results underscore the importance of stability in practice.

## C Stability of Dense Single-Vector Embeddings

**Theorem C.1** (Durrant and Kabán (2009))**.** *Let $(\mathcal{Q}, \mathcal{D}, \delta)$ be a nearest-neighbor search problem with $\delta(q, d) = \|q - d\|_2$, where queries $q \sim \mathcal{Q}$ and documents $d \sim \mathcal{D}$ are drawn independently from the same linear latent variable model: for a fixed number of latent factors $L$ (independent of $m$), each $x \in \mathbb{R}^m$ is generated as*

$$x_i = \sum_{l=1}^{L} w_{il}\, y_l + \varepsilon_i, \qquad i = 1, \ldots, m,$$

*where*

- *(i) the coefficients $w_{il} \in \mathbb{R}$ are constants independent of $m$ with $\sup_{i,l} |w_{il}| < \infty$;*
- *(ii) the latent factors $y_1, \ldots, y_L$ are mutually independent, non-degenerate, and have finite fourth moments;*
- *(iii) the noise terms $\varepsilon_1, \ldots, \varepsilon_m$ are mutually independent, independent of $(y_1, \ldots, y_L)$, have zero mean, and satisfy $\sup_i \mathbb{E}[\varepsilon_i^4] < \infty$.*

*If there exists a latent factor $l$ whose cumulative contribution grows linearly with the dimension, i.e.,*

$$\limsup_{m \to \infty} \frac{1}{m} \sum_{i=1}^{m} w_{il}^2 > 0,$$

| Corpus | Model | $m$ | $L$ | Relevant coords (%) |
|---|---|---|---|---|
| SciFact | bge-small-en-v1.5 | 384 | 256 | 100.0 |
| | all-MiniLM-L6-v2 | 384 | 256 | 99.2 |
| | e5-base-v2 | 768 | 512 | 99.9 |
| NFCorpus | bge-small-en-v1.5 | 384 | 256 | 100.0 |
| | all-MiniLM-L6-v2 | 384 | 256 | 99.2 |
| | e5-base-v2 | 768 | 512 | 99.9 |
| ArguAna | bge-small-en-v1.5 | 384 | 256 | 100.0 |
| | all-MiniLM-L6-v2 | 384 | 256 | 99.2 |
| | e5-base-v2 | 768 | 512 | 99.9 |

Table 1: Dense embeddings satisfy the linear latent variable condition of Theorem C.1. We use a PPCA fit $x = Ay + \delta$, $\delta \sim \mathcal{N}(0, \sigma^2 I)$, with $L$ chosen by held-out log-likelihood ($n = 4{,}000$ documents per corpus, 3,633 for NFCorpus). *Relevant coords*: fraction of coordinates whose systematic variance $\sum_l a_{il}^2$ exceeds the noise variance $\sigma^2$.

| Corpus | Model | $m$ | $L$ | Relevant coords (%) |
|---|---|---|---|---|
| SciFact | SPLADE | 15,697 | 512 | 17.5 |
| NFCorpus | SPLADE | 15,336 | 256 | 12.0 |
| ArguAna | SPLADE | 19,279 | 256 | 9.9 |

Table 2: SPLADE Formal et al. (2021) embeddings violate the linear latent variable condition of Theorem C.1. We define $m$ as the active vocabulary (coordinates non-zero in at least one document) out of $30{,}522$; measured over the full vocabulary, the relevant fraction is 6.0–9.0%.

*then* $\mathrm{Var}[\delta(q,d)^2]/\mathbb{E}[\delta(q,d)^2]^2$ *does not converge to* $0$ *as* $m \to \infty$. *Consequently, for sufficiently large databases, the nearest-neighbor problem is stable.*

Intuitively, Theorem C.1 states that distances do not concentrate as long as the number of "relevant" coordinates, meaning those carrying signal from the systematic latent factors rather than pure noise, grows no slower than the ambient dimension $m$.

# D Stability of Multi-Vector Search

## D.1 Proofs

***Theorem 3.5*** *For $Q \sim \mathcal{Q}$, let $D_{\min} = \min_j \mathrm{Agg}(Q, D_j)$ and $D_{\max} = \max_j \mathrm{Agg}(Q, D_j)$. Assume that $D_{\min} \leq \mathbb{E}[\mathrm{Agg}(Q, D)] \leq D_{\max}$ holds for all sufficiently large $m$. A multi-vector search instance is stable if and only if*

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

*Proof.* This result follows directly from the prior proofs of (Beyer et al., 1999) and (Durrant and Kabán, 2009) since the original proofs 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 $\mathrm{Agg}(Q, D)$ be the random variable for the aggregated distance in an $m$-dimensional space. Let its expected value be $\mathbb{E}[\mathrm{Agg}(Q, D)]$ and its variance be $\mathrm{Var}[\mathrm{Agg(Q, D)}]$. The database produces $n$ samples of this random variable. Let $\mathrm{DMIN} = \min_{D \in \mathcal{D}} \mathrm{Agg}(Q, D)$ and $\mathrm{DMAX} = \max_{D \in \mathcal{D}} \mathrm{Agg}(Q, D)$.

$\Longrightarrow$ This direction follows the argument from (Beyer et al., 1999). We will show that if $\lim_{m \to \infty} \frac{\mathrm{Var}[\mathrm{Agg}(Q,D)]}{\mathbb{E}[\mathrm{Agg}(Q,D)]^2} = 0$, then the multi-vector search problem is unstable. Define a normalized random variable $Y = \frac{\mathrm{Agg}(Q,D)}{\mathbb{E}[\mathrm{Agg}(Q,D)]}$. 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 $\min Y_i \to_p 1$ and $\max Y_i \to_p 1$. Since both the numerator and denominator converge in probability to 1, their ratio also converges in probability to 1. By definition, $\frac{\text{DMAX}}{\text{DMIN}} \to_p 1$ 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.

$\Longleftarrow$ This converse result follows the proof presented by (Durrant and Kabán, 2009). We proceed by proving the contrapositive: If the multi-vector search problem is unstable, then the relative variance must be zero. 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: $\frac{\text{DMAX}}{\text{DMIN}} \to_p 1$. For a sufficiently large database, we note 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:

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

This completes the proof of the contrapositive. □

***Lemma 3.10****[Chamfer Stability with Singleton Query Sets] Let $(\mathcal{Q}, \mathcal{D}, \delta, \text{Chamfer})$ be a multi-vector search problem where $|Q_i| = 1$ for each query set $Q_i \in \mathcal{Q}$. Let $(Q', D', \delta)$ 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 D.1** (Mediant Inequality (Spielman, 2018)). *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 3.11*** *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 $q_i \in Q$. The problem is stable if: (i) the induced single-vector search instance is $c$-strongly stable; (ii) the document sets satisfy $c$-non-degeneracy; and (iii) $\sum_{i<j} \text{Cov}(A_i, A_j) \geq 0$.*

*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 D.1 and Lemma 3.10 respectively. This completes the proof.

□

*Remark* D.2. Stability follows from applying the mediant inequality (Lemma D.1) to reduce the general multi-vector search problem to multi-vector search over singleton query sets, namely the setting in Lemma 3.10, at the cost of an $\frac{1}{k}$ overhead factor.

### D.2 Empirical Validation of Multi-Vector Theorem Assumptions

Theorem 3.11 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 (Khattab and Zaharia, 2020)

to construct query and document vector sets for five standard English-based information retrieval datasets from the BEIR benchmark (Thakur et al.). 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 nearest-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 3 where we see that all three theorem conditions are comfortably satisfied in practice.

| **Dataset** | **Stability Constant** ($c$) | **Non-Degeneracy Pass Rate (%)** | **Cov. Bound Pass Rate (%)** |
|---|---|---|---|
| MS MARCO[1] | 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 3: Validation of Theorem 3.11's assumptions across multiple datasets. All conditions are satisfied across all datasets, demonstrating the practical relevance of our theoretical results.

### D.3 Multi-Vector Search Counterexample Construction

To illustrate Theorem 3.11 and show a separation between Chamfer distance and average pooling for stability, we construct a family of multi-vector search instances whose dimensionality $m$ can be taken arbitrarily large while the conditions of the theorem continue to hold. All vectors lie on the unit sphere $\mathbb{S}^{m-1} \subset \mathbb{R}^m$ and we use cosine distance $\delta(x, y) = 1 - \langle x, y \rangle$.

**Generative model.** Query and document vectors are drawn from a mixture of $K$ locally low-rank components. The construction is parameterized by the ambient dimension $m$, the number of mixture components $K$, the query set size $N_q$, the document set size $N_d$ (with $N_q, N_d \leq K$), the local rank $r \ll m$, and an offset scale $\sigma > 0$.

1. **Modes.** Draw $K$ modes $\mu_1, \ldots, \mu_K$ independently and uniformly from $\mathbb{S}^{m-1}$ by sampling $g_k \sim \mathcal{N}(0, I_m)$ and setting $\mu_k = g_k / \|g_k\|_2$.
2. **Local subspaces.** For each mode $k$, draw a random matrix $M_k \in \mathbb{R}^{m \times r}$ with orthonormal columns (e.g., by orthonormalizing an $m \times r$ standard Gaussian matrix). The column space of $M_k$ is a randomly oriented $r$-dimensional subspace: a line through $\mu_k$ when $r = 1$, a disc when $r = 2$, and so on.
3. **Sampling a point from mode $k$.** Draw $z \sim \mathcal{N}(0, \sigma^2 I_r)$ and return
$$x = \frac{\mu_k + M_k z}{\|\mu_k + M_k z\|_2}.$$
Because $M_k$ has orthonormal columns, the offset norm $\|M_k z\|_2 = \|z\|_2$ has a distribution independent of $m$.
4. **Document sets.** Each document set is generated by choosing a subset of $N_d$ distinct modes uniformly at random from the $\binom{K}{N_d}$ possibilities and drawing one point from each chosen mode.

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

5. **Query sets.** Each query set is generated in the same way from a uniformly random subset of $N_q$ distinct modes.

In Figure 1 we use $n = 1000$ document sets, 100 query sets, and $N_q = N_d = 4$.

## E Stability of Filtered Vector Search

**Theorem E.1** (Stability of Filtered Search)**.** *Let $(\mathcal{Q}, \mathcal{D}, \delta', A_Q, A_D, f)$ be a filtered nearest-neighbor search problem with $\delta'(q, d) = \delta(q, d) + \alpha(1 - f(A_q, A_d))$ as the penalized distance function over a base vector distance function $\delta$. Suppose that the base nearest-neighbor search problem $(Q, D, \delta)$ is stable. Furthermore, suppose that the filter mismatch probability for any query-document pair is bounded by $0 < p_{\max} < 1$ for some constant $p_{\max}$ and that the expectation $\mathbb{E}[\delta'(q, d)]$ is bounded. If the penalty value satisfies $\alpha > \frac{2\Delta}{1-p_{\max}}$ where $\Delta = \max_{q,d} \delta(q, d)$, then the filtered search problem is stable under fusion.*

*Proof.* As a shorthand, let $\delta = \delta(q, d)$ and let the indicator variable $1_c = \mathbf{1}\{f(A_q, A_d) = 0\}$ denote a filter mismatch, and define $p_c = \mathbb{E}[1_c]$. In the remainder of the proof, we will assume $p_c > 0$ since if $p_c = 0$ then the filtered search problem reduces to the base nearest neighbor search problem which, by assumption, is stable.

Since the base problem $(Q, D, \delta)$ is stable, the contrapositive of Theorem 2.5 implies that $\mathrm{Var}[\delta]/\mathbb{E}[\delta]^2$ does not converge to 0; hence there exist a constant $v > 0$ and an infinite subsequence of dimensions along which $\mathrm{Var}[\delta] \geq v\mathbb{E}[\delta]^2$. We restrict attention to this subsequence for the remainder of the proof; all subsequent bounds hold along it. We will analyze the limiting behavior of the relative variance of the fused distance function $\delta'$ denoted by $\mathrm{RelVar}' = \frac{\mathrm{Var}[\delta']}{\mathbb{E}[\delta']^2}$. We will proceed by bounding the numerator and denominator terms separately.

We first observe that $\mathrm{Var}[\delta'] = \mathrm{Var}[\delta] + \alpha^2\mathrm{Var}[1_c] + 2\alpha\mathrm{Cov}(\delta, 1_c)$. Since $\delta 1_c \geq 0$ and $E[\delta] \leq \Delta$, it follows that

$$\mathrm{Cov}(\delta, 1_c) = \mathbb{E}[\delta 1_c] - \mathbb{E}[\delta]p_c \geq -\Delta p_c$$

and

$$\mathrm{Var}[1_c] = p_c(1 - p_c) \geq p_c(1 - p_{\max})$$

Therefore,

$$\mathrm{Var}[\delta'] \geq \mathrm{Var}[\delta] + \alpha p_c(\alpha(1 - p_{\max}) - 2\Delta) = \mathrm{Var}[\delta] + \alpha g p_c$$

where we define $g = \alpha(1 - p_{\max}) - 2\Delta$. We note that $g > 0$ by our assumption that $\alpha > \frac{2\Delta}{1-p_{\max}}$.

Next, for the denominator term, we note that

$$\mathbb{E}[\delta']^2 = (\mathbb{E}[\delta] + \alpha p_c)^2 \leq 2\mathbb{E}[\delta]^2 + 2\alpha^2 p_c$$

where the inequality follows from the fact that $(x + y)^2 \leq 2x^2 + 2y^2$ for all $x, y \in \mathbb{R}$ and that fact that $p_c^2 \leq p_c$ since $p_c \in [0, 1]$. Combining our bounds on the numerator and denominator, we have that

$$\mathrm{RelVar}' \geq \frac{v\mathbb{E}[\delta]^2 + \alpha g p_c}{2\mathbb{E}[\delta]^2 + 2\alpha^2 p_c}$$

Applying Lemma D.1, we have

$$\mathrm{RelVar}' \geq \min\left\{\frac{v}{2}, \frac{g}{2\alpha}\right\} > 0$$

Since this bound holds along the chosen subsequence and $\min\{\frac{v}{2}, \frac{g}{2\alpha}\} > 0$ is a constant independent of the dimension, the relative variance of $\delta'$ does not converge to 0 as $m \to \infty$. By Theorem 2.6, the filtered search problem is therefore stable.

□

### E.1 Empirical Validation of Filtered Search Theorem Assumption

Theorem E.1 relies on an assumption that the filter function is *discriminatory* meaning that the probability of a filter mismatch $p_{\max}$ is strictly between 0 and 1. In other words, for every filter predicate, there exist points that satisfy and do not satisfy the filter. We now validate that this assumption holds on a standard filtered vector search benchmark. We consider the popular Arxiv filtered search dataset (Iff et al., 2025) and specifically consider the categorical filters. We report our results in Table 4 across the low, medium, and high variants of the Arxiv dataset where we see that the theorem conditions indeed hold.

| Dataset | $p_{\max}$ | $p_{\text{mean}}$ | $\alpha$ |
|---|---|---|---|
| ArXiv [low] | 0.9565 | 0.9565 | 92.023 |
| ArXiv [medium] | 0.9565 | 0.9261 | 92.023 |
| ArXiv [high] | 0.8894 | 0.4398 | 36.160 |

Table 4: Validation of Theorem E.1's assumption of discriminatory filters on the Arxiv filtered datasets with categorial filters.

### E.2 Filtered Search Stability Simulation

Fix a penalty $\alpha > 0$. For each dimension $m$, the query is the origin $q = 0 \in \mathbb{R}^m$ with attribute set $A_q = \{\mathsf{far}\}$. A document is generated by drawing a radius $r \in \{1, \Delta_m\}$ uniformly at random, where $\Delta_m := 1+\alpha+m^{-1/2}$, and then sampling $d$ uniformly from the sphere of radius $r$ in $\mathbb{R}^m$. Documents on the outer shell carry $A_d = \{\mathsf{far}\}$ and documents on the inner shell carry $A_d = \emptyset$. With Euclidean distance and the containment filter $f(A_q, A_d) = \mathbf{1}\{A_q \subseteq A_d\}$, every document lies at distance exactly $r$ from the query, inner-shell documents fail the filter, and $\mathbf{1}_c = \mathbf{1}\{r = 1\}$, so $p_c = 1/2$ for every $m$. The assumptions of Theorem E.1 hold with constants independent of $m$: distances are bounded by $\Delta := 2 + \alpha$, the mismatch rate satisfies $p_c \le p_{\max}$ for any $p_{\max} \in [1/2, 1)$, and the base relative variance $(\Delta_m - 1)^2/(\Delta_m + 1)^2$ converges to $v = \alpha^2/(2+\alpha)^2 > 0$, so the base problem is stable. The penalty $\alpha$ is below the threshold $2\Delta/(1 - p_{\max}) \ge 4(2 + \alpha)$. Applying it lifts the inner shell to distance $1 + \alpha$, while the outer shell sits at $1 + \alpha + m^{-1/2}$. The fused distance therefore takes two values that are $m^{-1/2}$ apart, which gives $\mathrm{DMAX}'/\mathrm{DMIN}' \le 1 + m^{-1/2}/(1 + \alpha) \to 1$ and $\mathrm{RelVar}' = \Theta(1/m)$.

## F Stability of Sparse Vector Search

### F.1 Proofs

Recall that we are assuming $q, d \in \mathbb{R}^m_{\ge 0}$ and are normalized (i.e., $||q||_d = ||d||_d = 1$). We will start with the following lemma, which we use to assert that restricting our setting to $\mathbb{R}^m_{\ge 0}$ entails no loss of generality.

**Lemma F.1.** *Let $\phi : \mathbb{R}^m \to \mathbb{R}^{2m}$ be given by $\phi(x) = (x^+, x^-)$ with $x_i^+ = \max\{x_i, 0\}$ and $x_i^- = \max\{-x_i, 0\}$. Then, for all $x, y \in \mathbb{R}^m$ and $p \ge 1$,*

$$2^{-(1-1/p)}||x - y||_p \le ||\phi(x) - \phi(y)||_p \le 2^{1/p}||x - y||_p.$$

*Proof.* For $x, y \in \mathbb{R}^m$, define $\Delta_i \coloneqq |x_i - y_i|$, $A_i \coloneqq |x_i^+ - y_i^+|$, $B_i \coloneqq |x_i^- - y_i^-|$.

We will first show that for scalars $u, v \in \mathbb{R}$ and $p \ge 1$

$$2^{-(p-1)}|u - v|^p \le |u^+ - v^+|^p + |u^- - v^-|^p \le 2|u - v|^p. \tag{1}$$

To see why, set $a = u^+ - v^+$ and $b = u^- - v^-$ and observe that $u - v = a - b$. Using Hölder's inequality in $\mathbb{R}^2$ with conjugate exponents $p$ and $q = p/(p-1)$, we have that

$$|u - v| = |\langle (a, b), (1, -1)\rangle| \le ||(a, b)||_p||(1, -1)||_q.$$

Raising to the $p$-th power, we get

$$|u - v|^p \le (|a|^p + |b|^p)2^{p/q}.$$

Since $p/q = p - 1$, we have that $2^{-(p-1)}|u - v|^p \leq |a|^p + |b|^p$.

For the upper bound, it suffices to observe that the functions that map $t \to t^+$ and $t \to t^-$ for $t \in \mathbb{R}$ are both 1-Lipschitz. That is, $|u^+ - v^+| \leq |u - v|$ and $|u^- - v^-| \leq |u - v|$ from which we get $|u^+ - v^+|^p + |u^- - v^-|^p \leq 2|u - v|^p$.

With Equation 1 established, we can apply it coordinate-wise to $u = x_i, v = y_i$ to get

$$2^{-(p-1)} \sum_{i=1}^{m} \Delta_i^p \leq \sum_{i=1}^{m} (A_i^p + B_i^p) \leq 2 \sum_{i=1}^{m} \Delta_i^p.$$

The result follows after taking the $p$-th root and observing that

$$||x - y||_p^p = \sum_i \Delta_i^p, \qquad ||\phi(x) - \phi(y)||_p^p = \sum_i (A_i^p + B_i^p).$$

□

We now proceed to the main theorem on sparse vector search.

**Theorem 5.4.** *Let $(Q, D, ||\cdot||_p)$ be a nearest neighbor search problem where $Q, D \subset \mathbb{R}^m_{\geq 0}$ with $||q||_p = ||d||_p = 1$ for all $q \in Q, d \in D$ and $p \geq 1$. Fix $\alpha > 1/2$, let $Q_\alpha$ be the concentrated queries from definition 5.1, and let $(q, d) \sim P_m$ be a random query-document pair where $P_m$ is the joint distribution.*

*Assume that*

- *(a) The query marginal of $P_m$ is supported on $Q_\alpha$, i.e., $C_q(\kappa) \geq \alpha$ for every query.*
- *(b) $(Q_\alpha, D)$ satisfies overlap of importance with parameters $(\gamma, \pi)$ as stated in Definition 5.3.*
- *(c) Let $B$ be the event defined by $B \coloneqq \{T_q \cap T_d = \emptyset\} \cap \{C_d(R\kappa) \geq \alpha\}$ and assume that there exists $\beta > 0$ such that $\Pr(B) \geq \beta$.*

*Under assumptions $(a)$–$(c)$ above, define $X \coloneqq (2 - 2\gamma)^{1/p}$ and $Y \coloneqq 2^{1/p} \left(\alpha^{1/p} - (1 - \alpha)^{1/p}\right)$ and assume that $Y > X$. Then, $\liminf_{m \to \infty} \mathrm{RelVar}_m \geq \frac{\pi\beta}{2^{2/p}}(Y - X)^2 > 0$.*

*Proof.* Set $\delta(q, d) \coloneqq ||q - d||_p$ (for brevity, we will denote it as $\delta$) and define the two events from assumptions (b) and (c)

$$A := \left\{ \sum_{i \in T_q \cap T_d} \min\{q_i^p, d_i^p\} \geq \gamma \right\},$$
$$B := \{T_q \cap T_d = \varnothing\} \cap \{C_d(R\kappa) \geq \alpha\}.$$

Since $P_m$ is supported on $Q_\alpha \times D$, by assumption $(a)$, every query satisfies

$$\|q_{T_q}\|_p^p = C_q(\kappa) \geq \alpha \tag{2}$$

Now, note that for any $a, b \geq 0$ and $p \geq 1$,

$$|a - b|^p \leq a^p + b^p - 2\min\{a^p, b^p\}. \tag{3}$$

To verify this, assume without loss of generality that $a \geq b$. The right-hand side is $a^p - b^p$ while

$$a^p - b^p = \int_b^a p x^{p-1} dx \geq \int_0^{a-b} p x^{p-1} dx = (a - b)^p.$$

Applying 3 coordinate-wise and using normalization, we get

$$\begin{aligned}\delta^p &= \sum_{i=1}^{m} |q_i - d_i|^p \\ &\leq \sum_{i=1}^{m} (q_i^p + d_i^p - 2\min\{q_i^p, d_i^p\}) \\ &= 2 - 2\sum_{i=1}^{m} \min\{q_i^p, d_i^p\}\end{aligned}$$

On event $A$,

$$\sum_{i=1}^{m} \min\{q_i^p, d_i^p\} \geq \sum_{i \in T_q \cap T_d} \min\{q_i^p, d_i^p\} \geq \gamma.$$

Therefore, on $A$.

$$\delta \leq (2 - 2\gamma)^{1/p}. \tag{4}$$

On $B$, document concentration yields $\|d_{T_d}\|_p^P \geq \alpha$, and by 2, $\|q_{T_q}\|_p \geq \alpha^{1/p}$.

Since $T_q \cap T_d = \emptyset$ and $d$ has total $p$-mass one, the document can place at most $1 - \alpha$ mass on $T_q$. Thus,

$$\|d_{T_q}\|_p \leq (1 - \alpha)^{1/p}.$$

The reverse triangle inequality restricted to $T_q$ yields

$$\|(q - d)_{T_q}\|_p \geq \|q_{T_q}\|_p - \|d_{T_q}\|_p \geq \alpha^{1/p} - (1 - \alpha)^{1/p}. \tag{5}$$

Similarly, because the query has at least $\alpha$ mass on $T_q$ by 2 and $T_q \cap T_d = \emptyset$, the query can place at most $1 - \alpha$ mass on $T_d$, i.e., $\|q_{T_d}\|_p \leq (1 - \alpha)^{1/p}$. Applying the reverse triangle inequality restricted to $T_d$ yields

$$\|(q - d)_{T_d}\|_p \geq \|d_{T_d}\|_p - \|q_{T_d}\|_p \geq \alpha^{1/p} - (1 - \alpha)^{1/p}. \tag{6}$$

Since $T_q$ and $T_d$ are disjoint, their $p$-power distance contributions add. Combining equation 5 and 6 we get

$$\begin{aligned}\delta^p &\geq \|(q - d)_{T_q}\|_p^p + \|(q - d)_{T_d}\|_p^p \\ &\geq 2\left(\alpha^{1/p} - (1 - \alpha)^{1/p}\right)^p.\end{aligned}$$

Thus, on event $B$, taking $p$-th roots,

$$\delta \geq 2^{1/p}\left(\alpha^{1/p} - (1 - \alpha)^{1/p}\right) = Y. \tag{7}$$

The requirement that $\alpha$ be greater than $1/2$ ensures that this lower bound is positive.

Now let $(q', d') \sim P_m$ be an independent copy of $(q, d)$ (so that $q' \in Q_\alpha$)) and set $\delta' \coloneqq \|q' - d'\|_p$. Then, it follows that

$$\mathrm{Var}[\delta] = \frac{1}{2}\mathbb{E}[(\delta - \delta')^2].$$

Consider the two ordered events

$$E_1 \coloneqq \{(q, d) \in A, (q', d') \in B\} \quad \text{and} \quad E_2 \coloneqq \{(q, d) \in B, (q', d') \in A\}.$$

Since $\gamma > 0$, event $A$ requires $T_q \cap T_d \neq \emptyset$, so $A \cap B = \emptyset$ and hence $E_1 \cap E_2 = \emptyset$. On either event, equation 7 and 4 imply

$$|\delta - \delta'| \geq Y - X > 0.$$

The two sampled pairs are independent, so each ordered event has probability $\Pr(A)\Pr(B)$. Consequently,

$$\mathrm{Var}(\delta) \geq \frac{1}{2}\mathbb{E}[(\delta-\delta')^2(\mathbf{1}_{E_1}+\mathbf{1}_{E_2}] \tag{8}$$

$$\geq \frac{1}{2}\left(2\Pr(A)\Pr(B)(Y-X)^2\right) \tag{9}$$

$$= \Pr(A)\Pr(B)(Y-X)^2 \tag{10}$$

$$\geq \pi\beta(Y-X)^2. \tag{11}$$

Finally, non-negative unit $\ell_p$ vectors satisfy

$$\delta^p = \sum_{i=1}^{m}|q_i-d_i|^p \leq \sum_{i=1}^{m}(q_i^p+d_i^p) = 2.$$

Thus, $(\mathbb{E}[\delta])^2 \leq 2^{2/p}$. Dividing 11 by the squared mean yields

$$\mathrm{RelVar}(\delta) = \frac{\mathrm{Var}(\delta)}{\mathbb{E}[\delta]^2} \geq \frac{\pi\beta}{2^{2/p}}(Y-X)^2.$$

The right-hand side is strictly positive and dimension-independent. Taking the $\liminf$ completes the proof.

□

*Remark* F.2. We state the theorem using $\liminf \mathrm{RelVar}_m$ instead of $\lim \mathrm{RelVar}_m$. This is convenient because bounding $\liminf$ away from zero is sufficient to prove stability without needing to prove that the relative variance terms converge, which may or may not happen.

### F.2 Empirical Validation of Sparse Theorem Assumptions

We check the assumptions of Theorem 5.4 on $30,522$-dimensional SPLADE-V3 (Lassance et al., 2024) embeddings across five BEIR corpora, at $p = 2$ with $\ell_2$-normalized vectors. For each corpus, we draw $N = 2 \times 10^6$ query-document pairs, sampling the query and the document independently and uniformly, and measure the concentration of importance, the near and far regimes $A \coloneqq \sum_{i\in T_q\cap T_d}\min\{q_i^2, d_i^2\} \geq \gamma$ and $B \coloneqq \{T_q \cap T_d = \emptyset\} \cap \{C_d(R\kappa) \geq \alpha\}$, the gap condition $Y > X$, and the empirical stability ratio DMAX/DMIN over the sampled queries. We use one parameter setting for all corpora: $\kappa = 48$, $R = 8$, and $\alpha = 0.999$. Table 5 reports the results.

**Choice of $\alpha$.** At $p = 2$ we have $X = \sqrt{2-2\gamma}$ and $Y = \sqrt{2}\left(\sqrt{\alpha}-\sqrt{1-\alpha}\right)$, so $Y^2 = 2 - 4\sqrt{\alpha(1-\alpha)}$ and the gap condition $Y > X$ is equivalent to $\gamma > \gamma_{\min}(\alpha) \coloneqq 2\sqrt{\alpha(1-\alpha)}$. In other words, once $\gamma$ exceeds $\gamma_{\min}(\alpha)$, every pair in $A$ is at distance at most $X$, which is strictly less than the distance $Y$ of any pair in $B$. It decreases as $\alpha \to 1$, so stronger concentration permits a weaker overlap requirement. For a moderate value of $\alpha$ this threshold is large (e.g., $\gamma_{\min}(\alpha) = 0.6$ at $\alpha = 0.9$), and under independent sampling almost no pairs reach it. We therefore take $\alpha = 0.999$, where $\gamma_{\min}(\alpha) = 0.063$ and $Y = 1.369$, and a document head of $R\kappa = 384$ coordinates, which SPLADE documents concentrate on almost entirely.

**Concentration of importance.** The restriction to $Q_\alpha$ retains $81\%$ of queries on HotpotQA, whose multi-hop queries are longer, $96\%$ on TREC-COVID and $99.8\%$ on the other three corpora. On the document side, a fraction $\rho \geq 0.992$ of documents satisfies $C_d(R\kappa) \geq \alpha$ with our choice of $R$ and $\kappa$ above.

We estimate $\hat{\beta} = \Pr[B]$, the probability that the query head and the document head are disjoint and the document is concentrated. It ranges from 0.043 on TREC-COVID, whose long queries rarely miss a document's support, to 0.512 on MS-MARCO.

**Overlap of importance** . For each pair we compute the overlap statistic $S = \sum_{i \in T_q \cap T_d} \min\{q_i^2, d_i^2\}$ and set $\hat{\pi}(\gamma) = \Pr[S \geq \gamma]$. Since $\gamma$ is a free parameter of the theorem, we choose the value above $\gamma_{\min}(\alpha)$ that maximizes the lower bound in Theorem 5.4 evaluated at the estimates $\hat{\pi}\hat{\beta}(Y - X)^2/2$. This yields $\gamma \in [0.110, 0.177]$ and $\hat{pi} \in [8.6 \times 10^{-4}, 3.0 \times 10^{-2}]$. On every corpus, between 1,711 and 62,252 sampled pairs satisfy $S \geq \gamma$, so each $\hat{\pi}$ is estimated from many pairs rather than from a few outliers. MS-MARCO has the rarest overlap and TREC-COVID the most frequent.

Overall, SPLADE embeddings satisfy every theorem assumption with the chosen parameter configuration, and the stability ratios we observe in Table 5 confirm the sufficiency of our theorem in determining when search is stable. Most queries are concentrated, documents are concentrated almost without exception, and disjoint-head pairs occur with constant probability. Overlapping pairs have non-negligible mass above the required threshold, and the gap $Y - X$ is strictly positive.

| Corpus | qCoI | $\rho$ | $\hat{\beta}$ | $\gamma$ | $\hat{\pi}$ | $n_A$ | $Y - X$ | bound | DMAX/DMIN |
|---|---|---|---|---|---|---|---|---|---|
| HotpotQA | 0.808 | 0.998 | 0.221 | 0.120 | $3.5 \times 10^{-3}$ | 6,930 | 0.042 | $6.7 \times 10^{-7}$ | 1.28 |
| MS-MARCO | 0.998 | 0.998 | 0.512 | 0.177 | $8.6 \times 10^{-4}$ | 1,711 | 0.086 | $1.6 \times 10^{-6}$ | 1.61 |
| Natural Questions | 0.998 | 0.995 | 0.251 | 0.117 | $2.5 \times 10^{-3}$ | 4,929 | 0.040 | $5.0 \times 10^{-7}$ | 1.37 |
| NFCorpus | 0.998 | 0.996 | 0.209 | 0.110 | $9.9 \times 10^{-4}$ | 1,979 | 0.035 | $1.2 \times 10^{-7}$ | 1.17 |
| TREC-COVID | 0.960 | 0.992 | 0.043 | 0.170 | $3.1 \times 10^{-2}$ | 62,252 | 0.080 | $4.3 \times 10^{-6}$ | 1.78 |

Table 5: Validation of Theorem 5.4 on SPLADE ($p = 2$, $\kappa = 48$, $R = 8$, $\alpha = 0.999$, $Y = 1.369$), with queries restricted to $Q_\alpha$. qCoI is the retained fraction $|Q_\alpha|/|Q|$. $\rho$ is the fraction of documents with $C_d(R\kappa) \geq \alpha$ and $\hat{\beta}$ the fraction of pairs in event $B$. The overlap threshold $\gamma$ is chosen to maximize the bound subject to $Y > X$. $n_A$ is the number of pairs in event $A$ ($S \geq \gamma$) and $\hat{\pi} = n_A/N$. The bound column, given by $\hat{\pi}\hat{\beta}(Y - X)^2/2$, is the lower bound of Theorem 5.4 at these estimates, and DMAX/DMIN is the empirical stability ratio.

### F.3 Sparse Search Stability Simulation

In this section, we evaluate Theorem 5.4 using sparse embeddings constructed to reflect the structural patterns commonly observed in modern sparse neural embeddings. Two properties guide our design, namely *(1) head concentration* where a small set of coordinates carries most of the $\ell_p$ mass and *(2) semantic locality*, where high-importance dimensions cluster within coherent neighborhoods instead of being uniformly distributed across the embedding vector. To reproduce these behaviors, our data generator samples high-importance dimensions from a Zipf-shaped distribution and orients this selection towards a semantic center, which causes most of the mass to concentrate within a contiguous set of dimensions.

Figure 3 shows the results of the experiments when we set $\alpha = 0.83$, $\gamma = 0.2$, and $\pi = 0.5$ for the overlap probability. In particular, the CoI-only regime highlights that even strongly-concentrated embeddings exhibit distance collapse as the dimension grows when there is not enough overlap in their heads. These results confirm that neither concentration nor overlap of importance are sufficient conditions for stability on their own.