Our work: • for all modalities of single-vector embeddings

• "theoretical limitation": "the need to transform to use multimodality" as the space of things to represent grows.

• limitations of vectors in geometric space:

~~History~~: ① The newly proposed angle:

Not How many k-subsets that a given configuration can realize

, but rather: what is the necessary embedding dimension to realize all k-subsets under score margin γ

• sphere-packing volume argument

$\Downarrow$

obtain a lower bound on the embedding dimension (in terms of n, k, γ)

Setup: Let $\{\vec{v}_i\}_{i=1}^{n} \subseteq \mathbb{R}^d$ be unit vectors, each representing a document.

Fix the score margin $\gamma > 0$.

We say that a k-subset $\boxed{S} \subseteq [n]$ is "realized with margin γ" $\boxed{0 \le \gamma \le 1}$

if $\exists$ $\boxed{\text{unit query } u_S}$, such that

$$\left(\min_{i \in S} \underbrace{\langle u_S, v_i \rangle}_{\le \|u_S\| \times \|v_i\|}\right) \geq \left(\max_{j \notin S} \langle u_S, v_j \rangle\right) + 2 \times \gamma$$

$\|u_S\|, \|v_i\|$ both $= 1$

i.e., if the dot product between the specified query u_S and a candidate document (vector) is "~~[illegible]~~ the bigger, the better"

then "realization of S" is existence of a "magic" u_S

s.t. [lower "threshold" within S] exceeds [upper "threshold outside S] by $\geq 2\gamma$

T thm 1: (lower bound of $\underbrace{\text{dimension } d}_{:= \dim(\mathbb{R}^d)}$).

Assume $1 \le k < n$. Assume $\forall$ k-subset $S \subseteq [n]$,

then S can be realized with margin γ.

[this is an ideal assumption of document set $\{\vec{v}_i\}_{i=1}^n$]

Claim: $\binom{n}{k} \le \left(1+\frac{1}{\gamma}\right)^d$, thus $d \le \underbrace{\left(\log_{(1+\frac{1}{\gamma})} \binom{n}{k}\right)}_{??}$

proof: Consider $S, T \subseteq [n]$

where $|S| = |T| = k$

and $S \neq T$.

We can thus choose $(\exists)$ $i \in S \setminus T$ and $j \in T \setminus S$.

☆☆ In particular, by the assumption that every k-subset is γ-realizable, thus S and T are both γ-realizable

Thus $\exists$ unit query u_S, where

$$\left(\min_{\tilde{i} \in S} \langle u_S, v_{\tilde{i}} \rangle\right) - \left(\max_{\tilde{j} \notin S} \langle u_S, v_j \rangle\right) \ge 2\gamma$$

and $\exists$ unit query $\boxed{u_T}$, where

$$\left(\min_{\tilde{i} \in T} \langle u_T, v_i \rangle\right) - \left(\max_{\tilde{j} \notin T} \langle u_T, v_j \rangle\right) \ge 2\gamma$$

In particular,

dot product with u_S for some element in S

min in S

max not in S

dot prod. with u_S for some element not in S

Thus

$$\begin{cases} \langle u_S, v_i \rangle - \langle u_S, v_j \rangle \geq 2 \times \gamma \\ \langle u_T, v_j \rangle - \langle u_T, v_i \rangle \geq 2 \times \gamma \end{cases}$$

Thus

$$\begin{cases} \langle u_S, (v_i - v_j) \rangle \geq 2 \times \gamma \cdots ① \\ \langle u_T, (v_j - v_i) \rangle \geq 2 \times \gamma \cdots ② \end{cases}$$

Consider negative sign

adding ① and ②,

we have [whilst realizing $\langle u_T, (v_j - v_i) \rangle = \langle -u_T, (v_i - v_j) \rangle$]

$\therefore$ we have $\langle (u_S - u_T), (v_i - v_j) \rangle$

$$\geq 2\gamma + 2\gamma = 4\gamma$$

Note that $4\gamma \leq \langle u_S - u_T, v_i - v_j \rangle$

we want to "say sth." about u_S and u_T

$$\leq \| u_S - u_T \| \times \| v_i - v_j \|$$

$$\leq \| u_S - u_T \| \times [\underbrace{\| v_i \|}_{=1} + \underbrace{\| -v_j \|}_{= \| v_j \| = 1}]$$

$$\leq 2 \times \| u_S - u_T \|$$

(because we assume all the n documents are represented as unit vectors)

$\therefore \| u_S - u_T \| \geq 2\gamma.$

$M :=$ number of size-K subsets S of $[n] := \{1, 2, \cdots, n\}$

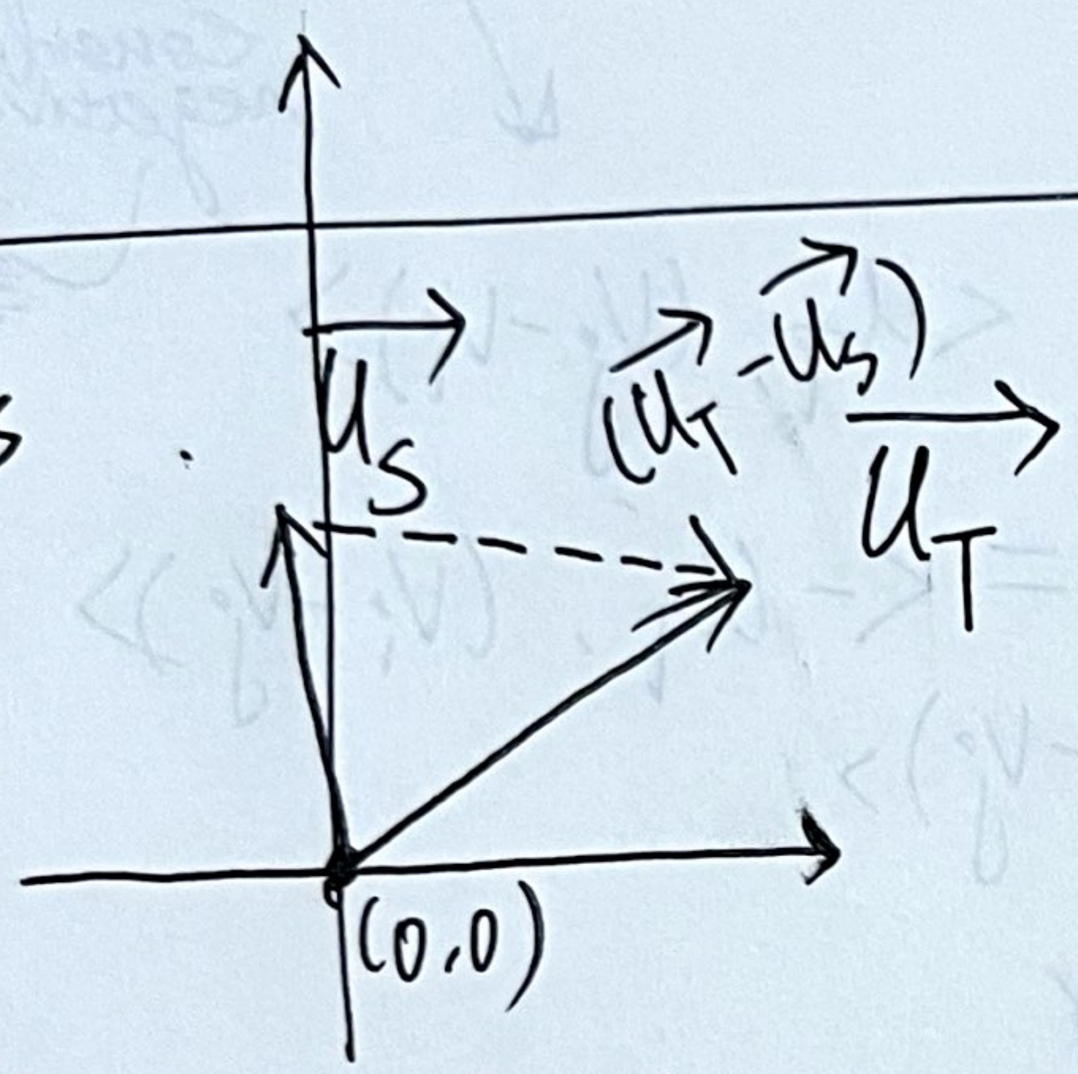
$= \#$ of queries u_S for such S $= \binom{n}{k}$

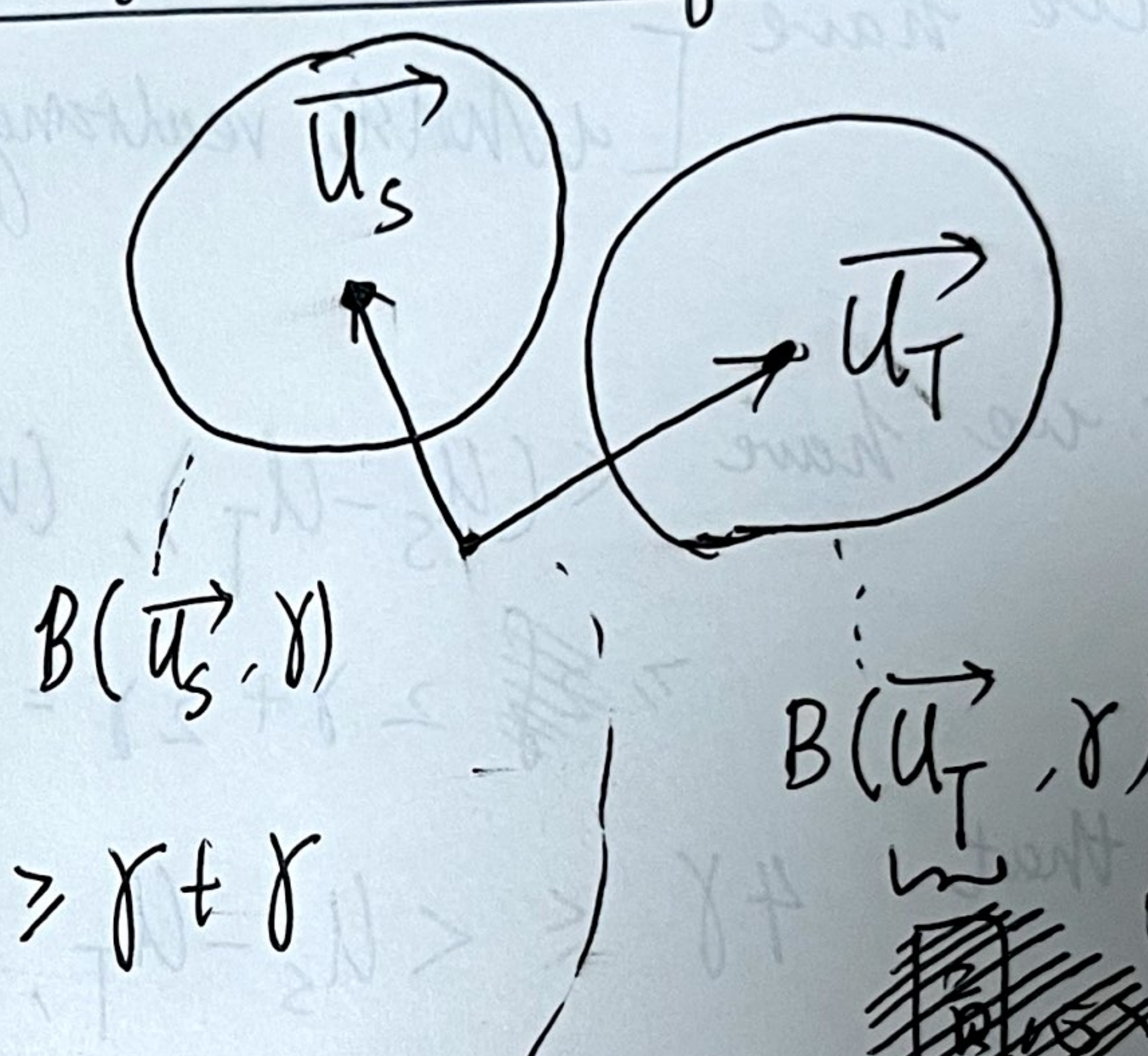
Recall that S and T are arbitrary size k subsets of $[n]$ $(S \neq T)$

we know $\|u_S - u_T\| \geq 2 \times \gamma$

$\therefore$ In fact, for any 2 (out of M total) k-subsets of S, their $\underbrace{\text{corresponding queries}}_{\text{under scheme of } \gamma\text{-realizability}}$ are separated by $\geq 2 \times \gamma$.

Thus





$\rightarrow \|\vec{u_T} - \vec{u_S}\| \geq \gamma + \gamma$

Proposition: $B(u_S, \gamma) \subseteq B(0, (1+\gamma))$

proof: suppose $\vec{x} \in B(u_S, \gamma)$ be arbitrary

Thus $\|\vec{x} - u_S\| \leq \gamma$

Thus $\|\vec{x} - \vec{0}\|$

$= \|\vec{x}\| = \|(\vec{x} - u_S) + u_S\|$

$\leq \|\vec{x} - u_S\| + \underbrace{\|u_S\|}_{=1}$

$= 1 + \|\vec{x} - u_S\|$

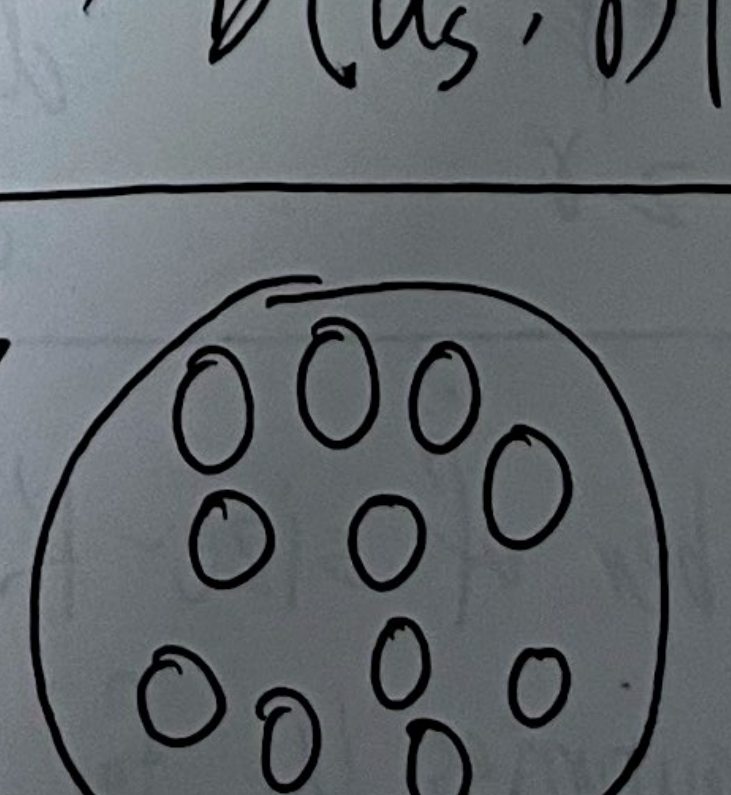
$\leq 1 + \gamma$

for $\forall S, T \subseteq [n]$ with $|S| = |T| = k$

~~that the~~ open balls $B(\vec{u_S}, \gamma)$ and $B(\vec{u_T}, \gamma)$ are disjoint

(i.e., $B(\vec{u_S}, \gamma) \cap B(\vec{u_T}, \gamma) = \emptyset$)

Thus



$M \times Vol(B_d(\gamma)) \leq Vol(B_d(1+\gamma))$

$$\therefore \quad M \times C_d \times \gamma^d \leq C_d \times (1+\gamma)^d$$

$$\therefore \quad M \leq \left(\frac{\gamma}{1+\gamma}\right)^d = \left(1 - \frac{1}{\gamma}\right)^d$$