How Many Features Can a Language Model Store Under the Linear Representation Hypothesis?

Nikhil Garg [1] **Jon Kleinberg** [1] **Kenny Peng** [1]

Abstract

We introduce a mathematical framework for the linear representation hypothesis (LRH), which asserts that intermediate layers of language models store features linearly. We separate the hypothesis into two claims: linear *representation* (features are linearly embedded in neuron activations) and linear *accessibility* (features can be linearly decoded). We then ask: How many neurons d suffice to both linearly represent and linearly access m features? Classical results in compressed sensing imply that for k-sparse inputs, $d = O(k \log(m/k))$ suffices if we allow non-linear decoding algorithms (Candes and Tao, 2006; Candès et al., 2006; Donoho, 2006). However, the additional requirement of linear decoding takes the problem out of the classical compressed sensing, into *linear* compressed sensing.

Our main theoretical result establishes nearly-matching upper and lower bounds for linear compressed sensing. We prove that $d = \Omega_\epsilon(\frac{k^2}{\log k} \log(m/k))$ is required while $d = O_\epsilon(k^2 \log m)$ suffices. The lower bound establishes a quantitative gap between classical and linear compressed setting, illustrating how linear accessibility is a meaningfully stronger hypothesis than linear representation alone. The upper bound confirms that neurons can store an exponential number of features under the LRH, giving theoretical evidence for the "superposition hypothesis" (Elhage et al., 2022).

The upper bound proof uses standard random constructions of matrices with approximately orthogonal columns. The lower bound proof uses rank bounds for near-identity matrices (Alon, 2003) together with Turán's theorem (bounding the number of edges in clique-free graphs). We also show how our results do and do not constrain the geometry of feature representations and extend our results to allow decoders with an activation function and bias.

1. Introduction

Towards understanding how language models work, a single principle has been remarkably useful: the Linear Representation Hypothesis (LRH) states that intermediate layers of language models store features linearly.

The observance of linear structure dates back to early work on word embeddings (Mikolov et al., 2013a;b; Arora et al., 2018). (Alain and Bengio, 2016) demonstrated empirically that in deeper layers of vision models, features (e.g., the presence of cats or dogs) could increasingly be extracted using linear classifiers trained on activations, suggesting that deep neural networks work to arrange features linearly. Since then, work has shown that a wide range of features can be accessed using these "linear probes," such as sentiment and game state (Nanda et al., 2023; Li et al., 2025). Leveraging the idea that features are stored as directions in activation space, a line of work has demonstrated the ability to steer language model behavior by manipulating activations along certain directions (Panickssery et al., 2023; Zou et al., 2023; Turner et al., 2023; Wang et al., 2025; Chen et al., 2025). Moreover, the assumption of an underlying linear structure has motivated the use of sparse autoencoders to automatically extract features from language models. Strikingly, these features appear to be highly-interpretable while capturing much of the original information (Cunningham et al., 2023; Gao et al., 2024; Movva et al., 2025). These linearly-represented features have formed the foundation for interpretable and causal explanations for how language models perform certain computations, like writing poems or answering basic questions (Lindsey et al., 2025).

Given the tremendous amount of activity that revolves around the LRH, recent efforts have aimed to give a formal account of the hypothesis and to derive and test implications (Elhage et al., 2022; Park et al., 2023; Costa et al., 2025). In particular, it has been hypothesized that linear representation allows for *superposition*, where d neurons can store $m \gg d$

[1]Cornell University. Correspondence to: Kenny Peng <kennypeng@cs.cornell.edu>.

Accepted to the 2nd Workshop on Compositional Learning at ICML 2026, Seoul, South Korea. Copyright 2026 by the author(s).

features (Elhage et al., 2022). Given the vast number of relevant concepts needed to generate text, such a capability appears to be necessary given the performance of language models. While results such as the Johnson-Lindenstrauss lemma (Johnson et al., 1984) and from compressed sensing (Candes and Tao, 2006; Candès et al., 2006; Donoho, 2006) are suggestive, and empirical evidence exists in toy settings (Elhage et al., 2022), current theoretical accounts do not clearly establish bounds on the amount of possibe superposition under the LRH.

In the present work, we aim to give a mathematical framework capable of addressing the following fundamental question (stated informally here):

> (*) Under the linear representation hypothesis, how many features can one layer of d neurons store?

More concretely, an intermediate layer of a language model embeds an input as a d-dimensional vector (the activations of the d neurons). Then how many features can this vector store, and under what conditions?

The present work illustrates one reason why there is a lack of clarity surrounding this basic question: different formulations of the LRH imply different answers to the question. In particular, we show a quantitative gap in the answer to (*) depending on if the LRH states that features are *represented* linearly or are also *accessible* linearly. These statements have often been treated as equivalent: for example, (Elhage et al., 2022) write that

> Linear representations make features "linearly accessible." A typical neural network layer is a linear function followed by a non-linearity. If a feature in the previous layer is represented linearly, a neuron in the next layer can "select it" and have it consistently excite or inhibit that neuron. If a feature were represented non-linearly, the model would not be able to do this in a single step.

Both linear representation and accessibility play an important role in work surrounding the LRH. Linear representation implies that model behavior can be altered by modifying activation space along linear directions. On the other hand, linear accessibility implies that features can be extracted using linear probes, and as (Elhage et al., 2022) note above, accessibility allows the next layer of a model to efficiently use a feature. Furthermore, both linear representation and accessibility are implicit in the design of sparse autoencoders used in practice, which utilize both linear encoding and decoding.

One key implication of the present work is that there is a meaningful way in which linear representation does not imply linear accessibility. We now describe our mathematical framework and results, illustrating this conceptual difference, and then showing how they lead to a quantitative difference.

1.1. Mathematical Framework

In this section, we introduce a mathematical framework that allows us to formally study implications of the linear representation hypothesis. In particular, the framework allows us to cleanly understand the difference between (linear) representation and (linear) accessibility, while also allowing us to formally quantify what it means for superposition to occur (for neuron activations to store more features than there are neurons).

The theoretical results, stated in Section 1.2, are self-contained. However, the framework here provides the more general context establishing the relationship between our mathematical results and the emerging scientific theory of language models.

Activations and features. We will assume that inputs comes from some set L, which we call a **language** (i.e., sequences of tokens). We let $f : L \to \mathbb{R}^d$ give the **activations** of d neurons in an intermediate layer. We let a **feature** be a function $z_i : L \to \mathbb{R}$. For example, a feature can capture if the text is a question, if it has positive sentiment, if it is about dogs, or if it should be continued with a number. Importantly, we think of the value of a feature $z_i(\ell)$ as depending only on the input text; in other words, features can exist independently of any language model.

Feature accessibility. A feature z_i is (ϵ, S)**-recovered** by a **probe** $g_i : \mathbb{R}^d \to \mathbb{R}$ if

$$|g_i(f(\ell)) - z_i(\ell)| < \epsilon \tag{1}$$

for all $\ell \in S$. Intuitively, this means that the value of a feature can be accessed from a text's activations $f(\ell)$, at least up to some approximation error.

The Linear Representation Hypothesis. We now formalize how the LRH places further restrictions on the above framework.

Linear representation for a set of features $z_1, \cdots, z_m : L \to \mathbb{R}$ implies that there exists a set of **representation vectors** $a_1, \cdots, a_m \in \mathbb{R}^d$, such that

$$f(\ell) = \sum_{i=1}^{m} z_i(\ell) a_i. \tag{2}$$

Equivalently, there exists $A \in \mathbb{R}^{d \times m}$ such that

$$f(\ell) = Az(\ell), \tag{3}$$

where $z(\ell) = [z_1(\ell), z_2(\ell), \cdots, z_m(\ell)] \in \mathbb{R}^m$. We refer to $z(\ell)$ as a text ℓ's **feature representation**. Under linear representation, activations are a function only of an input's feature representation.

Linear accessibility places a restriction on the probes g_i (in contrast to linear representation, which restricts f). Formally, under linear accessibility, we require the probe g_i to satisfy $g_i(x) = \langle b_i, x \rangle$ for some **probe vector** (or **linear probe**) $b_i \in \mathbb{R}^d$. Therefore, a feature z_i is (ϵ, S)**-linearly recovered** by a linear probe b_i if

$$|\langle b_i, f(\ell) \rangle - z_i(\ell)| < \epsilon \tag{4}$$

for all $\ell \in S$. Then if all features $z_1, \cdots, z_m$ can be (ϵ, S)-linearly recovered, this implies that there exists $B \in \mathbb{R}^{d \times m}$ such that

$$\left\| B^T f(\ell) - z(\ell) \right\|_\infty < \epsilon \tag{5}$$

for all $\ell \in S$.

Superposition. We can now see how this framework allows us to formally reason about (*), or about how much superposition can occur under the LRH. In particular, we are interested in understanding how many features m can be stored using activations in $\mathbb{R}^d$ (of d neurons) under the hypotheses of linear representation and linear recovery.

First note that under the linear representation, activations of an input text are a function only of its feature representation. Therefore, we can redefine activations to be given by $f : \mathbb{R}^m \to \mathbb{R}^d$. Under this definition, we restate linear representation to say that

$$f(z) = Az \tag{6}$$

for $z \in \mathbb{R}^m, A \in \mathbb{R}^{d \times m}$.

Now, assuming linear representation, we can ask two questions:

Q1. (General accessibility.) For a fixed m, how many dimensions d are required for there to exist $A \in \mathbb{R}^{d \times m}$ and $g : \mathbb{R}^d \to \mathbb{R}^m$ such that

$$\|g(Az) - z\|_\infty < \epsilon \tag{7}$$

for all $z \in S$.

Q2. (Linear accessibility.) For a fixed m, how many dimensions d are required for there to exist $A, B \in \mathbb{R}^{d \times m}$ such that

$$\left\| B^T Az - z \right\|_\infty < \epsilon \tag{8}$$

for all $z \in S$.

In both settings, we will generally assume S to be the set of k-sparse vectors (i.e., having at most k non-zero entries), and answer in terms of m, k, ϵ. In other words, we study settings where only a small number of features tend to be "active" for any particular input (intuitively, this is matches our intuition about language—that any given piece of text only expresses a small fraction of possible concepts).

A key result we show is that there is a quantitative gap in the answer to (1) and (2). In other words, there is a meaningful sense in which linear representation does not imply linear accessibility. In both settings, superposition can occur: an exponential number of features can be represented and accessed. However, we will show that the amount of tolerated sparsity differs significantly.

1.2. Results

We now state theoretical results that resolve the questions Q1 and Q2 posed above. First, we will show how classical results in compressed sensing resolve the general accessibility setting (Theorem 1.1). Then, we will show that the additional restriction to *linear* probes brings us into the different setting of *linear* compressed setting. The primary contributions of the present work are nearly-matching upper and lower bounds in the linear setting (Theorem 1.2 and Theorem 1.3).

Past Results: Compressed Sensing. Let us recall a classical result in compressed sensing (Candes and Tao, 2006; Candès et al., 2006; Donoho, 2006), which exactly deals with the setting in which features are represented linearly but accessed non-linearly, answering Q1.

Theorem 1.1 (Compressed Sensing)**.** *There exists a matrix $A \in \mathbb{R}^{d \times m}$ with $d = O\left(k \log \frac{m}{k}\right)$ such that for all k-sparse $z \in \mathbb{R}^m$,*

$$g(x) := \operatorname{argmin}_{z' \in \mathbb{R}^m : x = Az'} \|z'\|_1 \tag{9}$$

satisfies $g(Az) = z$.

This shows that when

$$d = O\left(k \log \frac{m}{k}\right), \tag{10}$$

k-sparse z in $\mathbb{R}^m$ can be linearly embedded in $\mathbb{R}^d$ in a way such that z can be *exactly* recovered from the embedding using ℓ_1-minimization (basis pursuit). However, basis pursuit is non-linear. We are interested in the case when features can be extracted linearly, which we discuss next.

Our Results: Linear Compressed Sensing. The requirement that the recovery function be linear takes us out of classical compressed sensing settings, which generally allows arbitrary (efficient) algorithms to perform recovery. The assumption of linear recovery (the setting of Q2) brings us into the regime of what could be called *linear compressed sensing*.

Let $d(m, k, \epsilon)$ be the smallest choice of d such that there exists $A, B \in \mathbb{R}^{d \times m}$ such that

$$\left\| B^T A z - z \right\|_\infty < \epsilon \tag{11}$$

for all k-sparse $z \in [-1, 1]^m$. (Here, any bounded interval would suffice.) Our main results provide nearly-matching upper and lower bounds on $d(m, k, \epsilon)$.

Theorem 1.2 (Upper Bound)**.**

$$d(m, k, \epsilon) = O_\epsilon \left(k^2 \log m \right). \tag{12}$$

This upper bound demonstrates that embeddings can in fact store an exponential number of features relative to embedding dimension.

Theorem 1.3 (Lower Bound)**.** *For* $\epsilon > \frac{k^{3/2}\sqrt{5}}{\sqrt{m}}$,

$$d(m, k, \epsilon) = \Omega_\epsilon \left(\frac{k^2}{\log k} \log \frac{m}{k} \right). \tag{13}$$

(Here, the restriction $\epsilon > \frac{k^{3/2}\sqrt{5}}{\sqrt{m}}$ is not very consequential, since we are generally interested in the regime where $k = O(\log m)$.) The lower bound establishes a quantitative gap between compressed sensing and linear compressed sensing. In particular, while d must be essentially quadratic in k in linear compressed sensing, it need only be linear in k in classical compressed sensing.

Further results. There has also been general interest in the "geometry" of features stored in language models (Park et al., 2023). For example, it is sometimes stated that (unrelated) features are represented as *orthogonal* directions under the LRH (Hindupur et al., 2025). In Section 4, we show what our theoretical results imply (and do not imply) about feature geometry. In particular, Proposition 4.1 shows that the lower bound in Theorem 1.3 can be achieved even for non-intuitive arrangements of feature directions: features can be represented and probed by approximately orthogonal directions, while different features can be represented by highly-correlated directions and also probed by highly-correlated directions. These results highlight the distinction between representation and probe vectors (also addressing uncertainty about whether encoder and decoder weights in sparse autoencoders correspond to a feature's "true" direction), and how the major requirement for linear feature representation and accessibility is that a feature's representation vector must be approximately orthogonal to other features' probe vectors (not other features' representation vectors). In Proposition 4.3, we then show how bounds on the magnitude of feature vectors limit the amount of "unusualness" in feature geometry.

In Section A, we then show that the lower bound in Theorem 1.3 also applies to more general settings. In particular, in Theorem A.1 we show that the bound applies in a setting where features are binary and the task is to linearly classify inputs according to each feature. As a corollary, the result implies that adding a bias and non-linear activation function to a linear probe does not significantly increase the number of features a model can store.

1.3. Intuition and Proof Techniques

We begin by developing some geometric intuition for the problem. Let $a_1, \cdots, a_m \in \mathbb{R}^d$ be the columns of $A \in \mathbb{R}^{d \times m}$ and $b_1, \cdots, b_m \in \mathbb{R}^d$ be the columns of $B \in \mathbb{R}^{d \times m}$. For each $i \in [m]$, we recall that a_i is feature i's **representation vector** and b_i is feature i's **probe vector**. Now let $\hat{z} = B^T A z$. We are interested in when

$$\|\hat{z} - z\|_\infty < \epsilon \iff |\hat{z}_i - z_i| < \epsilon, \quad \forall i \in [m]. \tag{14}$$

Note that

$$\hat{z}_i = \langle b_i, A z \rangle = \sum_{j=1}^{m} z_j \langle b_i, a_j \rangle. \tag{15}$$

Therefore, the value $\langle b_i, a_j \rangle$ tells us how much the probe b_i for the i-th feature responds to presence of the j-th feature. If $|\langle b_i, a_j \rangle|$ is large for $i \neq j$, that means that a text that contains feature j will contribute to the measurement of feature i. In this case, we say that feature j interferes with the measurement of feature i. If $\langle b_i, a_j \rangle > 0$, this is positive interference (presence of feature j will lead to inflated measurement of feature i). If $\langle b_i, a_j \rangle < 0$, this is negative interference (presence of feature j will lead to deflated measurement of feature i).

Roughly speaking, for (14) to hold, there must be low interference; equivalently, $B^T A$ must be small off-diagonal. A general step in our analysis is to consider how k-sparsity limits the amount by which interference across multiple features can compound.

Upper bound sketch. We can give a fairly straightforward construction to establish the upper bound. In particular, we can give a construction where $A = B$. Then it suffices to find a matrix A such that $\langle a_i, a_i \rangle = 1$ and such that $\langle a_i, a_j \rangle < \frac{\epsilon}{k}$. Leveraging the k-sparsity of z, this would show that the total interference for any feature is bounded by $k \cdot \frac{\epsilon}{k} = \epsilon$. We can then use standard constructions of incoherent matrices (i.e., matrices with approximately-orthonormal columns), such as appropriately-scaled random Gaussian matrices, to obtain the desired bound.

Initial attempt at lower bound. To prove the lower bound, a natural first inclination is to show that among a sufficiently large set of directions in $\mathbb{R}^d$, there must exist correlated directions. In particular, it is possible to show that for d small relative to m, there exists $i \in [m]$ and k choices

of j such that $|\langle a_i, a_j\rangle| > \frac{\epsilon}{k}$ and such that $\langle a_i, a_j\rangle$ have the same sign (and thus compound in interference). However, this approach does not generally work when $A \neq B$, since we do not strictly require representation directions of different features (or probe directions of different features) to be orthogonal. For $i \neq j$, it is only necessary for the inner product of feature i's representation vector to be orthogonal to feature j's *probe* vector. This means that there may exist "good" sets of representation directions and probe directions among which many are pairwise correlated.

In fact, in Section 4, we show that there are constructions with $B \neq A$ where probe directions are highly correlated with each other and representation directions are also highly correlated with each other. So while showing the existence of correlated vectors gives a lower bound in the case $A = B$, it does not suffice in general.

Lower bound sketch. It would suffice to show that if $\langle b_i, a_i\rangle$ is near 1 (as is required if $B^T A e_i \approx e_i$) and d is small, then there must exist many pairs $i \neq j$ such that $|\langle b_i, a_j\rangle|$ is large. In particular, we would like to find a feature i such that there are k choices of j such that $|\langle b_i, a_j\rangle| > \frac{\epsilon}{k}$ and such that the interference is all in the same direction. To do this, it suffices to find $2k$ such vectors a_j with $|\langle b_i, a_j\rangle| > \frac{\epsilon}{k}$ (by the pigeonhole principle).

The first idea is to use a result of (Alon, 2003), which demonstrates that low-rank matrices that are 1 on the diagonal must have a large off-diagonal entry. (This result also notably provides a near-optimal lower bound on the Johnson-Lindenstrauss Lemma.) In particular, we know that $B^T A$ has rank at most d, and we can minimally scale rows to allow diagonal entries to be 1. Note, however, that applying the result to $B^T A$ itself does not suffice since this only guarantees the existence of a single interfering pair of features.

So the second main idea is to apply the result instead to large principal submatrices of $B^T A$, which shows that any submatrix of $B^T A$ must have a large off-diagonal entry. Equivalently, among any large subset $R \subseteq [m]$, there must exist a pair $i, j \in R$ with $i \neq j$ such that $\langle b_i, a_j\rangle$ is large.

Finally, we consider a graph reduction and apply Turán's theorem. Consider a graph G with vertices $[m]$ and such that there is an edge between i and j if and only if $|\langle b_i, a_j\rangle| > \frac{\epsilon}{k}$ or $|\langle b_j, a_i\rangle| > \frac{\epsilon}{k}$. Then G has no independent set of size r. Then we can apply Turán's theorem (which gives an upper bound on the number of edges in a graph without an r-clique) on $\overline{G}$, hence providing a lower bound on the number of edges in G, which in turns gives a lower bound on the number of pairs of interfering features. Then we can use the pigeonhole principle to show that there must exist some probe vector that has large interference with *many* representation vectors, as desired.

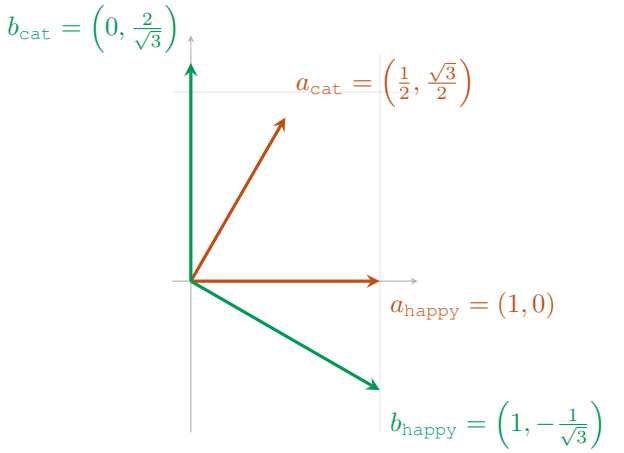


Figure 1. Representation vectors $a_{\texttt{cat}}, a_{\texttt{happy}}$ and probe vectors $b_{\texttt{cat}}, b_{\texttt{happy}}$ that yield perfect linear recovery of the `cat` and `happy` features. Notice that while the representation vectors are not orthogonal, the probe vectors are able perfectly extract features since the `cat` probe is orthogonal to the `happy` representation and the `happy` probe is orthogonal to the `cat` representation. Indeed, observe that $b_{\texttt{cat}}(z_{\texttt{cat}} a_{\texttt{cat}} + z_{\texttt{happy}} a_{\texttt{happy}}) = z_{\texttt{cat}}\langle b_{\texttt{cat}}, a_{\texttt{cat}}\rangle + z_{\texttt{happy}}\langle b_{\texttt{cat}}, a_{\texttt{happy}}\rangle = z_{\texttt{cat}} \cdot 1 + z_{\texttt{happy}} \cdot 0$.

1.4. Implications for the Linear Representation Hypothesis and the Theory of Language Models

In summary, this paper makes two main contributions:

First, we provide a mathematical framework for the linear representation hypothesis that cleanly distinguishes linear representation and linear accessibility, and allows us to formally analyze the amount of superposition possible under the LRH.

Second, we derive theoretical implications of the linear representation hypothesis. Mathematically, we establish a quantitative gap between compressed sensing and *linear* compressed sensing. Our main results demonstrate dependencies between superposition (the number of features stored relative to the number of neurons), sparsity (the number of features active per input), accessibility (the precision by which the next layer can access features stored in the previous layer), and feature geometry (the arrangement of linear directions corresponding to features).

2. Upper Bound Proof

In this section, we prove Theorem 1.2. In particular, we establish that there exists a matrix $A \in \mathbb{R}^{d \times m}$ with $d = O_\epsilon(k^2 \log m)$, such that

$$\left\|A^T A z - z\right\|_\infty < \epsilon \tag{16}$$

for all k-sparse $z \in [-1, 1]^m$. Intuitively, we want to find A such that $A^T A$ is close to the identity matrix—i.e., that A is approximately orthonormal. The following definition will

be useful.

Definition 2.1 (μ-incoherence)**.** Say that a matrix C with columns $c_1, \cdots, c_m$ is μ-incoherent if $\langle c_i, c_i \rangle = 1$ for all $i \in [m]$, and $|\langle c_i, c_j \rangle| < \mu$ for all $i \neq j$.

We can then state the following standard result, which demonstrates that random matrices are incoherent. (Equivalently, it shows that we can select an exponential number of approximately orthogonal directions.)

Lemma 2.2 (Incoherence of Random Matrices)**.** *Let $C \in \mathbb{R}^{d \times m}$ be a matrix with i.i.d. scaled Rademacher entries $\pm\frac{1}{\sqrt{d}}$. For $d = O\left(\frac{\log m + \log \frac{1}{\delta}}{\mu^2}\right)$, C is μ-incoherent with probability at least $1 - \delta$.*

The result guarantees the existence of a μ-incoherent matrix with $d = O(\frac{\log m}{\mu^2})$.

Proof of Theorem 1.2. Let $\hat{z} = A^T A z$ and suppose that A is μ-incoherent. Then if $z \in [-1, 1]^m$ has support T with $|T| \leq k$, we have that

$$\hat{z}_i = \left\langle a_i, \sum_{j=1}^{m} z_j a_j \right\rangle = z_i \langle a_i, a_i \rangle + \sum_{j \in T \setminus \{i\}} z_j \langle a_i, a_j \rangle \tag{17}$$

$$= z_i + \sum_{j \in T \setminus \{i\}} z_j \langle a_i, a_j \rangle. \tag{18}$$

Therefore, $|\hat{z}_i - z_i| < \sum_{j \in T \setminus \{i\}} |z_j \langle a_i, a_j \rangle| < k\mu$, where in the last inequality, we used that A is μ-incoherent, $|T| \leq k$, and $z_j \in [-1, 1]$. This implies that $\|\hat{z} - z\|_\infty < k\mu$. Taking $\mu = \frac{\epsilon}{k}$ and applying Lemma 2.2, we have

$$d(m, k, \epsilon) = O\left(\frac{k^2 \log m}{\epsilon^2}\right), \tag{19}$$

as desired.

Remark 2.3. The construction we have provided has $B = A$, meaning that features are represented with and probed with vectors in the same direction. In fact, as we show in Section 4, there exist constructions where a feature's representation and probe vectors are themselves approximately orthogonal, while all feature directions are highly correlated and all probe directions are also highly correlated.

The observation here also implies that attempts to show lower bounds by proving that sufficiently large sets of vectors must contain a pair of correlated vectors does not succeed, since collections of "good" representation and probe directions can in general contain many highly correlated directions.

3. Lower Bound Proof

In this section, we prove Theorem 1.3. In particular, we establish that for $\epsilon > \frac{k^{3/2}\sqrt{5}}{\sqrt{m}}$, if there exists $A, B \in \mathbb{R}^{d \times m}$ such that $\left\|B^T A z - z\right\|_\infty < \epsilon$ for all k-sparse $z \in [-1, 1]^m$, then

$$d = \Omega_\epsilon\left(\frac{k^2}{\log k} \log \frac{m}{k}\right). \tag{20}$$

Preliminary observations. Consider $A, B \in \mathbb{R}^{d \times m}$ such that $\left\|B^T A z - z\right\|_\infty < \epsilon$ for all k-sparse $z \in [-1, 1]^m$. Taking $z = e_i$, this implies that

$$|(B^T A e_i)_i - 1| < \epsilon \implies \langle b_i, a_i \rangle \in 1 \pm \epsilon. \tag{21}$$

for all $i \in [m]$. Furthermore, taking $z = \sum_{j \in T} e_j$ where $T \subseteq [m] \setminus \{i\}, |T| \leq k$ implies that

$$|(B^T A z)_i - 0| < \epsilon \implies |\langle b_i, \sum_{j \in T} a_j \rangle| < \epsilon. \tag{22}$$

To prove Theorem 1.3, our approach will be to show that if $A, B \in \mathbb{R}^{d \times m}$ satisfy (21) with $d < O_\epsilon(\frac{k^2}{\log k} \log \frac{m}{k})$, then we can find a contradiction to (22).

Finding many features with correlated interference. It would suffice to find $i \in [m]$ and a set $T^* \subseteq [m] \setminus \{i\}$ such that $|T^*| \geq 2k$ and $|\langle b_i, a_j \rangle| > \frac{\epsilon}{k}$ for all $j \in T$. This would imply (by the pigeonhole principle) the existence of a subset $T \subseteq T^*$ with $|T| = k$ such that $|\langle b_i, a_j \rangle| > \frac{\epsilon}{k}$ for all $j \in T$ and such that $|\langle b_i, a_j \rangle|$ has the same sign for all $j \in T$. In this case,

$$|\langle b_i, \sum_{j \in T} a_j \rangle| = \sum_{j \in T} |\langle b_i, a_j \rangle| > k \cdot \frac{\epsilon}{k} = \epsilon, \tag{23}$$

providing the needed contradiction. The remainder of the proof consists of showing that when $d < O_\epsilon(\frac{k^2}{\log k} \log \frac{m}{k})$, there must exist such i, T^*.

Bounding the rank of matrices with small off-diagonals. Note that the matrix $B^T A$ has entries $(B^T A)_{ij} = \langle b_i, a_j \rangle$. Also, $\text{rank}(B^T A) \leq d$. Our aim is to show that when d is small relative to m, under the assumption of (21), there must exist a row of $B^T A$ that has $2k$ off-diagonal entries with absolute value at least $\frac{\epsilon}{k}$. At a high level, the idea we pursue is that matrices that are near-zero off-diagonal (and large on the diagonal) must have large rank.

We begin by stating a result given by (Alon, 2003). The result shows that any square matrix that is one on the diagonal and bounded off-diagonal must have large rank. Of course, if the off-diagonal is 0, the matrix is the identity and thus has full rank. The result shows that we can bound the rank from below given small perturbations.

Lemma 3.1 ((Alon, 2003), Theorem 9.3)**.** *Suppose $\frac{1}{\sqrt{n}} < \epsilon < \frac{1}{2}$. Then let $D \in \mathbb{R}^{n \times n}$ such that $D_{ii} = 1$ for all $i \in [n]$ and $|D_{ij}| < \epsilon$ for all $i \neq j$. Then*

$$\text{rank}(D) \geq \Omega\left(\frac{1}{\epsilon^2 \log(\frac{1}{\epsilon})} \log n\right). \tag{24}$$

It is useful to state a corollary, which generalizes the result to the setting where diagonal entries are bounded away from 0.

Corollary 3.2. *Let $\gamma \in (0, 1]$ and $\frac{\gamma}{\sqrt{n}} < \epsilon < \frac{\gamma}{2}$. Then let $C \in \mathbb{R}^{n \times n}$ such that $C_{ii} \geq \gamma$ for all $i \in [n]$ and $|C_{ij}| < \epsilon$ for all $i \neq j$. Then*

$$\operatorname{rank}(C) \geq \Omega\left(\frac{\gamma^2}{\epsilon^2 \log(\frac{\gamma}{\epsilon})} \log n\right). \tag{25}$$

Proof. Let D be the matrix obtained by scaling each entry in the i-th row of C by $\frac{1}{C_{ii}}$. Then $\operatorname{rank}(C) = \operatorname{rank}(D)$. Then $D_{ii} = 1$ for all $i \in [m]$ and $|D_{ij}| = \frac{|C_{ij}|}{C_{ii}} \leq \frac{\epsilon}{\gamma}$ for all $i \neq j$. Then the result follows by applying Lemma 3.1 with $\epsilon = \frac{\epsilon}{\gamma}$. □

Now notice that $|(B^T A)_{ii}| > 1 - \epsilon$ for all $i \in [m]$ (by (21)). Therefore, we can apply Corollary 3.2 to $B^T A$, taking $\gamma = 1 - \epsilon$. However, applying Corollary 3.2 directly like this does not provide the result we desire, since it only shows the existence of a single large off-diagonal entry in $B^T A$. (A sufficiently large such entry would also contradict (22), but applying Corollary 3.2 in this way would only imply a $\Omega_\epsilon(\log m)$ lower bound, losing dependence on k.)

Applying Corollary 3.2 to submatrices. Instead, our approach is to apply Corollary 3.2 to all principal submatrices of $C = B^T A$ of a certain size. For a set $R \subseteq [m]$, let $C_R \in \mathbb{R}^{|R| \times |R|}$ denote the submatrix given by taking the rows and columns of C in R. Then C_R also has diagonal entries at least $\gamma = 1 - \epsilon$, and $\operatorname{rank}(C_R) \leq \operatorname{rank}(C) \leq d$.

Setting $r = \frac{m}{4k+1}$, Corollary 3.2 implies that there exists a constant $\alpha > 0$ such that if $|R| = r$ and

$$\operatorname{rank}(C_R) < \alpha \frac{(1-\epsilon)^2 \log r}{(\frac{\epsilon}{k})^2 \log(\frac{k(1-\epsilon)}{\epsilon})}, \tag{26}$$

then C_R has an off-diagonal entry with absolute value at least $\frac{\epsilon}{k}$. (Also note that $\epsilon > \frac{k^{3/2}\sqrt{5}}{\sqrt{m}}$, $\implies \frac{\epsilon}{k} > \frac{1}{\sqrt{\frac{m}{4k+1}}} = \frac{1}{\sqrt{r}}$ for $k \geq 1$, satisfying the assumption in Corollary 3.2.) Equivalently, for any set $|R| \subseteq [m]$, there exist $i, j \in R$ with $i \neq j$ such that $|C_{ij}| > \frac{\epsilon}{k}$. In other words, among any r features, there must exist an interfering pair.

Graph reduction and Turán's theorem. We would now like to show that the presence of a large off-diagonal entry in any submatrix implies the existence of a row with many large off-diagonal entries. Consider the undirected graph $G = ([m], E)$ such that $(i, j) \in E$ for $i \neq j$ if and only if $|\langle b_i, a_j \rangle| > \frac{\epsilon}{k}$ or $|\langle b_j, a_i \rangle| > \frac{\epsilon}{k}$. Then we know from above that if

$$d \leq \alpha \frac{(1-\epsilon)^2 \log r}{(\frac{\epsilon}{k})^2 \log(\frac{k(1-\epsilon)}{\epsilon})}, \tag{27}$$

then for every $R \subseteq [m]$ with $|R| = r$, there exists $i, j \in R$ with $i \neq j$ such that $(i, j) \in E$. In other words, G has no independent set of size r, or equivalently, $\overline{G}$ has no clique of size r. Recall that Turán's theorem shows that a graph with m edges without an r-clique has at most $(1 - \frac{1}{r-1})\frac{m^2}{2} < (1 - \frac{1}{r})\frac{m^2}{2}$ edges. This implies that G has at least

$$\binom{m}{2} - \left(1 - \frac{1}{r}\right)\frac{m^2}{2} = \frac{m^2}{2r} - \frac{m}{2} \tag{28}$$

edges. Now note that every edge in G implies at least one off-diagonal entry in $B^T A$ with absolute value at least $\frac{\epsilon}{k}$. This implies the existence of a row with at least $\frac{m}{2r} - \frac{1}{2}$ off-diagonal entries with absolute value at least $\frac{\epsilon}{k}$ (by the pigeonhole principle). Setting $r = \frac{m}{4k+1}$ implies a row in $B^T A$ with $2k$ entries with absolute values at least $\frac{\epsilon}{k}$, providing the desired contradiction.

Therefore, we must have that

$$d > \alpha \frac{\log \frac{m}{4k+1}}{(\frac{\epsilon}{(1-\epsilon)k})^2 \log \frac{(1-\epsilon)k}{\epsilon}} = \Omega_\epsilon\left(\frac{k^2}{\log k} \log \frac{m}{k}\right), \tag{29}$$

as desired.

4. The geometry of feature directions

One benefit of the linear representation hypothesis is that it provides a geometric conception of how language models represent concepts. In particular, features are thought of as directions in activation space. This allows us to state geometric relationships between different features. For example, it is often suggested that unrelated features are represented by *orthogonal* directions. In this section, we establish what is implied and not implied about the geometry of feature vectors under our theoretical framework.

Proposition 4.1 demonstrates two unintuitive facts. First, a feature can be represented and accessed by *different* vectors; in fact, a feature's representation and probe directions can be almost perfectly orthogonal. Second, different features can have representation vectors that are almost perfectly in the same direction; likewise, different features can have probe vectors that are almost perfectly in the same direction.

Proposition 4.3 then establishes that this unusual behavior can be eliminated if we further restrict the norms of both the representation and probe vectors. In particular, if we bound the norm of probe and representation vectors to be near 1, probe and representation directions must be closely aligned, and different features must represented and probed by approximately orthogonal directions.

Proposition 4.1. *Consider constants $\epsilon, \delta \in (0, 1)$. Then for $d = O_{\epsilon,\delta}(k^2 \log m)$, there exists $A, B \in \mathbb{R}^{d \times m}$ such that*

$$\left\| B^T A z - z \right\|_\infty < \epsilon \tag{30}$$

for all k-sparse $z \in [-1,1]^m$, and where

(i) $|\langle \frac{a_i}{\|a_i\|}, \frac{b_i}{\|b_i\|} \rangle| < \delta + o(1)$ *for all* $i \in [m]$,

(ii) $\langle \frac{a_i}{\|a_i\|}, \frac{a_j}{\|a_j\|} \rangle > 1 - \delta - o(1)$ *for all* $i \neq j$,

(iii) $\langle \frac{b_i}{\|b_i\|}, \frac{b_j}{\|b_j\|} \rangle > 1 - \delta - o(1)$ *for all* $i \neq j$.

We defer the proof of Proposition 4.1 to the appendix.

Remark 4.2. In practice, different representation and probe vectors are seen in different decoder and encoder weights in SAEs, respectively (e.g., (Templeton et al., 2024; Paulo and Belrose, 2025)). This had led to some confusion about which sets of weights are the "true" feature vectors. Proposition 4.1 illustrates that under the LRH, features have both a representation vector and a probe direction, and that these directions can be *distinct*.

We now show how the "unusual" behavior in Proposition 4.1 is controlled by the norm of the representation and probe vectors.

Proposition 4.3. *Consider any constants $\epsilon > 0, \gamma \geq 1$. Then if $\|a_i\|, \|b_i\| \leq \gamma$ for all $i \in [m]$ and*

$$\left\| B^T A z - z \right\|_\infty < \epsilon \tag{31}$$

for all k-sparse $z \in [-1,1]^m$, it holds that

(i) $\langle \frac{a_i}{\|a_i\|}, \frac{b_i}{\|b_i\|} \rangle \geq \frac{1-\epsilon}{\gamma^2}$ *for all* $i \in [m]$,

(ii) $\langle \frac{a_i}{\|a_i\|}, \frac{a_j}{\|a_j\|} \rangle \leq \frac{\epsilon\gamma^2}{1-\epsilon} + \sqrt{1 - \frac{(1-\epsilon)^2}{\gamma^4}}$ *for all* $i \neq j$,

(iii) $\langle \frac{b_i}{\|b_i\|}, \frac{b_j}{\|b_j\|} \rangle \leq \frac{\epsilon\gamma^2}{1-\epsilon} + \sqrt{1 - \frac{(1-\epsilon)^2}{\gamma^4}}$ *for all* $i \neq j$.

We defer the proof of Proposition 4.3 to the appendix.

Notice, in particular, taking $\epsilon \to 0, \gamma = 1$ in Proposition 4.3, we get that

(i) $\langle \frac{a_i}{\|a_i\|}, \frac{b_i}{\|b_i\|} \rangle \to 1$ for all $i \in [m]$,

(ii) $\langle \frac{a_i}{\|a_i\|}, \frac{a_j}{\|a_j\|} \rangle \to 0$ for all $i \neq j$, and

(iii) $\langle \frac{b_i}{\|b_i\|}, \frac{b_j}{\|b_j\|} \rangle \to 0$ for all $i \neq j$.

This shows that when representation and probe vectors are restricted to have unit norm, the clean behavior of equal representation and probe directions emerges (meaning that each feature is given by a single direction), and that different features have orthogonal directions.

5. Discussion

We introduced a mathematical framework in which we can formally state two pieces of the Linear Representation Hypothesis. First, that in intermediate layers of language models, features are *represented* linearly. Second, that these features can then be *accessed* linearly. We can then formally establish the number of features d neurons can linearly represent such that they can be accessed linearly. We show a quantitative gap between what can be achieved with linear and non-linear decoding.

Our work suggests a number of interesting further directions. First, while we assume linear representation and then establish results for non-linear and linear accessibility, it is interesting to consider when linear accessibility is possible under non-linear representation. Indeed, this would help establish if previous layers can arrange features in more subtle ways while still allowing efficient access by later layers. Similarly, analysis of more sophisticated decoders (like two-layer neural networks) would be interesting. More generally, other works have proposed alternate representation hypotheses (Engels et al., 2024; Csordás et al., 2024; Kantamneni and Tegmark, 2025). Formalizing these hypotheses and understanding their representational capacity could establish further insight into the plausibility of these hypotheses in comparison to the LRH. Finally, our analysis is limited to how a single layer of a language model stores features. Formalization of how multiple layers can leverage linear representations would be interesting. In particular, while we show that an exponential number of features can be represented and accessed linearly, the next layer of a neural network (say with d_2 neurons) consists only of d_2 linear probes, suggesting that each additional layer does more than extracting individual features. Formalizing what this additional layer of computation can accomplish given a linear representation (perhaps in combination with mechanisms like attention) would shed light on how language models make computations.

At a higher level, our work aims to help spur further work formalizing empirically-motivated hypotheses, and formally deriving implications of different theories of language models.

References

G. Alain and Y. Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

N. Alon. Problems and results in extremal combinatorics—i. *Discrete Mathematics*, 273(1-3):31–53, 2003.

S. Arora, Y. Li, Y. Liang, T. Ma, and A. Risteski. Linear algebraic structure of word senses, with applications to

polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495, 2018.

E. J. Candes and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE transactions on information theory*, 52(12):5406–5425, 2006.

E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

R. Chen, A. Arditi, H. Sleight, O. Evans, and J. Lindsey. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*, 2025.

V. Costa, T. Fel, E. S. Lubana, B. Tolooshams, and D. Ba. From flat to hierarchical: Extracting sparse representations with matching pursuit. *arXiv preprint arXiv:2506.03093*, 2025.

R. Csordás, C. Potts, C. D. Manning, and A. Geiger. Recurrent neural networks learn to store and generate sequences using non-linear representations. *arXiv preprint arXiv:2408.10920*, 2024.

H. Cunningham, A. Ewart, L. Riggs, R. Huben, and L. Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

D. L. Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

N. Elhage, T. Hume, C. Olsson, N. Schiefer, T. Henighan, S. Kravec, Z. Hatfield-Dodds, R. Lasenby, D. Drain, C. Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

J. Engels, E. J. Michaud, I. Liao, W. Gurnee, and M. Tegmark. Not all language model features are one-dimensionally linear. *arXiv preprint arXiv:2405.14860*, 2024.

L. Gao, T. D. la Tour, H. Tillman, G. Goh, R. Troll, A. Radford, I. Sutskever, J. Leike, and J. Wu. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

S. S. R. Hindupur, E. S. Lubana, T. Fel, and D. Ba. Projecting assumptions: The duality between sparse autoencoders and concept geometry. *arXiv preprint arXiv:2503.01822*, 2025.

W. B. Johnson, J. Lindenstrauss, et al. Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26(189-206):1, 1984.

S. Kantamneni and M. Tegmark. Language models use trigonometry to do addition. *arXiv preprint arXiv:2502.00873*, 2025.

B. Z. Li, Z. C. Guo, and J. Andreas. (how) do language models track state? *arXiv preprint arXiv:2503.02854*, 2025.

J. Lindsey, W. Gurnee, E. Ameisen, B. Chen, A. Pearce, N. L. Turner, C. Citro, D. Abrahams, S. Carter, B. Hosmer, J. Marcus, M. Sklar, A. Templeton, T. Bricken, C. McDougall, H. Cunningham, T. Henighan, A. Jermyn, A. Jones, A. Persic, Z. Qi, T. B. Thompson, S. Zimmerman, K. Rivoire, T. Conerly, C. Olah, and J. Batson. On the biology of a large language model. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013a.

T. Mikolov, W.-t. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751, 2013b.

R. Movva, K. Peng, N. Garg, J. Kleinberg, and E. Pierson. Sparse autoencoders for hypothesis generation. *arXiv preprint arXiv:2502.04382*, 2025.

N. Nanda, A. Lee, and M. Wattenberg. Emergent linear representations in world models of self-supervised sequence models. *arXiv preprint arXiv:2309.00941*, 2023.

N. Panickssery, N. Gabrieli, J. Schulz, M. Tong, E. Hubinger, and A. M. Turner. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2023.

K. Park, Y. J. Choe, and V. Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

G. Paulo and N. Belrose. Sparse autoencoders trained on the same data learn different features. *arXiv preprint arXiv:2501.16615*, 2025.

A. Templeton, T. Conerly, J. Marcus, J. Lindsey, T. Bricken, B. Chen, A. Pearce, C. Citro, E. Ameisen, A. Jones, et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. transformer circuits thread, 2024.

A. M. Turner, L. Thiergart, G. Leech, D. Udell, J. J. Vazquez, U. Mini, and M. MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

M. Wang, T. D. la Tour, O. Watkins, A. Makelov, R. A. Chi, S. Miserendino, J. Heidecke, T. Patwardhan, and D. Mossing. Persona features control emergent misalignment. *arXiv preprint arXiv:2506.19823*, 2025.

A. Zou, L. Phan, S. Chen, J. Campbell, P. Guo, R. Ren, A. Pan, X. Yin, M. Mazeika, A.-K. Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A. Activation functions and binary classification

In this section, we establish a generalization of our lower bound, which simultaneous resolves two additional questions: (1) What happens when we add an activation function or bias? (2) What about when features only take on binary values (as may be the case for many common features)?

In particular, we are curious how bounds may change if we allow $g(x) = \mathrm{ReLU}(Wx + b)$, since individual layers in a neural network consist of a linear transformation plus a non-linear activation. We are also interested in the setting in which $z \in \{0,1\}^m$ (i.e., features are binary).

Here, we prove that when feature representations are restricted to the Boolean cube $\{0,1\}^m$, $\Omega(k^2 \log(m/k))$ embedding dimensions are required to linearly separate embeddings with and without a feature. A corollary is that adding a bias and monotonic activation function does not increase the asymptotic capacity of embeddings.

Theorem A.1. *Suppose that there exist $A, B \in \mathbb{R}^{d\times m}$ and $t \in \mathbb{R}^m$ such that for all $i \in [m]$ and k-sparse $z \in \{0,1\}^m$,*

$$(B^T Az)_i > t_i \tag{32}$$

if and only if $z_i = 1$. Then for $k < \sqrt{m}$,

$$d = \Omega\left(\frac{k^2}{\log k} \log \frac{m}{k}\right). \tag{33}$$

It's worth noting here that taking $\epsilon < \frac{1}{2}$ in Theorem 1.2 provides a nearly-matching upper bound of $\Omega(k^2 \log m)$, since $\epsilon < \frac{1}{2}$ implies that $(B^T Az)_i < \frac{1}{2}$ when $z_i = 0$ and $(B^T Az)_i > \frac{1}{2}$ when $z_i = 1$. Similarly, Theorem A.1 implies Theorem 1.3 in the case $\epsilon < \frac{1}{2}$. However, the lower bound in Theorem 1.3 does not directly imply Theorem A.1 since we do not require that $\left\|B^T Az - z\right\|_\infty$ is small here, but rather that Az linearly separates positive and negative examples for each feature. That being said, the proof is very similar to that of Theorem 1.3, relying on the same overall strategy.

The following corollary makes the simple observation that if a feature cannot be linearly separated, then it also cannot be separated by adding a bias and monotonic activation function.

Corollary A.2. *Suppose that there exist $A, W \in \mathbb{R}^{d\times m}$, $b \in \mathbb{R}^m$, and a monotonically increasing activation function $\sigma : \mathbb{R} \to \mathbb{R}$ such that for all $i \in [m]$, k-sparse $z \in \{0,1\}^m$, and $x = Az$,*

$$\sigma(W^T x + b) > 0 \tag{34}$$

if and only if $z_i = 1$. (Here, σ is applied component-wise.) Then for $k < \sqrt{m}$,

$$d = \Omega\left(\frac{k^2}{\log k} \log \frac{m}{k}\right). \tag{35}$$

Proof. Consider any $i \in [m]$ and $z \in \{0,1\}^m$ such that $z_i = 1$ and z' such that $z'_i = 0$. Then $\sigma(W^T Az + b) > \sigma(W^T Az' + b)$. This implies that $(W^T Az)_i > (W^T Az')_i$. Therefore, there exists a threshold t_i such that for all $z \in \{0,1\}^m$, $(W^T Az)_i > t_i$ if and only if $z_i = 1$. The result follows by applying Theorem A.1. □

The proof establishes that probes of the form $g_i(x) = \mathrm{ReLU}(W^T x + b)$ implies the existence of a linear probe. In particular, this further establishes that for $\epsilon < \frac{1}{2}$, the bound in Theorem 1.3 cannot be lowered asymptotically by expanding the class of probes to allow a bias and activation function.

B. Proofs

Proof of Proposition 4.1. We give a construction. Let $c_1, c_2, \cdots, c_m, a^*, b^*$ be the columns of a μ-incoherent matrix $\mathbb{R}^{d\times m+2}$. Then set

$$a_i = c_i + \lambda a^* \tag{36}$$

$$b_i = c_i + \lambda b^* \tag{37}$$

for $\lambda = \sqrt{\frac{1}{\delta} - 1}$. We begin by showing that for an appropriately chosen μ, $\left\|B^T Az - z\right\|_\infty < \epsilon$ for all k-sparse $z \in [-1,1]^m$. Now let $\hat{z} = B^T Az$ for any $z \in [-1,1]^m$ with support T. Then

$$\hat{z}_i = \left\langle b_i, \sum_{j=1}^m z_j a_j \right\rangle = z_i \langle b_i, a_i\rangle + \sum_{j\in T\setminus\{i\}} z_j \langle b_i, a_j\rangle. \tag{38}$$

Therefore,

$$|\hat{z}_i - z_i| \leq |z_i(\langle b_i, a_i\rangle - 1)| + \sum_{j\in T\setminus\{i\}} |z_j\langle b_i, a_j\rangle| \quad (39)$$

$$\leq |z_i(\langle b_i, a_i\rangle - 1)| + \sum_{j\in T\setminus\{i\}} |\langle b_i, a_j\rangle|, \quad (40)$$

where the last inequality follows from observing that $z_j \in [-1, 1]$ for all $j \in [m]$. Now observe that

$$|z_i(\langle b_i, a_i\rangle - 1)| \quad (41)$$

$$= |z_i||\langle c_i + \lambda b^*, c_i + \lambda a^*\rangle - 1| \quad (42)$$

$$\leq |z_i|\left(|\langle c_i, \lambda a^*\rangle| + |\langle \lambda b^*, c_i\rangle| + |\langle \lambda b^*, \lambda a^*\rangle|\right) \quad (43)$$

$$= |z_i|(2\lambda + \lambda^2)\mu \quad (44)$$

$$< |z_i|(1 + \lambda)^2\mu, \quad (45)$$

where (43) follows because $|\langle c_i, c_i\rangle| = 1$ and (44) follows from μ-incoherence. Similarly,

$$|\langle b_i, a_j\rangle| = |\langle c_i + \lambda b^*, c_i + \lambda a^*\rangle| \quad (46)$$

$$= |\langle c_i, c_j\rangle| + |\langle c_i, \lambda a^*\rangle| + |\langle \lambda b^*, c_i\rangle| + |\langle \lambda b^*, \lambda a^*\rangle| \quad (47)$$

$$\leq (1 + 2\lambda + \lambda^2)\mu = (1 + \lambda)^2\mu. \quad (48)$$

This implies that

$$|z_i(\langle b_i, a_i\rangle - 1)| + \sum_{j\in T\setminus\{i\}} |\langle b_i, a_j\rangle| < k(1 + \lambda)^2\mu, \quad (49)$$

where we use the fact that if $|z_i| > 0$, then $|T \setminus \{i\}| \leq k - 1$. Therefore, it suffices to take $\mu = \frac{\epsilon}{k(1+\lambda^2)}$ to ensure that $\|\hat{z} - z\|_\infty < \epsilon$. From Lemma 2.2, there exists a μ-incoherent matrix in $\mathbb{R}^{d\times m+2}$ for

$$d = O\left(\frac{k^2(1 + \lambda)^4}{\epsilon^2}\log(m + 2)\right) \quad (50)$$

$$= O\left(\frac{k^2(1 + \lambda)^4}{\epsilon^2}\log m\right) \quad (51)$$

$$= O_{\epsilon,\delta}(k^2 \log m). \quad (52)$$

Establishing (i). We now show that the construction satisfies (i). Observe that for all $i \in [m]$,

$$\left|\left\langle \frac{a_i}{\|a_i\|}, \frac{b_i}{\|b_i\|}\right\rangle\right| \quad (53)$$

$$= \frac{|\langle c_i + \lambda a^*, c_i + \lambda b^*\rangle|}{\|a_i\|\,\|b_i\|} \quad (54)$$

$$\leq \frac{|\langle c_i, c_i\rangle| + |\langle c_i, \lambda b^*\rangle| + |\langle \lambda a^*, c_i\rangle| + |\langle \lambda a^*, \lambda b^*\rangle|}{\|a_i\|\,\|b_i\|} \quad (55)$$

$$\leq \frac{1 + (2\lambda + \lambda^2)\mu}{\|a_i\|\,\|b_i\|} = \frac{1 + \frac{(2\lambda+\lambda^2)\epsilon}{k(1+\lambda^2)}}{\|a_i\|\,\|b_i\|} = \frac{1 + o(1)}{\|a_i\|\,\|b_i\|}. \quad (56)$$

Note that

$$\langle a_i, a_i\rangle = \langle c_i + \lambda a^*, c_i + \lambda a^*\rangle \quad (57)$$

$$= \langle c_i, c_i\rangle + 2\langle c_i, \lambda a^*\rangle + \langle \lambda a^*, \lambda a^*\rangle \quad (58)$$

$$\leq 1 + \lambda^2 - 2\lambda\mu = 1 + \lambda^2 - \frac{2\lambda\epsilon}{k(1 + \lambda^2)} \quad (59)$$

$$= 1 + \lambda^2 - o(1) \quad (60)$$

Therefore, $\|a_i\| \leq \sqrt{1 + \lambda^2 - o(1)}$. Analogously, $\|b_i\| \leq \sqrt{1 + \lambda^2 - o(1)}$. So we have, plugging into (56),

$$\left\langle \frac{a_i}{\|a_i\|}, \frac{b_i}{\|b_i\|}\right\rangle \leq \frac{1 + o(1)}{1 + \lambda^2 - o(1)} \quad (61)$$

$$= \frac{1}{1 + \lambda^2} + o(1) = \delta + o(1), \quad (62)$$

where we recall that we set $\lambda = \sqrt{\frac{1}{\delta} - 1}$.

Establishing (ii,iii). We now show that the construction satisfies (ii). (iii) follows analogously. For all $i \neq j$,

$$\left|\left\langle \frac{a_i}{\|a_i\|}, \frac{a_j}{\|a_j\|}\right\rangle\right| \quad (63)$$

$$= \frac{|\langle c_i + \lambda a^*, c_j + \lambda a^*\rangle|}{\|a_i\|\,\|a_j\|} \quad (64)$$

$$= \frac{\langle c_i, c_j\rangle + \langle c_i, \lambda a^*\rangle + \langle \lambda a^*, c_j\rangle + \langle \lambda a^*, \lambda a^*\rangle}{\|a_i\|\,\|a_j\|} \quad (65)$$

$$\geq \frac{\lambda^2 - (1 + 2\lambda)\mu}{1 + \lambda^2 + o(1)} \quad (66)$$

$$= \frac{\lambda^2 - o(1)}{1 + \lambda^2 + o(1)} \quad (67)$$

$$= 1 - \frac{1}{1 + \lambda^2} - o(1) = 1 - \delta - o(1), \quad (68)$$

where we again recall that where we recall that $\lambda = \sqrt{\frac{1}{\delta} - 1}$.

□

Proof of Proposition 4.3. We must have that $\langle a_i, b_i\rangle \geq 1 - \epsilon$ since $\left\|B^T A e_i - e_i\right\|_\infty \leq \epsilon$. It follows directly that

$$\langle a_i, b_i\rangle > 1 - \epsilon \implies \left\langle \frac{a_i}{\|a_i\|}, \frac{b_i}{\|b_i\|}\right\rangle \geq \frac{1 - \epsilon}{\gamma^2}, \quad (69)$$

establishing (i).

We now show (ii). By Cauchy-Schwarz, $\|a_i\|\,\|b_i\| \geq \langle a_i, b_i\rangle > 1 - \epsilon$ for all $i \in [m]$. So since $\|a_i\|, \|b_i\| \leq \lambda$ for all $i \in [m]$, we have that $\|a_i\|, \|b_i\| \geq \frac{1-\epsilon}{\gamma^2}$ for all $i \in [m]$. Therefore, we have that

$$\left|\left\langle \frac{b_i}{\|b_i\|}, \frac{a_j}{\|a_j\|}\right\rangle\right| \leq |\langle b_i, a_j\rangle|\frac{\gamma^2}{1 - \epsilon} < \frac{\epsilon\gamma^2}{1 - \epsilon}. \quad (70)$$

Let $\tilde{a_i}$ denote $\frac{a_i}{\|a_i\|}$ and $\tilde{b_i}$ denote $\frac{b_i}{\|b_i\|}$ for all $i \in [m]$. Then write

$$\tilde{a_i} = \alpha_i \tilde{b_i} + \tilde{a_{i\perp}} \tag{71}$$

$$\tilde{a_j} = \alpha_j \tilde{b_i} + \tilde{a_{j\perp}}, \tag{72}$$

where $\alpha_i = \langle \tilde{a_i}, \tilde{b_i} \rangle$,$\alpha_j = \langle \tilde{a_j}, \tilde{b_j} \rangle$, and $\tilde{a_{i\perp}}, \tilde{a_{j\perp}} \perp \tilde{b_i}$. Then

$$\langle \tilde{a_i}, \tilde{a_j} \rangle = \alpha_i \alpha_j + \langle \tilde{a_{i\perp}}, \tilde{a_{j\perp}} \rangle. \tag{73}$$

By Cauchy-Schwarz,

$$|\langle \tilde{a_{i\perp}}, \tilde{a_{j\perp}} \rangle| \leq \|\tilde{a_{i\perp}}\| \left\|\tilde{a_{j\perp}}\right\| = \sqrt{(1-\alpha_i^2)(1-\alpha_j^2)}. \tag{74}$$

Therefore,

$$|\langle \tilde{a_i}, \tilde{a_j} \rangle| \leq |\alpha_i||\alpha_j| + \sqrt{(1-\alpha_i^2)(1-\alpha_j^2)}. \tag{75}$$

We have that $1 \geq \alpha_i > \frac{1-\epsilon}{\gamma^2}$ and $0 < \alpha_j < \frac{\epsilon\gamma^2}{1-\epsilon}$. Therefore,

$$|\langle \tilde{a_i}, \tilde{a_j} \rangle| \leq \frac{\epsilon\gamma^2}{1-\epsilon} + \sqrt{1 - \frac{(1-\epsilon)^2}{\gamma^4}}, \tag{76}$$

as desired. (iii) follows analogously. □

Proof of Theorem A.1. Suppose that there existed $A, B \in \mathbb{R}^{d\times m}$ and $t \in \mathbb{R}^m$ such that for all $i \in [m]$ and k-sparse $z \in \{0,1\}^m$, $(B^T Az)_i > t_i$ if and only if $z_i = 1$. Then notice that we could scale the i-th column of B and γ_i by any constant positive factor to get another solution. In particular, we can obtain a solution where $B^T A$ is 1 on the diagonal. Therefore, it suffices to show that for $d \leq O(\frac{k^2}{\log k} \log \frac{m}{k})$, there does not exist B, A, t satisfying (32) and such that $(B^T A)_{ii} = 1$ for all $i \in [m]$.

Then we have that $t_i \leq 1$, since $(B^T Ae_i) > t_i$ Then, along the lines of the proof of Theorem 1.3, it suffices to find $i \in [m]$ and $T^* \subseteq [m] \setminus \{i\}$ with $|T^*| = 2k$ and

$$|\langle b_i, a_j \rangle| > \frac{1}{k} \tag{77}$$

for all $j \in T^*$. Indeed, this would imply the existence of a pair $i \in [m], T \subseteq [m] \setminus \{i\}$ (with $|T| = k$), such that

$$(B^T Az)_i > k \cdot \frac{1}{k} = 1 \geq t_i, \tag{78}$$

giving a contradiction.

We will again apply Lemma 3.1 to submatrices of C of size r. Indeed, Lemma 3.1 implies that there exists a constant $\alpha > 0$ such that if $|R| = r$ and

$$\operatorname{rank}(C_R) < \alpha \frac{\log r}{(\frac{1}{k})^2 \log k} = \alpha \frac{k^2}{\log k} \log r, \tag{79}$$

then C_R has an off-diagonal entry with absolute value at least $\frac{1}{k}$. Then we can proceed as in the proof of Theorem 1.3, taking $r = \frac{m}{4k+1}$ and applying Turán's theorem. □