On the Expressive Flexibility of Self-Attention Matrices

Valerii Likhosherstov\textsuperscript{1*}, Krzysztof Choromanski\textsuperscript{2*}, Adrian Weller\textsuperscript{1,3}
\textsuperscript{1}University of Cambridge
\textsuperscript{2}Google Brain
\textsuperscript{3}The Alan Turing Institute
vl304@cam.ac.uk

Abstract
Transformer networks are able to capture patterns in data coming from many domains (text, images, videos, proteins, etc.) with little or no change to architecture components. We perform a theoretical analysis of the core component responsible for signal propagation between elements, i.e. the self-attention matrix. We ask the following questions: Can a self-attention matrix approximate arbitrary patterns? How small is the query dimension \(d\) required for such an approximation? Our first result shows that the task of deciding whether approximation of a given pattern is possible or not is \textsc{NP-hard} for a fixed \(d > 1\). In practice, the self-attention matrix typically exhibits two properties: it is sparse, and it changes dynamically depending on the input to the module. Motivated by this observation, we show that the self-attention matrix can provably approximate sparse matrices. While the parameters of self-attention are fixed, various sparse matrices can be approximated by only modifying the inputs. Our proof is based on the random projection technique and uses the seminal Johnson-Lindenstrauss lemma. In particular, we show that, in order to approximate any sparse matrix up to a given precision defined in terms of preserving matrix element ratios, \(d\) grows only logarithmically with the sequence length \(n\) (i.e. \(d = \mathcal{O}(\log n)\)).

Introduction
Transformer networks have demonstrated strong performance in the area of large-scale deep learning, coming close to or beating the state of the art in a wide range of tasks. Initially proposed in the context of neural machine translation (Vaswani et al. 2017), Transformers were found to generalize well across a variety of natural language processing tasks when pretrained on large text corpora (Devlin et al. 2019; Radford et al. 2019; Brown et al. 2020). These successes facilitated the application of Transformers in other domains. For instance, in biology, Transformers pretrained on large corpora of proteins were shown to predict proteins’ structure and function (Elnaggar et al. 2019; Rives et al. 2021), and to generate protein sequences with specific properties (Madani et al. 2020). Another exciting advancement was the emergence of Vision Transformers (Dosovitskiy et al. 2021) and, later, Video Vision Transformers (Arnab et al. 2021).

Thus, Transformers appear domain-agnostic and can learn any priors once a suitable large-scale dataset is provided. Finally, Transformers were recently shown to be applicable for end-to-end training on large-scale multimodal data of images with textual annotations extracted from the Internet (Radford et al. 2021; Jia et al. 2021). The resulting models are highly generalizable and perform very well in zero-shot classification from scratch, and when fine-tuned on standard benchmarks of a smaller scale.

The omnivorous nature of these models suggests that Transformers and their core component, self-attention, have an inherent ability to capture useful patterns in data regardless of the domain. A thorough analysis is required to gain a deeper understanding of this remarkable phenomenon. We take a step in this direction by analyzing the expressiveness of the self-attention module.

In self-attention, dependencies between elements of the input are propagated via a self-attention matrix, which can be thought of as an input-dependent linear projection applied to the input. By the definition, this right stochastic matrix (i.e. having nonnegative elements with rows summing up to 1) encodes input-dependent patterns in the data. Therefore, we aim to analyze the expressiveness of this matrix, to understand how flexible these input-dependent patterns can be. Importantly, we consider the setup when the \textbf{query dimension of self-attention} \(d\) is much smaller than the sequence length \(n\) aiming to characterize relationships between the two. We focus on the case when \(d < n\) because it is typical in practice, e.g. \(d = 64, n = 512\) in BERT (Devlin et al. 2019). Also, the case \(d \geq n\) can model any right stochastic matrix as shown in (Bhojanapalli et al. 2020). Finally, smaller \(d\) facilitates computational efficiency of Transformers, which are notorious for their high compute demand and CO\(_2\) footprint (Strubell, Ganesh, and McCallum 2019).

Our first contribution shows that understanding which right stochastic matrices can be approximated by a self-attention matrix is a challenging task. Namely, we show that the decision task which accepts the right stochastic matrix and outputs whether approximation is possible or not for a given \(d \geq 1\) is \textsc{NP-hard}. We further show that the case \(d = 1\) is tractable and give an algorithm for the decision problem. Our derivations are inspired by the theory of dot product graphs (Kang and Müller 2012).

Since it is hard to tackle the general setup of approxim-
Let \( n \) be the length of a processed sequence and \( d_{hid} \) be the size of a hidden representation passed through the neural network. We define the unnormalized self-attention matrix as a parametrized mapping from the current hidden state \( X \in \mathbb{R}^{n \times d_{hid}} \) into \( \mathbb{R}^{n \times n} \). The mapping depends on two learnable parameter matrices \( W_Q, W_K \in \mathbb{R}^{d_{hid} \times d}, d \leq d_{hid} \), where \( d \leq d_{hid} \) is the query dimension. The mapping is defined as

\[
\text{USAM}(X; d, d_{hid}, W_Q, W_K) = \exp[ X W_Q W_K^\top X^\top], \tag{1}
\]

where \( \exp[\cdot] \) is an elementwise exponent. Next, we define the (normalized) self-attention matrix as

\[
\text{SAM}(X; d, d_{hid}, W_Q, W_K) = \text{RN}(\text{USAM}(X; d, d_{hid}, W_Q, W_K)). \tag{2}
\]

Here, \( \text{RN}(\cdot) \) (row normalization) divides each row by the sum of its elements, so that the result is right stochastic meaning that all rows are nonnegative and sum to 1.

Finally, self-attention is defined as a parametrized mapping from \( X \) into \( \mathbb{R}^{n \times d} \) with parameters \( W_Q, W_K, W_V \in \mathbb{R}^{d_{hid} \times d} \). It has the form:

\[
\text{SA}(X; d, d_{hid}, W_Q, W_K, W_V) = \text{SAM}(X; d, d_{hid}, W_Q, W_K) X W_V. \tag{3}
\]

Self-attention has the form of a differentiable dictionary, where the output at each position \( 1 \leq i \leq n \) is a sum of all values \( W_Q^i X_{i'}, 1 \leq i' \leq n \), weighted proportionally to exponentiated dot products of the query \( W_Q^i X_i \) and the key vectors \( W_K^{i'} X_{i'} \). Usually (Vaswani et al. 2017), these dot products are also divided by \( \sqrt{d} \), since this empirically facilitates stable training. Without loss of generality, we do not include this linear scaling factor in our definition (1), since it can be fused into one of the matrices \( W_Q \) or \( W_K \).

**NP-Hardness of Approximation**

Below we define a notion of \( \text{SAM}(X; d, d_{hid}, W_Q, W_K) \) weakly approximating a right stochastic matrix \( A = \text{RN}(S) \),
where $S \in \{0,1\}^{n \times n}$ is a matrix of zeros and ones with at least one 1 in every row so that its row normalization is well-defined. The goal of this notion is to capture a minimal set of conditions which make it possible to say that $\text{SAM}(X; d, d_{hid}, W_Q, W_K)$ approximates $A$:

**Definition 0.1.** Suppose $S \in \{0,1\}^{n \times n}$ has at least one 1 in every row and $X \in \mathbb{R}^{n \times d_{hid}}$, $W_Q, W_K \in \mathbb{R}^{d_{hid} \times d}$, $d_{hid} \geq d$, are some matrices. $M = \text{SAM}(X; d, d_{hid}, W_Q, W_K)$ weakly approximates $A = \text{RN}(S)$ when

$$\forall 1 \leq i \leq n : \min_{j, A_{i,j}>0} M_{i,j} > \max_{j, A_{i,j}=0} M_{i,j},$$

(4)

where we assume that $\max$ over an empty set is $-\infty$. If such $X, W_Q, W_K$ exist for $A$, then we say that $A$ can be $(d, d_{hid})$-weakly SAM-approximated.

According to this definition, $M$ approximates $A$ if in each row, the values of $M$ corresponding to zero positions in $A$ are smaller than those corresponding to nonzero positions in $A$. For integers $d_{hid} \geq d \geq 1$, by $(d, d_{hid})$-WEAK-SAM-APPROX we denote the algorithmic problem of deciding, given an integer $n > 0$ and a matrix $S \in \{0,1\}^{n \times n}$ vertices with at least one 1 in each row, whether $\text{RN}(S)$ can be $(d, d_{hid})$-weakly SAM-approximated or not. Consequently, $(d, d_{hid})$-WEAK-SAM-APPROX returns either True if the approximation is possible or False otherwise. Our first contribution is as follows:

**Theorem 0.2.** $(d, d_{hid})$-WEAK-SAM-APPROX is NP-hard for all $d_{hid} \geq d > 1$.

**Proof.** See Appendix. \(\square\)

Theorem 0.2 states that the approximation task with a small $d$ (since the size of the graph $n$ can be arbitrarily large in $(d, d_{hid})$-WEAK-SAM-APPROX) is very hard and possibly unsolvable in polynomial time. This supports our initial claim that the problem of right stochastic matrix approximation by self-attention is a challenging task. The proof is inspired by a similar result in the theory of dot product graphs (Kang and Müller 2012).

For completeness, we also show that the case of $(1, d_{hid})$-WEAK-SAM-APPROX, $d_{hid} \geq 1$, has a positive characterization:

**Theorem 0.3.** Algorithm 1 solves $(1, d_{hid})$-WEAK-SAM-APPROX in the time which is polynomial of the input’s size.

**Proof.** See Appendix. \(\square\)

Since the general case of right stochastic matrix approximation by self-attention is very hard and potentially unsolvable, in the next section we focus on a special case of sparse matrices. For this class of matrices, we derive a lower bound on $d$ which gives a provable approximation.

**Approximating Sparse Matrices by Self-Attention Matrix**

**The Main Result**

We will call the square matrix $k$-nonzero-bounded if for each row or column of the matrix, the total number of nonzero elements is no more than $k$.

**Algorithm 1:** Algorithm for solving $(1, d_{hid})$-WEAK-SAM-APPROX ($\mathcal{V}$ denotes $\{1,\ldots,n\}$). **Input:** $S \in \{0,1\}^{n \times n}$ with at least one 1 in each row. **Output:** True or False.

1. For $1 \leq i \leq n$, set $I^{(i)} = \{1 \leq j \leq n \mid M_{i,j} = 1\}$.
2. Set $K_0 = \{1 \leq i \leq n \mid I^{(i)} \neq \{1,\ldots,n\}\}$. If $K_0 = \emptyset$, return True.
3. Let $H$ be a graph on $K_0$ as vertices, such that there is an edge between $v$, $w$ if $I^{(v)} \subseteq I^{(w)}$ or $I^{(w)} \subseteq I^{(v)}$.
4. If $H$ doesn’t consist of two (possibly empty) connected components $K_1, K_2 \subseteq K_0$, return False.
5. For $1 \leq i \leq n$, set $I^{(i)} = I^{(i)}$ if $i \notin K_2$ and $I^{(i)} = \mathcal{V} \setminus I^{(i)}$ if $i \in K_2$.
6. If there exist $i, j \in K_0$ such that neither $I^{(i)} \subseteq I^{(j)}$, nor $I^{(j)} \subseteq I^{(i)}$, return False.
7. If $d_{hid} > 1$, return True.
8. Let $i_1, \ldots, i_n$ be an ordering of $1, \ldots, n$ such that $I^{(i_1)} \subseteq \cdots \subseteq I^{(i_n)}$.
9. If $K_2 \subseteq I^{(i_r)}$, or $I^{(i_{n-r})} \subseteq K_2$ or there is $1 \leq r < n$ such that $I^{(i_{r+1})} \subseteq K_2 \subseteq I^{(i_{n-r})}$, return True.
10. If $K_2 \subseteq \mathcal{V} \setminus I^{(i_r)}$, or $\mathcal{V} \setminus I^{(i_{n-r})} \subseteq K_2$, or there is $1 \leq r < n$ such that $\mathcal{V} \setminus I^{(i_{r+1})} \subseteq K_2 \subseteq \mathcal{V} \setminus I^{(i_{n-r})}$, return True.
11. Return False.

Apart from the notion of the bounded number of nonzero elements, we also define matrices with elements of a bounded variation. For $\gamma \geq 1$, we call the matrix $A \in \mathbb{R}^{n \times n}$ with nonnegative elements $\gamma$-variation-bounded, if for every row $1 \leq i \leq n$ and every column indices $1 \leq j_1, j_2 \leq L$ such that $A_{i,j_1}, A_{i,j_2} \neq 0$,

$$\gamma^{-1} \leq \frac{A_{i,j_1}}{A_{i,j_2}} \leq \gamma.$$  

(5)

For instance, all nonzero entries of a $1$-variation-bounded matrix are the same for each row of the matrix.

The following theorem is the main result about approximation of sparse matrices by self-attention:

**Theorem 0.4.** Let $n > 1$, $k, d_{hid}, d \leq \min(d_{hid}, 2n)$ be natural numbers, $0 < \epsilon_1 < 1 < \epsilon_2 > 0$, $\gamma \geq 1$ be real numbers,

$$d = 2[16k^2 \left( \frac{\log \gamma - \log \epsilon_1}{\epsilon_2} + 1 \right)^2 \times (2 \log n + \log(n-1) + \log 2)].$$  

(6)

Then there exist $W_Q, W_K \in \mathbb{R}^{d_{hid} \times d}$, such that for any right stochastic, $k$-nonzero-bounded, $\gamma$-variation-bounded matrix $A \in \mathbb{R}^{n \times n}$, there is $X \in \mathbb{R}^{n \times d_{hid}}$ and $M = \text{SAM}(X; d, d_{hid}, W_Q, W_K)$ satisfying

1. For all row indices $1 \leq i \leq n$ and column indices $1 \leq j_1, j_2 \leq n$ such that $A_{i,j_1} = 0, A_{i,j_2} \neq 0$, it holds that

$$\frac{M_{i,j_1}}{M_{i,j_2}} < \epsilon_1;$$  

(7)
2. For all row indices $1 \leq i \leq n$ and column indices $1 \leq j_1, j_2 \leq n$ such that $A_{i,j_1} \neq 0, A_{i,j_2} \neq 0$, it holds that
\[
\frac{A_{i,j_1}}{A_{i,j_2}} \cdot \exp(-\varepsilon_2) < \frac{M_{i,j_1}}{M_{i,j_2}} < \frac{A_{i,j_1}}{A_{i,j_2}} \cdot \exp(\varepsilon_2).
\]

$W_Q, W_K$ can be constructed in $O(d_{hid} \cdot d)$ time. For any $A$, $X$ and $M$ can be computed in randomized time polynomial in $n, d_{hid}, k$.

Informally, Theorem 0.4 states that for given sizes $d_{hid}, d$ growing only logarithmically with the sequence length $n$ when the sparsity and variability parameters $(k, \gamma)$ are fixed, there exist fixed parameter matrices $W_Q, W_K$ such that for any non-zero-bounded matrix $A$ there is a self-attention input $X$ such that $M = \text{SAM}(X; d, d_{hid}, W_Q, W_K)$ approximates $A$ very well. The quality of approximation is characterized by upper and lower bounds on ratios of elements located in the same row of $M$:

1. Equation (23) means that zero elements of $A$ are approximated by elements of $M$ which are small compared to nonzero elements of the same row when $\varepsilon_1$ is chosen small. By definition $M$ is a strictly positive matrix, therefore in principle we can only approximate zero elements of $A$ by very small positive numbers.

2. Equation (24) means that ratios of nonzero elements of the same row in $M$ are in a close multiplicative neighborhood of the corresponding ratios in $A$ when $\varepsilon_2$ is chosen small. Since rows of both matrices $A$ and $M$ sum up to 1, similar enough ratios of element pairs also imply element similarity in terms of their absolute magnitude.

Remark 0.5. In the statement of Theorem 0.4, $M$ weakly approximates $A$ for any $0 < \varepsilon_1 < 1, \varepsilon_2 > 0$ which can be checked by Definition 0.1 (Equation 23).

Finally, as the proof is constructive, we will obtain an algorithm for computing $W_Q, W_K$, which turn out to be matrices of a simple structure. For any $A$ from the theorem statement, the probabilistic algorithm induced by the proof enables $X$ and $M$ to be computed in randomized polynomial time in $n, d_{hid}, k$.

In the rest of the section we describe the detailed proof and intuition behind it.

Proof of Theorem 0.4: Matrix $B$ and the Intuition Behind the Proof

Define vector $\alpha \in \mathbb{R}^n$ so that for each row index $1 \leq i \leq n$, the minimal nonzero element in this row is $\alpha_i$. Define matrix $B \in \mathbb{R}^{n \times n}$ as follows. For all $1 \leq i, j, n$,
\[
B_{i,j} = \begin{cases} 
0 & \text{if } A_{i,j} = 0; \\
\log(A_{i,j}/(\varepsilon_1 \alpha_i)) + \varepsilon_2 & \text{otherwise.}
\end{cases}
\]

Observe, that $C = \varepsilon_1 \exp(-\varepsilon_2) \text{diag}(\alpha) \exp(B)$ can be thought of as an approximation of $A$:
\[
C = \varepsilon_1 \exp(-\varepsilon_2) \text{diag}(\alpha) \exp(B) \approx A.
\]

Indeed, for any $1 \leq i, j \leq n$ such that $A_{i,j} \neq 0, C_{i,j} = A_{i,j}$ by definition of $B$ and $C$ (Equations 9, 10). On the other hand, when $A_{i,j} = 0, C_{i,j} = \varepsilon_1 \exp(-\varepsilon_2) \alpha_i \leq \varepsilon_1$, where we use $\varepsilon_2 > 0$ and $\alpha_i \leq 1$, since $A$ is a right stochastic matrix. Hence, smaller $\varepsilon_1$ yields a better approximation (10).

Suppose we find $X, W_Q, W_K$ such that
\[
B \approx X W_Q W_K^T X^T.
\]

Intuitively, if the approximation (11) is sufficiently good then
\[
\text{USAM}(X; d, d_{hid}, W_Q, W_K) = \exp(X W_Q W_K^T X^T) \\
\approx \exp(B) = \varepsilon_1^{-1} \exp(\varepsilon_2) \text{diag}(\alpha)^{-1} C \\
\approx \varepsilon_1^{-1} \exp(\varepsilon_2) \text{diag}(\alpha)^{-1} A,
\]

where in the last transition we use (10). Since the unnormalized self-attention matrix $\text{USAM}(X; d, d_{hid}, W_Q, W_K)$ is a good approximation for $A$ with rescaled rows (recall $\varepsilon_1^{-1} \exp(\varepsilon_2) \text{diag}(\alpha)^{-1}$ multipliers), the normalized self-attention matrix $\text{SAM}(X; d, d_{hid}, W_Q, W_K)$ should be a good approximation for $A$, which is itself row-normalized (right stochastic):
\[
\text{SAM}(X; d, d_{hid}, W_Q, W_K) \approx A.
\]

Next, we formally construct such $X, W_Q, W_K$ and derive tight error bounds for the approximation (13) in terms of matrix element ratios (23, 24).

Proof of Theorem 0.4: Construction of $X, W_Q, W_K$ through Random Projections

Consider a singular value decomposition (SVD, Trefethen and Bau 1997) of the matrix $B = U \Sigma V^T$, where $U, V \in \mathbb{R}^{n \times n}$ are orthogonal matrices and $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_n)$, $\sigma_1 \geq \ldots \geq \sigma_L > 0$ are singular values of $B$. Define $D = U \Sigma$, then $B$ can be decomposed as $B = D V^T$.

We will use random projections to compress $D$ and $V$ into matrices of shape $n \times d/2$ ($d$ is even according to (22)). Namely, let $Y \in \mathbb{R}^{n \times d/2}$ be a random matrix sampled from a uniform distribution (Haar measure) on a set of Stiefel matrices $\Omega \in \mathbb{R}^{n \times d/2}|\Omega^\top \Omega = I_{d/2}|$. Here, $I_{d/2}$ is a $(d/2) \times (d/2)$ identity matrix. Then we set $X^{(1)} = (2n/d)^{1/2} D Y \in \mathbb{R}^{n \times d/2}$, $X^{(2)} = (2n/d)^{1/2} V Y \in \mathbb{R}^{n \times d/2}$, $X^{(1)}$, $X^{(2)}$ can be considered as compressions of $D, V$ since $X^{(1)} X^{(2)^\top}$ is an unbiased approximation of $B = D V^T$:
\[
\mathbb{E} X^{(1)} X^{(2)^\top} = D \times \mathbb{E} [(2n/d) \cdot YY^\top] \times V^T \\
= D \times \mathbb{E} [n \cdot Y_i Y_i^\top] \times V^T = D V^T = B,
\]

where we use the fact that columns of $Y$ are uniformly distributed on $S^{d-1}$. See Figure 2a for an illustration. We set $X, W_Q, W_K$ as
\[
X = \begin{bmatrix} X^{(1)} & X^{(2)} \end{bmatrix}, \\
W_Q = [I_d \ 0_{d \times (d_{hid} - d)}],
\]

\[
W_K = [I_d \ 0_{d \times (d_{hid} - d)}]^T,
\]

While $Y$ can be defined as a matrix with i.i.d. sub-Gaussian entries (Kaban 2015), in general, orthogonal projections outperform unstructured ones in theory and practice (Choromanski, Rowland, and Weller 2017; Choromanski et al. 2021; Lin et al. 2020). We also manage to obtain better dot product concentration results for Stiefel projections compared to unstructured ones (see discussion in Appendix).
By definition of $B$, the right-hand side inequality in (19) is rewritten as $\exp(\frac{1}{2}d/i\alpha i - \log \alpha_i - \log \epsilon_1 \leq \log \epsilon_1)$. Figure 2: Illustration of the approximation scheme. (a) $\exp(B)$ is a row-rescaled approximation of $A$ (Equation 14), whereas $X(1)^{\top}X(2)^{\top}$ is an unbiased approximation to $B$ (14). (b) Representation of $X(1)^{\top}X(2)$ as $XWQW^\top X^\top$ according to (15,16,17).

$W_K = \left[ \begin{array}{c c c} \mathbf{0}_{d/2 \times d/2} & I_{d/2} & \mathbf{0}_{d \times (d_{hid}-d)} \end{array} \right]^\top$, \hspace{1cm} (17)

where $\mathbf{0}_{m \times n}$ denotes a zero matrix of the corresponding shape. It is easy to see that in this case $XWQW^\top X^\top = X(1)^{\top}X(2)^{\top}$ (see Figure 2b).

Our next step is to prove that with a nonzero probability, differences of elements in $XWQW^\top X^\top$ concentrate near the same differences in $B$:

**Lemma 0.6.** With probability greater than $(n+2)^{-1}$ it holds that

$$\forall 1 \leq i, j, j_2 \leq n, i, j \neq j_2 : \begin{cases} (XWQW^\top X^\top)_{i,j} - (XWQW^\top X^\top)_{i,j_2} < -\epsilon_2, & (XWQW^\top X^\top)_{i,j_2} - B_{i,j_2} + B_{i,j_2} + \epsilon_2, \end{cases} \hspace{1cm} (18)$$

The proof (in Appendix) uses a corollary of the seminal Johnson-Lindenstrauss lemma (Johnson and Lindenstrauss 1984) about inner product preservation under random projections (Kaban 2015). Two crucial observations are that

- $W_Q$ and $W_K$ do not depend on $A$ by construction (16,17);
- according to Lemma 0.6, $X$ satisfying (18) can be found with any probability by redrawing $Y \sim O(n)$ times.

Suppose that (15,16,17) hold. Then for any $1 \leq i, j, j_2 \leq n, i, j \neq j_2$:

$$B_{i,j_2} - B_{i,j_2} - \epsilon_2 < (XWQW^\top X^\top)_{i,j_2} - (XWQW^\top X^\top)_{i,j_2} < -\epsilon_2.$$ \hspace{1cm} (19)

By definition of $B$ (9), whenever $A_{i,j_2} = 0, A_{i,j_2} \neq 0$, the right hand side inequality in (19) is rewritten as

$$(XWQW^\top X^\top)_{i,j_2} - (XWQW^\top X^\top)_{i,j_2} < -\log A_{i,j_2} + \log \alpha_i + \log \epsilon_1 \leq \log \epsilon_1.$$ \hspace{1cm} (20)

Here we also used $A_{i,j_2} \geq \alpha_i$, (20) is equivalent to (23) after exponentiating, since exponents of $(XWQW^\top X^\top)_{i,j_2}$ and $(XWQW^\top X^\top)_{i,j_2}$ are $M_{i,j_2}$ and $M_{i,j_2}$ rescaled by the same factor.

Similarly to (20), whenever $A_{i,j_2}, A_{i,j_2} \neq 0$, by expanding $B$’s definition, (19) is rewritten as

$$\log A_{i,j_2} - \log A_{i,j_2} - \epsilon_2 < (XWQW^\top X^\top)_{i,j_1}$$

which is equivalent to (24) after exponentiating. This concludes the proof of Theorem 0.4. \hfill \Box

**The Case of Causal Self-attention**

Another popular type of self-attention mechanism is causal self-attention, when each position $i$ only attends to elements $j \leq i$. This modification is required for autoregressive language modelling (Radford et al. 2019; Brown et al. 2020) when each token is modelled as depending only on previous tokens in the sequence. We define the causal self-attention matrix CSAM and causal self-attention CSA as

$$\text{CSAM}(X; d, d_{hid}, W_Q, W_K) = \text{diag}(\mathcal{M}'_1)^{-1}\mathcal{M}'_1,$$

$$\text{CSA}(X; d, d_{hid}, W_Q, W_K, W_V) = \text{CSAM}(X; d, d_{hid}, W_Q, W_K) X W_V,$$

where $\text{tril}(\cdot)$ is the lower triangular part of the argument matrix, meaning that it zeroes out all elements strictly above the main diagonal.

A natural question is whether the analog of Theorem 0.4 holds for causal self-attention matrices. Since these matrices are lower triangular, we should only attempt to approximate lower-triangular right-stochastic matrices $A$. In fact, we obtain the following result (differences with Theorem 0.4 are highlighted in bold).

**Corollary 0.7.** Let $n > 1, k, d_{hid}, d \leq \min(d_{hid}, 2n)$ be natural numbers, $0 < \epsilon_1 < 1, \epsilon_2 > 0, \gamma \geq 1$ be real numbers,

$$\frac{1}{d} \leq 2\left[ 16k^2 - \frac{\log \gamma - \log \epsilon_1}{\epsilon_2} + 1 \right]^2 \times (2 \log n + \log (n-1) + \log 2).$$ \hspace{1cm} (22)

Then there exist $W_Q, W_K \in \mathbb{R}^{d_{hid} \times d}$, such that for any lower triangular, right stochastic, $k$-nonzero-bounded, $\gamma$-variation-bounded matrix $A \in \mathbb{R}^{n \times n}$, there is $X \in \mathbb{R}^{n \times d_{hid}}$ and $M = \text{CSAM}(X; d, d_{hid}, W_Q, W_K)$ satisfying

1. For all row indices $1 \leq i \leq n$ and column indices $1 \leq j_1, j_2 \leq i$ such that $A_{i,j_1} = 0, A_{i,j_2} \neq 0$, it holds

2. For all row indices $1 \leq i \leq n$ and column indices $1 \leq j_1, j_2 \leq i$, it holds
that
\[
\frac{M_{i,j_1}}{M_{i,j_2}} < \epsilon_1; \quad (23)
\]

2. For all row indices \(1 \leq i \leq n\) and column indices \(1 \leq j_1, j_2 \leq i\) such that \(A_{i,j_1} \neq 0, A_{i,j_2} \neq 0\), it holds that
\[
\frac{A_{i,j_1}}{A_{i,j_2}} \cdot \exp(-\epsilon_2) < \frac{M_{i,j_1}}{M_{i,j_2}} < \frac{A_{i,j_1}}{A_{i,j_2}} \cdot \exp(\epsilon_2). \quad (24)
\]

\(W_Q, W_K\) can be constructed in \(O(d_{hid} \cdot d)\) time. For any \(A, X, M\) can be computed in randomized time polynomial in \(n, d_{hid}, k\).

**Proof.** The proof is unchanged compared to the proof of Theorem 0.4 with the only change that \(j_1, j_2\) are considered in the range \(1 \leq j_1, j_2 \leq i\) when computing difference bounds (20, 21). Cases when column indices \(j_1\) or \(j_2\) are bigger than the row index \(i\) are redundant, since both \(A\) and \(M = \text{CSAM}(X; d, d_{hid}, W_Q, W_K)\) have zero entries above the main diagonal.

We conclude that the same logarithmic dependence \(d = O(\log n)\) holds for causal self-attention.

**Experiments**

Theorem 0.4 suggests an upper bound (rhs. in Equation 22) for the \(d_{\min}(\epsilon_1, \epsilon_2)\) – i.e. the minimal \(d\) which results in \(M\) satisfying (23, 24) for fixed \(\epsilon_1, \epsilon_2\). A question which we address in the experimental section is, therefore, “What is the actual \(d_{\min}(\epsilon_1, \epsilon_2)\) in practice? Does it satisfy the logarithmic law \(d = O(\log n)\)?”

To answer this question, we perform the following simulation. We select a range of \((k, \gamma, \epsilon_1, \epsilon_2)\) parameters. For each set of parameters, we iterate over \(n\) on a uniform grid from 512 to 3072 with a step size 256. For each \(n\) we sample the matrix \(A\) and iterate over a uniform grid of \(d\) values in ascending order until we find such \(d\) which results in \(M\) satisfying (23, 24). We sample nonzero positions of \(A\) by taking a union over \(k\) random permutation matrices. The nonzero value is set to either 1 or \(\gamma\) by a coin flip.

To check whether for the current \(d\) there is \(M\) satisfying (23, 24), we construct \(Y, X(1), X(2)\) and \(M\) using the algorithm implied by the proof of Theorem 0.4. To sample Stiefel matrices \(Y\), we use the algorithm based on QR decompositions of random Gaussian matrices from (Stewart 1980). We redraw the \(Y\) matrix \(Qn\) times, \(Q = 1\), in the spirit of Lemma 0.6 suggesting that \(O(n)\) redraws should be enough to find the right \(Y, X(1), X(2)\) with a constant probability (when \(d\) is big enough).

Figure 3 illustrates the results. A remarkable observation is that, although empirical \(d_{\min}(\epsilon_1, \epsilon_2)\) (red circles) grows slower than the theoretical upper bound (blue dashed line shows the angle of this line), it nicely fits the logarithmic curve \(d = O(\log n)\) (black dotted line) in all two evaluated setups. The fact that the true \(d_{\min}(\epsilon_1, \epsilon_2)\) grows slower than (22) is natural, since (22) is an upper bound on it. Though, as experiments reveal, both curves differ only by a constant multiplicative factor.

We run an additional experiment to reveal how \(d_{\min}(\epsilon_1, \epsilon_2)\) depends on the number of samples \(Qn\) used to find \(M\) satisfying (23, 24). We take 2 setups and try a range of \(Q\) values from 0.1 to 5.0. Results are shown in Figure 4. We observe that \(d_{\min}(\epsilon_1, \epsilon_2)\) does not depend a lot on the choice of \(Q\) and is roughly unchanged. Therefore, we conclude that our findings regarding the behaviour of empirical \(d_{\min}(\epsilon_1, \epsilon_2)\) do not depend on \(Q\) much and \(Q = 1\) is a reasonable choice. Additional experimental details and results (plots for bigger \(k\) values) can be found in Appendix.

**Related Work**

**Expressivity of Transformers.** As Transformers gained popularity, more theoretical results have emerged to explain their expressivity. Transformers were shown to be universal approximators (Yun et al. 2020a), Turing-complete (Bhattamishra, Patel, and Goyal 2020) and able to recognize counter languages (Bhattamishra, Ahuja, and Goyal 2020). Furthermore, Transformer modifications such as Big-Bird (Zaheer et al. 2020), Transformers with hard attention (Pérez, Marinković, and Barceló 2019) and sparse Transformers (Yun et al. 2020b) were shown to be universal approximators. Note that (Yun et al. 2020a; Bhattamishra, Patel, and Goyal 2020) rely on multilayer constructions, whereas we consider a single self-attention module, and (Zaheer et al. 2020; Pérez, Marinković, and Barceló 2019; Yun et al. 2020b) analyze unconventional forms of self-attention. Dong, Cordoñnier, and Loukas (2021) analyze limitations of a pure self-attention Transformer, i.e. without feedforward blocks and skip connections. Cordoñnier, Loukas, and Jaggi (2020) show that multi-head self-attention can provably model any image convolution layer. Bhujanapalli et al. (2020) show that for large \(d (d \geq n)\) and fixed inputs, there exist \(W_Q, W_K\) which approximate any positive right stochastic matrix via self-attention. In contrast, we analyze expressivity when \(d\) is very small (\(d = O(\log n)\)).

**Random projections and Johnson-Lindenstrauss lemma.** Our proof techniques rely on the seminal Johnson-Lindenstrauss tranformation (JLT, Johnson and Lindenstrauss 1984) used for dimensionality reduction (Federok et al. 2018). A random projection approach similar to ours was used by Frankl and Maehara (1987) to lower-bound graph sphericity – a characteristic which is NP-hard to compute in general. A related random features technique, relying on random projections, was originally introduced to improve efficiency of kernel SVMs (Rahimi and Recht 2008), but recently found application in speeding up long-sequence Transformers (Choromanski et al. 2020, 2021). We use Stiefel matrices as random projections, which in general result in tighter approximations than unconstrained projections (Lin et al. 2020). Ensembles of orthogonal random projections were shown to provide much better concentration results for the estimators relying on them in various other contexts, in particular: kernel approximation (Choromanski and Sindhwani 2016; JLT can be considered a special instantiation with a dot-product kernel), estimation of the gradients of Gaussian smoothings with evolution strategy methods (Choromanski et al. 2018), kernel ridge
Figure 3: Finding empirical dependency of $d_{\text{min}}(\epsilon_1, \epsilon_2)$ on $n$ given the fixed set of parameters $(k, \gamma, \epsilon_1, \epsilon_2)$. Each plot corresponds to one out of six sets of tested parameters. Red circles correspond to simulation results: we redraw matrix $A$ 5 times for each $n$, resulting in 5 red circles per $n$. The horizontal $n$ axis is in a logarithmic scale, so that $O(\log n)$ corresponds to a straight line. The black dotted line corresponds to a $O(\log n)$ fit for the dots (linear when x axis scale is logarithmic). The blue dashed line indicates the slope of the theoretical upper bound on $d_{\text{min}}(\epsilon_1, \epsilon_2)$ (right hand side in Equation 22). We experiment with $k$ from $\{1, 2\}$ (results for bigger values of $k$ can be found in Appendix). The slope of the observations is lower than for the theoretical blue line, confirming our theoretical result.

Figure 4: Empirical dependence of $d_{\text{min}}(\epsilon_1, \epsilon_2)$ on $Q$ – the factor defining the number of samples $Q \cdot n$ (rounded to an integer) used to find the right matrix $M$. For each $Q$ we repeat the procedure to generate empirical values of $d_{\text{min}}(\epsilon_1, \epsilon_2)$ (red circles from Figure 3) and connect them into a line for the better visualization.

regression (Choromanski, Downey, and Boots 2018), sliced Wasserstein distance (Rowland et al. 2019) and more.

**Conclusion**

We have analyzed the expressive flexibility of the self-attention matrix as a mechanism to approximate sparse patterns. First, we show that even deciding whether the approximation is possible or not for a given pattern is NP-hard in general. Then we prove that weights of self-attention can be constructed in such a way that any sparse matrix can be approximated with certain input to the self-attention module. We show that, when error and other parameters are fixed, $d$ grows only logarithmically with the sequence length $n$, i.e. $d = O(\log n)$ when other matrix parameters are fixed. We hope our work will facilitate further in-depth theoretical analysis of self-attention and Transformers to understand better their remarkable performance across a variety of tasks.
Acknowledgements
V. L. acknowledges support from the Cambridge Trust and DeepMind. V. L. was part-time employed by Google while a PhD student. A.W. acknowledges support from a Turing AI Fellowship under EPSRC grant EP/V025279/1, The Alan Turing Institute, and the Leverhulme Trust via CFI.

References