Memorization in Attention-only Transformers

Léo Dana

June 14, 2024

Abstract

Summary

1 Introduction

Several papers have shown that memorization, especially of factual information, happens in the MLPs of a Transformer [ROME, MEMIT, Fact Finding].

In this picture, attention layers are considered to solely move information between each token's stream [Look Before You Leap].

But is this all that attention do regarding memorization ? Although it has not been shown that attention layers are memorizing empirically, one can wonder if they have the capacity to do so. And if so, what is the algorithm implemented by the attention ?

To answer these questions, we focus on an Attention-only Transformer (AoT) with only 1 layer and no MLP, so that only the attention memorizes.

We call an AoT with dimension d "fully expressive" if it can represent any linear map of rank d. When a Transformer is fully expressive, this means we only need to understand the best linear mappings.

Our main results are:

- A proof, that, given sufficiently many attention modules, the AoT can be fully expressive.
- Bounds on the KL divergence of the best linear map with π in two cases: when the distribution to learn has rank almost d, and when the distribution is a look-up table. For the former, memorizing means exaclty learning the d most important singular directions, for the latter it means memorizing every example but with some noise.

2 Related work

Associative memory: in there paper, they present a model of memory and show that the model can memorize at most d features. Yet, there approach of memory seems to be valid only in the regime where d is close to N. Thus we want to know what is the more general solution when d is much lower. Works have shown that models can have an exponential memory using superposition. To study these phenomenons, we generalize the papers' setting. This makes it hard to compare results. For example, on there task, our model doesn't exhibit a maximum storage capacity, which makes the comparison meaningless.

Memorization in the MLP -> several papers have shown that memorization in real transformers happend in MLPs (ROME / MEMIT / Factual recall), but what about memorization in Transformers ?

Attention Only Transformers -> the mathematical equivalence between Transformers and Attention only Transformers has been shown in X, Y. Yet there construction are very costly because one simulate an MLP using an Attention instead of using fully the attentional representation power. We will prove that MLP and Attention layers are equaly good when learning distribution.

Low-Rank Bottleneck in Multi-head Attention Models -> They prove that with $d < S$ the Transformer cannot produce arbitrary attention pattern and thus cannot become fully expressive. We provide a proof that if the total cumulated head dimension is large enough, then this restriction disappear.

On the Expressive Power of Self-Attention Matrices -> They show that it is sufficient to have $k^2 \log(S)$ head dimension to approximate attention pattern which are k sparce. Theorem X in appendix also approximate sparce attention pattern but uses S dimension to do so. Combining both ideas could provide a better approximation theorem !

3 Attention-only Transformers

Let π a distribution on sequences of N possible tokens, of lenght S. We generate each new token conditional on the last k

$$t_i \sim \pi(\cdot\,|t_{i-1:i-k})$$

We call $\pi(t_{i-1:i-k})$ the prior distribution and $\pi(\cdot\,|t_{i-1:i-k})$ the conditional distribution. We take the conditionnal distribution such that it doesn't have any zero probabilities. Most results will generalize when the conditional contains some zeros, we describe the changes in appendix ???.

We are interested in the problem of learning a function to approximate the distribution π for token of index $> k$. Our goal is to minimize the KL-divergence for sequences of lenght S, defined as

$$d_{KL}^{S}(\pi, f) = \mathbb{E}_{t\sim\pi}\left[\frac{1}{S-k}\sum_{i=k+1}^{S}\log\left(\frac{\pi(\cdot|t_{i-1:i-k})}{f(t_{1:i-1})}\right)\right]$$

3.1 Attention-only Transformer's Formalism

We consider a variant of the transformer architecture T with only 1 layer, 1 head per attention mechanism, but P attention layers computed in parallel containing one heads each. Although all results will be stated for one head per attention module, they hold for any number of heads. The embedding dimension is d. Let $t_{1:S} = \{t_i\}_{i=1:S}$ a sequence of S tokens. The transformer's computation writes as follows,

$$T(t_{1:S})_s = W_U\left(e(t_s) + \sum_{p=1}^{P} W_{OV}^{p} A^{p}(t_{1:S})\right)_s$$

where each Attention block can be written

$$A^p(t_{1:S})_s = \sum_{j=1}^{s} a_s^p(t_{1:S})_j(e(t_j) + pos_j)$$

and a is the softmaxed attention score

$$a_s^p(T_i(t_{1:S}))_j = \text{softmax}\left(\frac{1}{\sqrt{d}}(e(t_s) + pos_s)^T W_{QK}^p(e(t_i) + pos_i),\ i = 1:s\right)_j$$

pos_s are the additive positional embeddings, e is the token embedding matrix and W_U the unembedding matrix. We have factorized matrices W_Q^p and W_K^p as well as W_V^p and W_O^p, following the intuition from [Mathematical framework for Transformers].

By concatenating the matrices W_{OV}^p and A^p, we can factorize the attention asa sum of linear maps,

$$T(t_{1:S})_s = W_U(e(t_s) + W_{OV}A(t_{1:S})_s)$$

with $W_{OV} \in \mathbb{R}^{d,dP}$ and $A(t_{1:S}) \in \mathbb{R}^{dP}$.

We also introduce $f_W(t_{1:S}) = \text{softmax}(W\text{onehot}(t_{1:S}))$, the low rank linear map on token sequences. We can rewrite $T = f_{W_U W_E}$, were W_E depends on e, pos, W_{OV}^p and W_{QK}^p. We say that an AoT is *fully expressive* when its architecture can represent any low rank linear map.

3.2 Universal approximation

Since any AoT can be seen as a low rank linear map from the token sequences to logit, and so are Transformers in general, the class of function is at least as expressive as the class of linear map.

Proposition 1. *Let T be any transformer with embedding dimension d, and let W^* of rank d minimizing the KL-divergence with π. Then we have that for any $S \geq k+1$,*

$$d_{KL}^S(\pi, T) \geq d_{KL}^{k+1}(\pi, f_{W^*})$$

In fact, for the class of AoT, they are as expressive as the class of low rank linear maps. Moreover, we can control the approximation error when the AoT doesn't have enough heads to be fully expressive.

Theorem 1. *Let $\epsilon \geq 0$, N_ε the smallest number of questions whose cumulative probability is greater than $1 - \varepsilon$, and f_{W^*} the optimal linear mapping of rank d. There exist an AoT T^* with embedding dimension d and $\lceil \frac{N_\epsilon}{d} \rceil$ total parallel attention modules such that its divergence with π is bounded*

$$\left| d_{KL}^{k+1}(\pi, f_{W^*}) - d_{KL}^{k+1}(\pi, T^*) \right| \leq \epsilon \sigma_1(f_{W^*}) C(d, N, SN_\epsilon)$$

Here, taking $\varepsilon = 0$ gives a fully expressive AoT Transformer. In the bound, the constant $\sigma_1(f_{W^*})$ means the biggest prediction made by f_{W^*}, since in the worst case, by not controling the predictions of the low probability sequences, the AoT might predict a logit of very high norm.
This theorem holds if we try to predict several tokens one after the other, meaning that we bound $d_{KL}^S(\pi, T^*)$ at the cost of having a bigger residual stream.

[TODO] Learning with MLPs uses the same number of parameters, and has a lower bound. This means that attention is actually as expressive as MLPs.

4 Optimal Linear Mappings

Let $f : \mathbb{R}^N \to \mathbb{R}^M$ a linear function of rank d, S the softmax function, and $\pi \in \Delta^{N,M}$ a bigram distribution. We want to find the optimal mapping f that minimizes the KL-divergence $d_{KL}(\pi, S \circ f) = d_{KL}(\pi, f) = \mathbb{E}_{t_i, t_o}\left[\log\left(\frac{\pi(t_o|t_i)}{(S \circ f)(t_i)_{t_o}}\right)\right]$. We call it the *Bigram Problem*: given an inpout token in $[N]$ the task is to predict the correct probability distribution on output tokens in $[M]$.

Notations: t_i always represent the index of the input token and t_o of the output token. We denote π_{t_i} the conditional distribution $(\pi(t_o|t_i))_{t_o}$, $\Pi_{t_i,t_o} = \pi(t_o|t_i)$ the matrix whose columns are the conditional probabilities, and similarly $L_{t_i,t_o} = \log(\pi(t_o|t_i))$. We let $I_{\pi_{t_i}}$ be the diagonal matrix with value $\pi(t_o|t_i)$ and let $H(\pi)$ the entropy of π.

The solutions to this problem depend on the rank of $L - \mathbb{E}[L]$, which has coefficient $L_{t_i,t_o} - \mathbb{E}_{t_o}[L_{t_i,t_o}]$. In particular, if $d \geq rank(L - \mathbb{E}[L])$, then we can take $f = L - \mathbb{E}[L]$ which give 0 divergence. Otherwise, the problem is harder to solve because of the non-linearity of the softmax. To go one step further, we rewrite the divergence as

$$d_{KL}(\pi, f) = \mathbb{E}_{t_i}\left[\log\left(\mathbb{E}_{t_o}\left[e^{\bar{f}(t_i)_{t_o}}\right]\right)\right]$$

with $\bar{f}(t_i)_{t_o} = f(t_i)_{t_o} - L_{t_i,t_o} - \mathbb{E}_{t_o}[(t_i)_{t_o} - L_{t_i,t_o}]$. Note that this expression shows we can choose $w_E(t_i)$ independantly for each t_i by minimizing the sum of exponential. This will be the basis of our analysis. To find W_U, however, requires to find the matrice that take into account all the prior and the posterior probability π. In very specific case, the optimal W_U can be guessed.

The following sections propose progress on two settings: when $L - \mathbb{E}[L]$ has almost rank d, meaning that L's $d+1$ largest singular value is small, and when π has low entropy, meaning it is almost a one-to-one mapping.

4.1 Almost rank d

Let σ_j the j-th largest singular value of $L - \mathbb{E}[L]$. We treat here the case where σ_{d+1} is small. Intuitively, this means that solving the least-square problem will give an a good solution to the problem. We use this fact to obtain a bound on the divergence in the Lemma below.

Lemma 1. *Let*

$$\bar{f}(t_i)_{t_o} = f(t_i)_{t_o} - L_{t_i,t_o} - \mathbb{E}_{t_o}[(t_i)_{t_o} - L_{t_i,t_o}]$$

we have the following bound on the KL*-divergence*

$$\left|d_{KL}(\pi, f) - \mathbb{E}_{t_i}\left[\log\left(1 + \frac{1}{2}\mathbb{V}_{t_o}\left(\bar{f}(t_i)_{t_o}\right)\right)\right]\right| \leq \mathbb{E}_{t_i}[\exp_3(||\bar{f}(t_i)||_2)]$$

with $\exp_3(x) = e^x - 1 - x - \frac{x^2}{2}$.

Lemma 1 gives a bound valid when the L^2 norm of $\bar{f}(t_i)$ is small. Yet, in the above equation, $\exp_3(\mathbb{E}_{t_i}[||\bar{f}(t_i)||_2])$ is of order 3 while $\mathbb{V}_{t_o}\left(\bar{f}(t_i)_{t_o}\right)$ is of order 2. We would like to optimize for the highest order term in order to have the most precise bound. We thus propose to choose f by optimizing $\mathbb{V}_{t_o}\left(\bar{f}(t_i)_{t_o}\right)$ which is also a least-square problem.

Lemma 2. *For each t_i, and $f = W_U^T W_E$, the function*

$$w_E(t_i) \to \mathbb{V}_{t_o}(\bar{f}(t_i)_{t_o}) = ||I_{\sqrt{\pi_{t_i}}}(I_d - \mathbb{1}\Pi_{t_i}^T)(W_U w_E(t_i) - L_{t_i})||_2^2$$

is minimized by taking $\mathbb{1} \notin Im(W_U)$ and

$$w_E(t_i) = [W_U^T(I_{\pi_{t_i}} - \Pi_{t_i}\Pi_{t_i}^T)W_U]^{-1}W_U^T(I_{\pi_{t_i}} - \Pi_{t_i}\Pi_{t_i}^T)L_{t_i}$$

We end up with the function

$$f(t_i) = P_{t_i}L_{t_i}, \quad P_{t_i} = W_U[W_U^T(I_{\pi_{t_i}} - \Pi_{t_i}\Pi_{t_i}^T)W_U]^{-1}W_U^T(I_{\pi_{t_i}} - \Pi_{t_i}\Pi_{t_i}^T)$$

where P_{t_i} is a projection.

Now, using Lemma 2, we can derive a bound on the KL-divergence which is in some cases better.

Theorem 2. *For f_{ls} being the least-square solution of the problem $||f(t_i) - L_{t_i}||_2$, and $C = \left(1 + \frac{N||\pi_{t_i}||_{+\infty}}{\sqrt{2}}\right)$, we have the bound*

$$d_{KL}(\pi, f_{ls}) \leq \mathbb{E}_{t_i}\left[\log\left(1 + \frac{||\pi_{t_i}||_{+\infty}}{2}C\sigma_{d+1}^2\right) + \exp_3\left(\sqrt{C}\sigma_{d+1}\right)\right]$$

For f_{wls} being the weighted least-square solution of Lemma 2, we have the bound

$$d_{KL}(\pi, f_{wls}) \leq \mathbb{E}_{t_i}\left[\log\left(1 + \frac{||\pi_{t_i}||_{+\infty}}{2}\sigma_{d+1}^2\right) + \exp_3\left(\sqrt{\frac{||\pi_{t_i}||_{+\infty}}{||\pi_{t_i}||_{-\infty}}}\sigma_{d+1}\right)\right]$$

Since the choice of W_U is the same in both cases, we can choose the most optimal $w_E(t_i)$ between the two bounds, so taking the minimum of the bounds in the expectation over t_i still holds.

Remarks: The above theorem proposes two bounds, which are in general tradeoffs depending on the shape each distributions π_{t_i}. If we have $||\pi_{t_i}||_{-\infty} \simeq \frac{C}{N}$ the optimal choice if to take f_{wls}, and if $||\pi_{t_i}||_{+\infty} \simeq \frac{C}{N}$ then the optimal choice is f_{ls}. For langage modeling, and with huge dictionnary sizes, we always end up in the situation where $||\pi_{t_i}||_{-\infty} \simeq 0$, so in the absence of a titgher bound not involving $||\pi_{t_i}||_{-\infty}$, the weighted least-square solution should be expected.

To understand what the solution to the weigthed least-square do is quite tricky in the general case. Corrolary 1 presents a special case in which the difference between the two proposed solution is easily understandable. This case is also very nice since one doesn't have the term in $||\pi_{t_i}||_{-\infty}$.

Corrolary 1. *Let $L - \mathbb{E}[L] = V\Sigma U$ the singular value decomposition, and suppose that for ξ a permutation we have $V = I_\xi$ a permutation matrice. Let $S_d(\xi) = \{\xi(j), 1 \leq j \leq d\}$, $\pi(S|t_i) = \sum_{t_o \in S} \pi(t_o|t_i)$, and $H(\pi_{t_i}, S) = \frac{1}{\pi(S|t_i)} \sum_{t_o \in S} \pi(t_o|t_i) \log(\pi(t_o|t_i))$ Then we have*

$$f_{wls}(t_i)_{t_o} = \begin{cases} \log(\pi(t_o|t_i)) & \text{if } t_o \in S_d(\xi) \\ H(\pi_{t_i}, S_d(\xi)^c) + \log(\pi(S_d(\xi)^c|t_i)) & \text{if } t_o \in S_d(\xi)^c \end{cases}$$

and the divergence is exactly

$$d_{KL}(\pi, f_{wls}) = \mathbb{E}_{t_i}\left[\log\left(\pi(S_d(\xi)|t_i) + \pi(S_d(\xi)^c|t_i)(N-d)e^{H(\pi_{t_i}, S_d(\xi)^c)}\right)\right]$$

while

$$f_{ls}(t_i)_{t_o} = \begin{cases} \log(\pi(t_o|t_i)) & \text{if } t_o \in S_d(\xi) \\ H(\pi_{t_i}) & \text{if } t_o \in S_d(\xi)^c \end{cases}$$

and

$$d_{KL}(\pi, f_{ls}) = \mathbb{E}_{t_i}\left[\log\left(\pi(S_d(\xi)|t_i) + (N-d)e^{H(\pi_{t_i})}\right)\right]$$

Remarks: Corrolary 1 states that when the singular values are aligne with the output space, then f_{ls} and f_{wls} implement exact memorization of the d most important output coordinates. The only difference in that case is that f_{wls} predicts for the other tokens a uniform probability over the local mean $H(\pi_{t_i}, S_d(\xi)^c)$ of these tokens, whereas f_{ls} predicts the global mean $H(\pi_{t_i})$, failing to use adapt to the already learned features.
In this special case, the weighted least-square solution is strictly better than the least-square.

4.2 Look-up table

When the distribution has low-entropy, this means that their is one next-token to predict while the other are mostly noise. If the entropy of the conditional distributions are all exactly 0, then $\pi(t_o|t_i) = \mathbb{1}_{g(t_i)=t_o}$ for some function $g : [N] \to [M]$. In that setting, there exists a sequence of mappings which converges to a 0 divrgence. This means that despite having an embedding dimension of $d \geq 2$, we can remember an arbitrarely large mapping.

When the entropy of the conditional distribution is not 0, but is low, we have $\pi(t_o|t_i) \simeq \mathbb{1}_{g(t_i)=t_o}$. Here we can remember exaclty the most important prediction, and uniformly approximate the rest. The following theorem is a bound based on this intuition.

Theorem 3. *Let $g : [N] \to [M]$. There exists W_U be such that*

$$||W_U W_U^T - I_d||_\infty \leq C = \sqrt{\frac{32 \log(M+1)}{d}}$$

We choose $f(t_i) = \lambda(t_i) W_U^T w_U(g(t_i))$, with $\lambda(t_i)$ the solution to the equation

$$-H(\pi_{t_i}) = \log\left(\sum_j e^{\lambda(t_i)(w_U(j) - w_U(g(t_i)))^T w_U(g(t_i))}\right)$$

we obtain the bound

$$d_{KL}(\pi, f) \leq \mathbb{E}_{t_i}\left[(1 - \pi(g(t_i)|t_i)) \log\left(\frac{M-1}{e^{-H(\pi_{t_i})} - 1}\right)\left(\frac{1 + 2C + C d_{TV}(\tilde{\pi}_{t_i}, \pi_{unif})}{1 - 2C}\right)\right]$$

Here we make use of the Johnson-Linderstrauss theorem to make the vector from W_U interact as slightly as possible.

5 Conclusion

TODO

A Proof of Universal approximations

To prove Theorem 1, we will first show that the Attention matrix, with enough heads can be chosen invertible.

Lemma 3 (Rank of the attention matrix)**.** *Let L the size of the sequence, d the embedding dimension, and P the number of parallel attention heads. Let $A \in \mathbb{R}^{N^L, Pd}$ as defined in section 3.1. For $Pd \geq N^k$, there exists matrices W_{QK}^p and embeddings $e(t_j)$ and pos_j such that A is invertible.*

Proof : Lemma 4. Let us choose e and pos with the wollowing properties

1. They have strictly positive cefficients,
2. For any non-zero polynomial with coefficient in $\mathbb{Z}$ of degree d on $(N+L)d$ variables, then the polynomial is non zero on the $(N+L)d$ coefficients of $e(t_j)$ and pos_j,

Such embeddings exists using transcendental numbers. We note $x(t,j) = e(t_j) + pos_j$, and define

$$f : W \to \left(\frac{\sum_i^L e^{x(t,L)^T W x(t,i)} x(t,i)}{\sum_i^L e^{x(t,L)^T W x(t,i)}} \right)_{t \in \mathcal{T}_L}$$

To prove that A is invertible is equivalent to showing that its rows span $\mathbb{R}^{N^L}$, and thus to showing that the image of f is not contained in any hyperplan.

Let us prove this by absurd and take $(v_t)_{t \in \mathcal{T}_L}$ such that $\sum_{t \in \mathcal{T}_L} v_t^T f_{x(t,j)} = 0$ as functional equality. Let $i_{\max}(t) = \arg\max_i (x(t,i)_1)$ the indice of the greatest x on the first dimension. We can rewrite the equality as

$$\sum_{t \in \mathcal{T}_L} \frac{\sum_i^L e^{x(t,L)^T W (x(t,i) - x(t, i_{\max}(t)))} v_t^T x(t,i)}{\sum_i^L e^{x(t,L)^T W (x(t,i) - x(t, i_{\max}(t)))}} = 0$$

Let us only consider matrices W such that $W_{i,j} = 0$ if $i \neq 1 \neq j$. In this case, we have $x(t,L)^T W (x(t,i) - x(t, i_{\max}(t))) = x(t,L)_1^T W_{1,1} (x(t,i)_1 - x(t, i_{\max}(t))_1)$ so by definition of $i_{\max}(t)$, this is negative when $W_{1,1} = w$ is positive. Thus taking $W_{1,1} \to +\infty$ makes every exponential go to 0 or stay at 1. Thus, for a large enough w we can use the Taylor approximation of $x \to \frac{1}{1+x}$ around 0. and we have

$$\frac{1}{1 + \sum_{i \neq i_{\max}(t)}^L e^{x(t,L)^T W (x(t,i) - x(t, i_{\max}(t)))}} = 1 + \sum_{j=1}^{+\infty} \left(\sum_{i \neq i_{\max}(t)}^L e^{x(t,L)^T W (x(t,i) - x(t, i_{\max}))} \right)^j$$

Finally, we let $S(t) = \{i_1, ..., i_i \neq i_{\max}(t), \text{unordered}\}$ and call $N(j_1, ..., j_i)$ the number of possible permutations of this sequence. Now developping the product of sum into a sum of product gives us the following equality

$$\sum_{t\in\mathcal{T}_L}\sum_{j=0}^{+\infty}\sum_{i_1,...,i_j\in S(t),i_0}(-1)^j N(j_1,...,j_i)v_t^T x(t,i_0)e^{x(t,L)^T W\sum_{k=0}^{j} x(t,i_k)-x(t,i_{max}(t))}=0$$

Since the family of exponential function is free, we can identify the coefficient in front of each different exponent as 0.

To end the proof, we show that for each sequence t, there exists d coefficients in the exponential that are unique to this token sequence. Consider $n\in[2,d+1]$, and the exponent generated by the n^i times the indice $i\neq i_{\max}(t)$. This gives the exponent $x(t,L)_1 w\sum_{i\neq i_{\max}(t)} n^i(x(t,i)_1-x(t,i_{\max}(t))_1)$. Now let t' another token sequence and $c(t)_i$ other coefficient for the number of time that the indice i was chosen, such that

$$x(t,L)_1\sum_{i\neq i_{\max}(t)} n^i(x(t,i)_1-x(t,i_{\max}(t))_1)=x(t',L)_1\sum_{i\neq i_{\max}(t')} c(t')_i(x(t',i)_1-x(t',i_{\max}(t'))_1)$$

We can use the property (2) of the embedding to identify the coefficients on each side. First let us start with the positional embeddings. We only have each time the coefficients $pos_L^T pos_i$, so we can identify each. This means that $i_{\max}(t)=i_{\max}(t')$ and that $c(t')_i=n^i$. Finaly, to indentify the word embedding, we can start by looking at $e(t_L)_1*\left(\sum_{i\neq i_{\max}(t)}^{L} n^i(pos_{i,1}-pos_{i_{\max}(t),1})\right)$ and compare it to its counterpart, this gives us $t_L=t'_L$. Then we remove the extra information we already deduced and have

$$\sum_{i\neq i_{\max}(t)}^{L} n^i(e(t_i)_1-e(t_{i_{\max}(t)})_1)=\sum_{i\neq i_{\max}(t)}^{L} n^i(e(t'_j)_1-e(t'_{i_{\max}(t)})_1)$$

We can take any word embedding and look at its integer coefficient. If that number if negative, it means that this word embedding is also the maximum of the sequence, so it should be as well for the sequence of the right. If positive, it is a sum of power of n, so it can be written in base n uniquely. This unicity mean that the sequence on the right should produce the same, and thus we have $t_j=t'_j$. If each side is null, this means that every token is the same, but since we already deduced t_L we have all the information we need.

This shows that $t=t'$, that the number of indices is $j=\sum_{i\neq i_{\max}(t)}^{L} n^i$ of $j=\sum_{i\neq i_{\max}(t)}^{L} n^i+1$ if $i_0=i_{\max}(t)$. So we get the coefficient in front of the exponential

$$v_t^T\sum_{i=1}^{N} x(t,i)N_i=0$$

where N_i is the number of permutation of the sequence $(i_1,...,i_j)$ with $n^k-\mathbb{1}_{i=k\neq i_{\max}(t)}$ times the indice k. The matrice of this system has non-zero determinant by property (2), so the system solves for $v_t=0$. □

Proof : Theorem 1. Let us first prove the theorem when $\epsilon = 0$. Let N_0 be the number of sentences with non-zero prior probability. We take $P = \lceil \frac{N_0}{d} \rceil$. We have $T(t_{1:k}) = W_U e(t_k) + W_U W_{OV} A(t_{1:k})$, we choose e, pos and W_{QK}^p such that A is invertible like in Lemma 4. Now denote W_U^* and W_E^* the matrices that minimize $d_{KL}(\pi, f_{W_U^* W_E^*})$. We take $W_U = W_U^*$, and $W_{OV} = (W_E^* - e \otimes 1 \otimes ... \otimes 1)A^{-1}$, with $(e \otimes 1 \otimes ... \otimes 1)(t_{1:k}) = e(t_k)$, which gives $T^* = W_U^* W_E^*$ the fully expressive Transformer.

TODO □

A.1 Auto-regressive AoT

The setting of the paper focuses on one single next token prediction. However, one has Theorem 1 for several prediction in a row. The scaling in d needed is the sequence size, which is worst than [Expressivity of the Attention] to produce the a matrice that has the same focus. By using their results, one might trade the scaling $d \sim k^2 \log(S)$ for another error term.

Corrolary 2. *There exist a transformers T_λ^* with embedding dimension $d \geq S+1$, $\left\lceil \frac{N^k}{d} \right\rceil$ parallel attention module and 1 layer, such that*

$$d_{KL}^S(\pi, T_\lambda^*) + o(e^{-\lambda}) = \min_{r(W)=d} d_{KL}^{k+1}(\pi, f_W)$$

Proof : Corrolary 2. We keep the same notations as in the proof of Theorem 2, but we let $t_{j:s} = \{t_i\}_{i=j:s}$. We have that for all $k+1 \leq s \leq S$

$$T_\lambda(t_{1:s}) = W_U W_E(t_s) + W_U W_O \tilde{a}(t_{1:s})$$

and with the same reasoning, it is sufficient to have

$$W_E^*(t_{s-k:s}) = W_E(t_s) + W_O \tilde{a}(t_{1:s})$$

to have optimality of the prediction. Thus, if we can show that

$$\tilde{a}(t_{1:s}) = \tilde{a}(t_{s-k:s}) + o(e^{-\lambda})$$

this will conclude the proof, as the rest is the same as Theorem 2.

To prove this, we will focus on the attention pattern, and make sure the model only look at the k last tokens only. Let $(e_i)_{i=1:d}$ an orthonormal family of the embedding space, and we define $\mathcal{H}_1 = Vect(e_i, i = 1 : S-k)$, and $\mathcal{H}_2 = Vect(e_i, i = S-k+1 : d)$. Let $pos_s = e_s + \tilde{pos}_s$. We know that the matrices $W_{QK}^{p,h}$ are of dimension $d' = \frac{d}{H}$. We can thus fix

$$W_{QK}^{p,h} = \lambda \sum_{s=k+1}^{S} e_s \left(\sum_{i=s-k}^{s} e_i \right)^T + \tilde{W}_{QK}^{p,h}$$

where $\tilde{W}^{p,h}_{QK}$ is of rank $d' - S - k$, and of kernel included in $\mathcal{H}_2$. With this choice, we have that $pos_j^T W^{p,h}_{QK} pos_i = \lambda$ if both $k+1 \leq j \leq S$ and $j - k \leq i \leq j$, and 0 otherwise.

For $k+1 \leq s \leq S$, and $1 \leq i \leq s$, we have that

$$(W_E(t_s)+pos_s)^T W^{p,h}_{QK}(W_E(t_i)+pos_i) = \lambda\delta_{s-k\leq i\leq s}+(W_E(t_s)+\tilde{pos}_s)^T \tilde{W}^{p,h}_{QK}(W_E(t_i)+\tilde{pos}_i)$$

Intuitively, what this construction does is use S dimension to make the network focus on the right tokens at each step, and uses the rest of the dimenions to produce an invertible matrice $\tilde{a}$. Now, we can see that after the softmax, we get

$$a^{p,h}(t_{1:s})_j = a^{p,h}(t_{s-k:s})_j(1 + o(e^{-\lambda})) + o(e^{-\lambda})$$

□

B Proofs for Optimal Linear Mappings

Proof : Lemma 2. Recall that $\bar{f}(t_i)_{t_o} = f(t_i)_{t_o} - \mathbb{E}_{t_o}[f(t_i)]_{t_o} - (L_{t_i,t_o} - \mathbb{E}_{t_o}[L_{t_i,t_o}])$.

$$\begin{aligned}
d_{KL}(\pi, f) &= \mathbb{E}_{t_i}[H(\pi_{t_i})] - \mathbb{E}_{t_i,p}\left[\log\left(\frac{e^{f(t_i)_p}}{\sum_{t_o} e^{f(t_i)_{t_o}}}\right)\right] \\
&= \mathbb{E}_{t_i}[H(\pi_{t_i})] - \mathbb{E}_{t_i,p}[f(t_i)_p - \mathbb{E}_{t_o}[f(t_i)_{t_o} - L_{t_i,t_o}]] \\
&- \mathbb{E}_{t_i}\left[\log\left(\sum_{t_o} e^{f(t_i)_{t_o} - \mathbb{E}_{t_o}[f(t_i)_{t_o} - L_{t_i,t_o}]}\right)\right] \\
&= \mathbb{E}_{t_i}\left[\log\left(\sum_{t_o} \pi(t_o|t_i) e^{\bar{f}(t_i)_{t_o}}\right)\right] \\
&= \mathbb{E}_{t_i}[\log(1 + \mathbb{E}_{t_o}[e^{\bar{f}(t_i)_{t_o}} - 1|t_i])]
\end{aligned} \tag{1}$$

For the bound we use the finite increment inequality for $x \to \log(1+x)$

$$\left|\frac{\log(1+y) - \log(1+x)}{y - x}\right| \leq \frac{1}{1 + \min(y, x)} \leq 1$$

with $y = \mathbb{E}_{t_o}[e^{\bar{f}(t_i)_{t_o}} - 1]$ and $x = \frac{1}{2}\mathbb{V}_{t_o}\left(\bar{f}(t_i)_{t_o}\right)$ both positive. □

Proof : Lemma 3. We decompose the variance as

$$\begin{aligned}
\mathbb{V}_{t_o}(\bar{f}(t_i)_{t_o}) &= \mathbb{V}_{t_o}(f(t_i)_{t_o}) - 2Cov_{t_o}(L_{t_o}, f(t_i)_{t_o}) + \mathbb{V}_{t_o}(L_{t_o}) \\
&= w_E^T(t_i)\mathbb{V}_{t_o}(w_U(t_o))w_E(t_i) - 2Cov_{t_o}(L_{t_o}, w_U^T(t_o))w_E(t_i) + \mathbb{V}_{t_o}(L_{t_o})
\end{aligned} \tag{2}$$

by taking the gradient on $w_E(t_i)$, we get the equation

$$\mathbb{V}_{t_o}(w_U(t_o))w_E(t_i) = Cov_{t_o}(L_{t_o}, w_U(t_o))$$

So using that $\mathbb{V}_{t_o}(w_U(t_o)) = W_U(I_\pi - \Pi\Pi^T)W_U^T$ is symetric and invertible if $\mathbb{1} \notin Im(W_U^T)$, which is a reasonnable assumption to make: since the softmax function erase the direction $Vect(\mathbb{1})$, W_U^T has no reason to learn this in its span. We find

$$w_E(t_i) = \mathbb{V}_{t_o}(w_U(t_o))^{-1}Cov_{t_o}(L_{t_o}, w_U(t_o))$$

□

Proof : Theorem 2.

$$\begin{aligned}
d_{KL}(\pi, f) &= \mathbb{E}_{t_i}[H(\pi_{t_i})] - \mathbb{E}_{t_i,t_o}\left[\log\left(\frac{f_{t_o}(t_i)}{\sum_j e^{f_j(t_i)}}\right)\right] \\
&= \mathbb{E}_{t_i}[H(\pi_{t_i})] + \mathbb{E}_{t_i,t_o}\left[\log\left(\sum_j e^{f_j(t_i)-f_{t_o}(t_i)}\right)\right] \\
&= \mathbb{E}_{t_i}[H(\pi_{t_i})] + \mathbb{E}_{t_i}\left[\log\left(\sum_j e^{f_j(t_i)-f_{g(t_i)}(t_i)}\right)\right] + \mathbb{E}_{t_i,t_o}[f_{g(t_i)}(t_i) - f_{t_o}(t_i)] \\
&= \mathbb{E}_{t_i}[H(\pi_{t_i})] + \mathbb{E}_{t_i}\left[\log\left(\sum_j e^{\lambda(t_i)(w_U(j)-w_U(g(t_i)))^T w_U(g(t_i))}\right)\right] + \mathbb{E}_{t_i,t_o}[f_{g(t_i)}(t_i) - f_{t_o}(t_i)]
\end{aligned}$$

For each t_i, $H(\pi_{t_i}) \in [0, \log(M)]$ and

$$\lambda \to \log\left(\sum_j e^{\lambda(w_U(j)-w_U(g(t_i)))^T w_U(g(t_i))}\right)$$

is decreasing from $\log(M)$ to 0. Thus there exists a solution $\lambda(t_i)$ to equation (1). Moreover since $w_u(t_o)w_U(j) \leq C$ for $j \neq t_o$ and $||w_U(g(t_i))||_2^2 \geq 1 - C$, we obtain the bound

$$\lambda(t_i) \leq \frac{1}{1-2C}\log\left(\frac{M-1}{e^{H(\pi_{t_i})-1}}\right)$$

Taking $\lambda(t_i)$ to be this solution leaves us with

$$\begin{aligned}
d_{KL}(\pi, f) &= \mathbb{E}_{t_i,t_o}[f_{g(t_i)}(t_i) - f_{t_o}(t_i)] \\
&= \sum_{t_i} \pi(t_i) f_{g(t_i)}(t_i) - \sum_{t_i,t_o} \pi(t_i)\pi(t_o|t_i) f_{t_o}(t_i) \\
&= \sum_{t_i} \pi(t_i)(1 - \pi(g(t_i)|t_i)) f_{g(t_i)}(t_i) - \sum_{t_i, t_o \neq g(t_i)} \pi(t_i)\pi(t_o|t_i) f_{t_o}(t_i) \\
&= \sum_{t_i} \pi(t_i)(1 - \pi(g(t_i)|t_i)) \left(f_{g(t_i)}(t_i) + \frac{\sum_{t_o} f(t_i)_{t_o}}{M-1} \right) - \sum_{t_i,t_o} \pi(t_i) \left(\pi(t_o|t_i) - \frac{(1 - \pi(g(t_i)|t_i))}{M-1} \right) f_{t_o}(t \\
&\leq \sum_{t_i} \pi(t_i)(1 - \pi(g(t_i)|t_i))\lambda(t_i) \left(1 + 2C + C d_{TV}(\tilde{\pi}_{t_i}, \pi_{\text{unif}})\right) \\
&\leq \mathbb{E}_{t_i} \left[(1 - \pi(g(t_i)|t_i)) \log \left(\frac{M-1}{e^{H(\pi_{t_i})} - 1} \right) \left(\frac{1 + 2C + C d_{TV}(\tilde{\pi}_{t_i}, \pi_{\text{unif}})}{1 - 2C} \right) \right]
\end{aligned}$$

With $\tilde{\pi}_{t_i}(t_o) = \pi(t_o|t_i, t_o \neq g(t_i))$, and π_{unif} the uniform distribution over $M-1$ tokens. □