How Do Transformers Learn Topic Structure: Towards a Mechanistic Understanding

Yuchen Li [1] **Yuanzhi Li** [1 2] **Andrej Risteski** [1]

Abstract

While the successes of transformers across many domains are indisputable, accurate understanding of the learning mechanics is still largely lacking. Their capabilities have been probed on benchmarks which include a variety of structured and reasoning tasks—but mathematical understanding is lagging substantially behind. Recent lines of work have begun studying representational aspects of this question: that is, the size/depth/complexity of attention-based networks to perform certain tasks. However, there is no guarantee the learning dynamics will converge to the constructions proposed. In our paper, we provide fine-grained mechanistic understanding of how transformers learn "semantic structure", understood as capturing co-occurrence structure of words. Precisely, we show, through a combination of mathematical analysis and experiments on Wikipedia data and synthetic data modeled by Latent Dirichlet Allocation (LDA), that the embedding layer and the self-attention layer encode the topical structure. In the former case, this manifests as higher average inner product of embeddings between same-topic words. In the latter, it manifests as higher average pairwise attention between same-topic words. The mathematical results involve several assumptions to make the analysis tractable, which we verify on data, and might be of independent interest as well.

1. INTRODUCTION

The transformer architecture (Vaswani et al., 2017) is a critical building block of many leading approaches to natural language processing (Devlin et al., 2019; Brown et al., 2020), and other domains such as vision (Dosovitskiy et al., 2021) and protein structure prediction (Jumper et al., 2021). While the NLP community has produced a large body of work on probing and visualizing trained networks (Hewitt & Manning, 2019; Clark et al., 2019; Tenney et al., 2019; Kovaleva et al., 2019), we still have little formal understanding of the mechanisms by which transformers, trained with simple gradient-descent based algorithms, learn from their training data. The challenge is that the training dynamics are non-trivial, even for relatively simple structured data distributions, and even for simple (e.g. 1-layer) transformers.

In particular, we study *semantic structure*, as understood through the lens of *co-occurrences* of words, and their topical structure. Precisely, if we fit topics to a real-life corpus like Wikipedia using a *Latent Dirichlet Allocation* (LDA, Blei et al., 2003) model, we find a pretrained BERT model produces token embeddings that are more similar (in terms of inner product) if they belong to the same topic, and more different if they belong to different topics (see e.g. Figure 3).

Inspired by these observations, we study LDA-generated data as a sandbox to understand—both through experiments on such synthetic data, and theoretical results—the process by which the embeddings and attention learn the topical structure. We find that the above observations from Wikipedia data are even more pronounced on synthetic LDA data. Moreover, we mathematically prove why such structure arises by analyzing a simplified *two-stage training dynamics* for a single-layer transformer trained under the masked language modeling objective. We also verify the two-stage nature of training dynamics obtains for a wide variety of optimizers and hyperparameter settings. [1]

2. OVERVIEW OF RESULTS

We focus on understanding the optimization dynamics of transformers in a simple sandbox: a single-layer transformer trained on (synthetic) data following a topic model distribution—and validate that our results robustly transfer to real data (Wikipedia WikimediaFoundation, 2023). We show that topic structure can be encoded both in the embedding layer, and in the attention mechanism of the network. Moreover, even if one of these components is not trained (i.e. handicapped), the other can "compensate" for it.

[1]Carnegie Mellon University [2]Microsoft Research. Correspondence to: Yuchen Li <yuchenl4@cs.cmu.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

[1]Code is released at https://github.com/YuchenLi01/transformer_topic_model_LDA

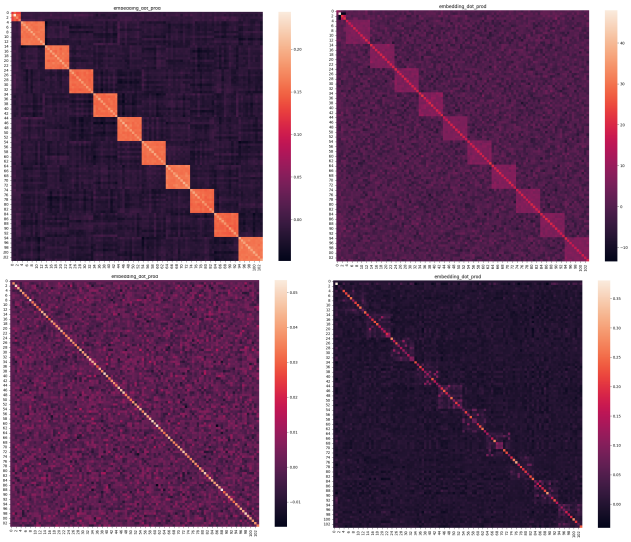

Figure 1. Embedding weight dot product of models trained on synthetic topic modeling data (Section 6.1). Top: with cross-entropy loss, Bottom: with squared loss. Left: using SGD optimizer, Right: using Adam optimizer, both with learning rate 0.01. The block-wise pattern verifies our theory in Section 4. The 10 blocks correspond to the 10 topics in the data distribution. In particular, a diagonal pattern is a special case of the block-wise optima that we prove (see Theorem 1).

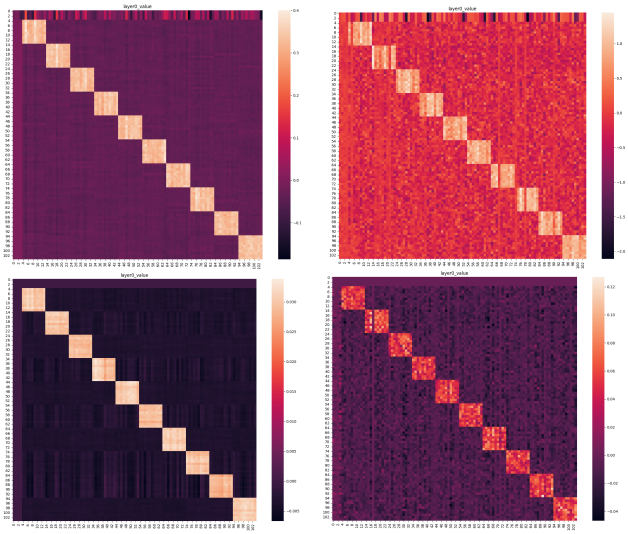

Figure 2. Convergence point of trained $\boldsymbol{W}^V$ (with L_2-regularization) when freezing uniform attention weights and one-hot word embedding. The four plots correspond to different combinations of loss function and optimizer. Top: with cross-entropy loss, Bottom: with squared loss. Left: using SGD optimizer, Right: using Adam optimizer, both with learning rate 0.01. The block-wise pattern verifies our theory in Section 5.2. The 10 blocks correspond to the 10 topics in the data distribution. Results are qualitatively similar without L_2-regularization, or if we train $\boldsymbol{W}^K$ and $\boldsymbol{W}^Q$ instead of freezing them (see Appendix G.1).

Theoretically, we characterize precisely how the topic structure is learned in the two extremal cases: when the attention mechanism is frozen to be uniform, and the only model parameters that are trained are the token embeddings; and when the token embeddings are frozen to be one-hot vectors, and the attention parameters (the key, query, and value matrices) are trained. We empirically verify our characterization on synthetic LDA-generated data, and also show that on real Wikipedia data, topic structure is learned both in the embeddings, and the attention mechanism.

2.1. Topic structure is encoded in token embeddings

In the first extremal case, we analyze the optima when we solely train the embedding layer. Precisely, we show that even when we freeze the attention scores to be uniform and all other elements of the transformer are set to identity, the model can still achieve near optimal loss by "encoding" the topic structure in the embedding weights:

Theorem (Optimal word embedding, informal)**.** *If the training data follows a topic model data distribution, and the transformer has trainable embedding layer, frozen (uniform) attention scores, and all other components set to identity, then the optimal embedding layer of a single layer transformer is such that the inner product of the embeddings of a pair of words is larger when the words belong to the same topic, and smaller when they belong to different topics.*

Intuitively, this result states that words of the same topic, after training, have more similar embeddings than words of different topics. In this sense, the embedding layer captures the topic structure. We also empirically show (Section 6 and Figure 1) that this phenomenon is robust to differences in loss function and optimization method. See Section 4 for the formal theorem and Appendix D for the proof.

2.2. Topic structure is encoded in self-attention

In the second extreme, we study the behavior of the self-attention in a transformer trained on a topic modeling distribution, without the aid of trained token embeddings — i.e. when we use hard-coded, *one-hot* embeddings. To make the analysis feasible, we break down the training process into two separate stages, and characterize the optima in each stage. In the *first stage*, the attention is frozen to be uniform, and the matrix $\boldsymbol{W}^V$ is trained. In the *second stage*, the matrix $\boldsymbol{W}^V$ is frozen to the optimal value from the first stage, and the optimal attention weights is analyzed. Intuitively, such a two-stage approximation is reasonable, because in the initial stages of training, the gradients for the value matrix are much larger than those for the key and query matrices (see Section 8). While this is an approximation, this two-stage phenomenon can be observed empirically for a wide variety of hyperparameter settings (see Section 5.1 and in particular Figure 4). We also provide empirical evidence that the optima characterized in our analysis closely track the actual convergence points of models.

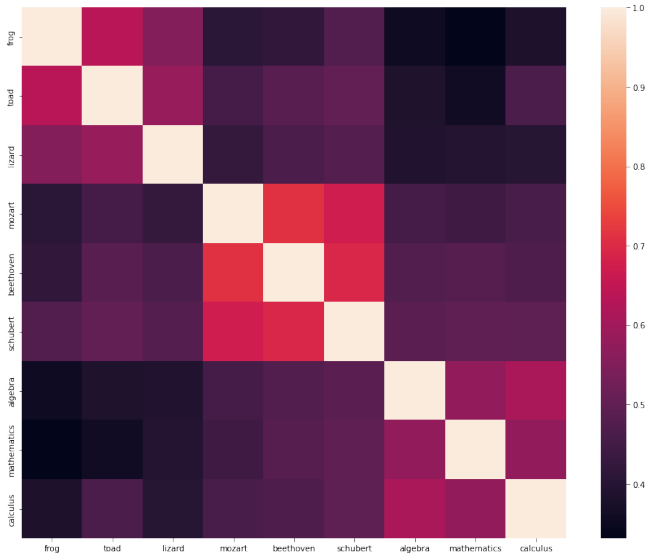

Figure 3. For a pre-trained BERT on Wikipedia corpus, the cosine similarity of the word embeddings encodes topical structures, i.e. it is larger if the two words belong to the same topic, and smaller if they belong to different topics. This phenomenon is more pronounced for words that are very likely only under a few topics. In this figure, the nine words fall into three topics: {frog, toad, lizard} are animals, {mozart, beethoven, schubert} are musicians, and {algebra, arithmetic, calculus} are mathematical concepts.

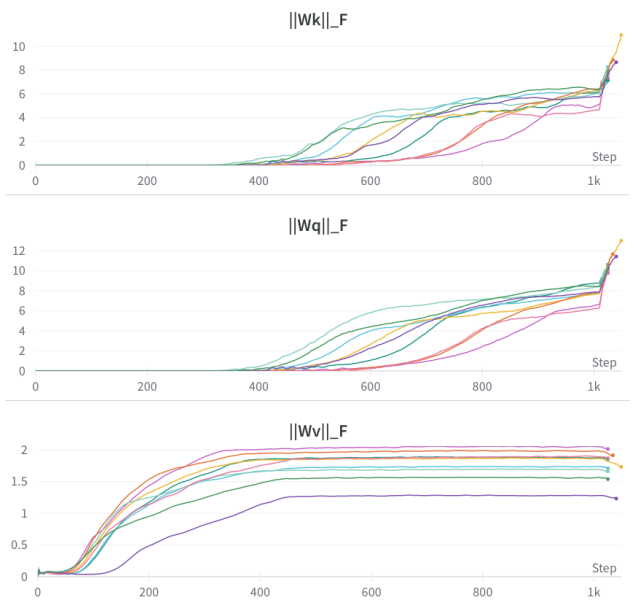

Figure 4. Two-stage learning dynamics of a single-layer transformer trained on LDA data distribution. All weight matrices are initialized to random matrices near zero, and *simultaneously trained*. The learning dynamics naturally exhibits a *two-stage* phenomenon: in **Stage 1** (steps 0-400), the norms of the key matrix (W^K, top) and the query matrix (W^Q, middle) stay close to 0, while the norm of the value matrix (W^V, bottom) increases significantly. In **Stage 2** (steps 400-1000), the norms of W^K and W^Q start increasing significantly, while the norm of W^V stays relatively flat. Different curves in the figure correspond to different settings of the hyperparameters as well as different runs in each setting. (See Section 8 for more details.)

In brief, the self-attention function is $\mathrm{Attn}(Z) := \boldsymbol{W}^V \boldsymbol{Z} A(\boldsymbol{Z})$ in which $A(\boldsymbol{Z})$ denotes the attention weights, and $\boldsymbol{W}^V$ is the value matrix weight. Intuitively, $A(Z)_{ij}$ is the importance of the i-th word for predicting the j-th word, and $\boldsymbol{W}^V$ is aggregates the word embeddings in a sentence, weighted by the attention weights $A(\boldsymbol{Z})$. The formal definition of the model architecture is in Section 3.3.

2.2.1. Optimal W^V in Stage 1

We characterize the optimal $\boldsymbol{W}^V$ in the initial stage of training: $\boldsymbol{W}^V$ will learn a block-wise structure (see Figure 2), in which each block corresponds to a topic:

Theorem (Optimal $\boldsymbol{W}^V$, informal)**.** *If the training data follows a topic model data distribution, when the token embeddings are frozen to be one-hot vectors and attention scores are frozen to be uniform, under mild L_2 regularization, the optimal $\boldsymbol{W}^V$ for the masked language modeling objective has block-wise structure, namely the (i, j)-th entry of $\boldsymbol{W}^V$ is on average larger when the words i and j belong to the same topic, and on average smaller when the words i and j belong to different topics.*

For the formal theorem statement, see Section 5. The proof is deferred to Appendix F. We also empirically show (Section 6 and Figure 2) that this phenomenon is robust to differences in training loss and optimization method.

2.2.2. Optimal attention weights in Stage 2

For the second stage of the training dynamics, we assume $\boldsymbol{W}^V$ is frozen to the optimal value in the first stage, and train the attention weights.

Theorem (Optimal attention weights, informal)**.** *If the training data follows a topic model data distribution, with a single layer transformer, the optimal attention weight for the masked language modeling objective after freezing block-wise $\boldsymbol{W}^V$ is such that on average: a convex combination of* same-word *attention and* same-topic-different-words *attention should be relatively large, compared to* different-topic *attention.*

For the formal assumption and theorem statements, see Section 5. The proof is deferred to Appendix F.

We empirically show (in Section 6) that even when the all the self-attention weight matrices are *jointly* trained (instead of trained with the two-stage process described), the behavior of attention weights still follows the relations that the above theorem describes.

2.3. Empirical results

We provide empirical evidence that the main conclusions in our theoretical findings remain robust even under settings that are more complex and realistic than our theoretical setup, and under variations of the training algorithm and loss. For example, we also test on synthetic data using a Latent Dirichlet Allocation (LDA) topic model (Blei et al., 2003) instead of our simplified topic modeling distribution; finally, we report results for a model pre-trained on the Wikipedia textual corpus, and discuss the connections with our conclusions derived in the synthetic setting. We describe detailed experimental setup and results in Section 6, as well as Appendix G.

3. PROBLEM SETUP

3.1. Topic models

For our theoretical analysis, in order to have a well-defined notion of a "ground truth", we will consider data distribution generated by a topic model consisting of T topics $\{1, \cdots, T\}$ and Tv words $\{1, \cdots, Tv\}$. We will in fact, consider a special case of an LDA (Latent Dirichlet Allocation) model (Blei et al., 2003). Precisely, each document $\boldsymbol{w}$ is a sequence of words $w_1, \cdots, w_N$, and is generated by: [2]

1. Randomly choose τ distinct topics $t_1, \cdots, t_\tau$ from $[T]$.
2. For $n \in [N]$:

(a) Randomly choose a topic t from $\{t_1, \cdots, t_\tau\}$.
(b) Randomly choose w_n from $\{(t-1)v+1, \cdots, tv\}$.

Note, under this data distribution, each word belongs to exactly one topic, and different topics do not share common words.

Definition 1 (Topic-word indicator)**.** *A word i belongs to topic t (denoted as $i \in t$) if $i \in \{(t-1)v+1, \cdots, tv\}$. Correspondingly, $\texttt{topic}(i) := \lceil \frac{i}{v} \rceil$*

Let $\mathcal{D}_{\boldsymbol{w}}$ denote the distribution of documents following the above generative process. Furthermore, for each document $\boldsymbol{w}$, let $\boldsymbol{X} \in \{0,1\}^{(Tv+1)\times N}$ denote its *one-hot* encoding, in which $X_{ij} = 1$ if $w_j = i$, and 0 otherwise. Analogous to $\mathcal{D}_{\boldsymbol{w}}$, let $\mathcal{D}_{\boldsymbol{X}}$ denote the distribution of document one-hot encodings.

To simplify our theoretical analysis, we consider the *infinitely-long-document* setting, such that within each document, the empirical token distribution is equal to the groundtruth token distribution:

[2]Our theoretical results crucially depend on all topics being disjoint, i.e. they do not share common words. It is not crucial that the words in the same topic all have the same probabilities. Allowing these probabilities to be different would lead to results of similar flavor, but complicates the notation.

Assumption 1 (Infinitely-long documents)**.** *Each document $\boldsymbol{w}$ consists of exactly τ topics $\{t_1, \cdots, t_\tau\}$. Moreover, for each word $i \in \{1, \cdots, Tv\}$ in the vocabulary, its empirical probability in the document*

$$\boldsymbol{p}_{\boldsymbol{w}}(i) = \frac{\sum_{n=1}^{N} \mathbb{1}_{w_n = i}}{N} = \begin{cases} \frac{1}{\tau v}, & \text{if } i \in \cup_{j=1}^{\tau} t_j \\ 0, & \text{otherwise} \end{cases}$$

In our synthetic data experiments, we use a finite N and generate data using an LDA model (Blei et al., 2003) which allows for slightly more variability—and demonstrates that our results are robust to changes in the setting. Detailed experimental setup is described in Section 6.

3.2. Training objective

Given data following the distribution defined in Section 3.1, we train a transformer network using the masked language modeling objective (Devlin et al., 2019). We first define the token $\texttt{[MASK]} = 0$ in addition to the words $\{1, \cdots, Tv\}$ of the topic model. Three constant probabilities $p_m, p_c, p_r \in (0,1)$ specify the masking scheme:

1. For the original document $\boldsymbol{w} = w_1 \cdots w_N$, first randomly choose a set of masked indices $M(\boldsymbol{w}) \subset [N]$ such that $\forall i \in [N]$, with probability p_m, $i \in M(\boldsymbol{w})$.
2. Define the masked document $\tilde{\boldsymbol{w}} = \tilde{w}_1 \cdots \tilde{w}_N$ such that for each $i \in [N]$,

   (a) If $i \notin M(\boldsymbol{w})$, then $\tilde{w}_i = w_i$.
   (b) If $i \in M(\boldsymbol{w})$, then $\tilde{w}_i =$

$$\begin{cases} w_i, \text{with probability } p_c \\ \text{random word in } [Tv], \text{with probability } p_r \\ \texttt{[MASK]} = 0, \text{with probability } 1 - p_c - p_r \end{cases}$$

Given a document $\boldsymbol{w}$ and its masked version $\tilde{\boldsymbol{w}}$, the model f_θ (parameterized by θ) observes $\tilde{\boldsymbol{w}}$ and is trained to predict the original words at the masked positions M. More formally, given the one-hot encoding of the masked document $\tilde{\boldsymbol{X}}$, and the model prediction $\hat{\boldsymbol{X}} = f_\theta(\tilde{\boldsymbol{X}}) \in \mathbb{R}^{(Tv+1)\times N}$, letting $\boldsymbol{X}_{:j}$ denote the j-th column of matrix $\boldsymbol{X}$, for some loss function $l(\cdot,\cdot) \to \mathbb{R}$, the training objective is $\min_\theta L(\theta)$ for

$$L(\theta) = \mathbb{E}_{\boldsymbol{X} \sim \mathcal{D}_{\boldsymbol{X}}} \mathbb{E}_M \frac{1}{|M|} \sum_{j \in M} l(f_\theta(\tilde{\boldsymbol{X}})_{:j}, \boldsymbol{X}_{:j}) \quad (1)$$

Motivated by the empirical success of applying weight decay to training transformers, we also consider a regularized version of the above masked language modeling objective. For L_2-regularization [3] with parameter $\lambda > 0$:

[3]When θ is a vector, L_2-regularization penalizes $\|\theta\|_2$. When θ is a matrix, the correct norm to regularize is $\|\theta\|_F$.

$$L_{\text{l2reg}}(\theta) = L(\theta) + \lambda\|\theta\|_2^2 \tag{2}$$

Our theoretical analysis uses the squared loss: given a prediction vector $\boldsymbol{x} \in \mathbb{R}^d$ and an one-hot label vector $\boldsymbol{y} \in \{0,1\}^d$ in which $y_i = 1$ and $\forall j \neq i, y_j = 0$

$$l := l_{\text{sq}}(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \tag{3}$$

Our experiments additionally study the cross entropy loss:

$$l := l_{\text{ce}}(\boldsymbol{x}, \boldsymbol{y}) = -\log \frac{\exp(\boldsymbol{x}_i)}{\sum_{j=1}^d \exp(\boldsymbol{x}_j)} \tag{4}$$

Remark 1. *We give results for both types of loss functions because the cross-entropy loss, albeit practically more commonly used, is theoretically less convenient. Concretely, it involves the softmax operation which is invariant under addition by the same constant in each dimension (implying that the minimizer is not necessarily unique); moreover, the optima are often at infinity. By contrast, with squared loss, the set of optima is more easily characterized using some finite-valued closed form expressions.*

Empirically, we will show (in Section 6) that the conclusions in our theoretical analyses hold for both the cross-entropy loss and the squared loss, as well as with variants of the training algorithm like SGD and Adam.

3.3. Transformer network architecture

To theoretically reason about the role played by the embedding layer and the self-attention layer, we consider a one-layer transformer model (Vaswani et al., 2017) with the simplification that the residual connection and normalization layers are removed. Precisely:

$$f(\boldsymbol{Z}) = \boldsymbol{W}^{\text{pred}}(\boldsymbol{W}^V\boldsymbol{Z})\sigma\left(\frac{(\boldsymbol{W}^K\boldsymbol{Z})^\top(\boldsymbol{W}^Q\boldsymbol{Z})}{\sqrt{d_a}}\right) + \boldsymbol{b}^{\text{pred}}$$

$\boldsymbol{Z} \in \mathbb{R}^{d\times N}$ is the input representation. d is the embedding dimension. $\boldsymbol{W}^{\text{pred}} \in \mathbb{R}^{V\times d}$ and $\boldsymbol{b}^{\text{pred}} \in \mathbb{R}^V$ are the prediction head weights and biases. V is the vocabulary size. In our masked language modeling setting (Section 3.2), $V = Tv + 1$. $\boldsymbol{W}^V \in \mathbb{R}^{d\times d}$ is the value matrix weight. $\sigma : \mathbb{R}^{N\times N} \mapsto (0,1)^{N\times N}$ is the column-wise softmax operation, such that $\sigma(A)_{ij} = \frac{\exp(A_{ij})}{\sum_{l=1}^N \exp(A_{lj})}$. d_a is the attention head size. $\boldsymbol{W}^K \in \mathbb{R}^{d_a\times d}$ is the key matrix. $\boldsymbol{W}^Q \in \mathbb{R}^{d_a\times d}$ is the query matrix. Let $A(\boldsymbol{Z})$ denote the attention weights:

$$A(\boldsymbol{Z}) := \sigma\left(\frac{(\boldsymbol{W}^K\boldsymbol{Z})^\top(\boldsymbol{W}^Q\boldsymbol{Z})}{\sqrt{d_a}}\right) \in (0,1)^{N\times N} \tag{5}$$

Appendix C includes additional remarks on the architecture.

In our setting, the input $\boldsymbol{Z}$ is the embedding of the masked document, i.e. $\boldsymbol{Z} = \boldsymbol{W}^E\tilde{\boldsymbol{X}}$ for some embedding weights $\boldsymbol{W}^E \in \mathbb{R}^{d\times(Tv+1)}$. Moreover, following empirical best practice (Press & Wolf, 2017) and standard implementation in (Wolf et al., 2020), we weight-tie the prediction head weight $\boldsymbol{W}^{\text{pred}}$ and the embedding weight $\boldsymbol{W}^E$:

$$f(\tilde{\boldsymbol{X}}) = \boldsymbol{W}^{E\top}\boldsymbol{W}^V\boldsymbol{W}^E\tilde{\boldsymbol{X}}A(\boldsymbol{W}^E\tilde{\boldsymbol{X}}) + \boldsymbol{b}^{\text{pred}} \tag{6}$$

In part of our theoretical analysis (in Section 5) and experiments (in Section 6), we freeze *one-hot* word embeddings, to study the mechanism that self-attention represents the topic structures without the aid of trained token embeddings. That is, set $d = Tv + 1$ and $\boldsymbol{W}^E = I$:

$$f(\tilde{\boldsymbol{X}}) = \boldsymbol{W}^V\tilde{\boldsymbol{X}}A(\tilde{\boldsymbol{X}}) + \boldsymbol{b}^{\text{pred}} \tag{7}$$

4. TOPIC STRUCTURE CAN BE ENCODED IN TOKEN EMBEDDINGS

The first result shows that, under the topic model data distribution, even if we freeze the self-attention to be uniform, the embedding layer can encode the topic structure. Precisely:

Theorem 1 (Optimal token embedding)**.** *Suppose the data distribution follows the topic modeling assumption in Section 3.1 and Assumption 1. Suppose we train a single layer transformer given by equation 6 with $\boldsymbol{W}^K = 0, \boldsymbol{W}^Q = 0, \boldsymbol{W}^V = I$ and $\forall i, \boldsymbol{b}_i^{pred} = -\frac{p_m p_r}{(1-(1-p_c)p_m)Tv}$, under the masked language modeling objective (equation 1) with the squared loss (equation 3). Then, there exist constants $u_0, \cdots, u_{Tv} \in \mathbb{R}$ such that the optimal word embedding weight $\boldsymbol{W}^E$ and $\boldsymbol{E} := \boldsymbol{W}^{E\top}\boldsymbol{W}^E$ satisfy:*

1. *The 0-th row of $\boldsymbol{E}$ satisfies:*
   - *(a)* $\boldsymbol{E}_{00} = -\left(\frac{1}{p_m(1-p_c-p_r)} - 1\right)\cdot u_0$
   - *(b)* $\forall t \in [T], \sum_{l\in t}\boldsymbol{E}_{0l} = u_0 v$
2. *The 0-th column of $\boldsymbol{E}$ satisfies $\forall i \in \{1, \cdots, Tv\}$:*
   - *(a)* $\boldsymbol{E}_{i0} = -\left(\frac{1}{(1-p_c-p_r)p_m} - 1\right)u_i$
3. *$\boldsymbol{E}_{ij}$ ($\forall i, j \in \{1, \cdots, Tv\}$) satisfy:*
   - *(a)* $\sum_{l\in \texttt{topic}(i)}\boldsymbol{E}_{il} = u_i v + \frac{1}{1-(1-p_c)p_m}$
   - *(b)* $\forall t \in [T]$ *such that* $\texttt{topic}(i) \neq t$, $\sum_{l\in t}\boldsymbol{E}_{il} = u_i v$

Remark 2. *Point 3 is the important one among the list of conclusions. The way to read the theorem is that, among the entries of an optimal $\boldsymbol{E}$: for i and j corresponding to the indices of tokens of the **same topic**, $\boldsymbol{E}_{ij}$ is (on average) larger, meaning that the embeddings of same-topic tokens are more similar; for i and j corresponding to **different topics**, $\boldsymbol{E}_{ij}$ is (on average) smaller, meaning that the embeddings of different-topic tokens are less similar. In particular,*

when the constants $u_0, \cdots, u_{Tv}$ are all zero, then the above larger-vs-smaller difference becomes a positive-vs-zero difference, which we roughly observe in practice.

Remark 3. *Intuitively, the setting of the bias $\boldsymbol{b}^{pred}$ is used to "denoise" the masked sequence, i.e. to subtract the probability caused by filling in random words in the masking process (described in Section 3.2).*

The proof of this theorem is deferred to Appendix D.

Proving comparable results under cross-entropy loss (equation 4) is more challenging considering Remark 1. However, we empirically show that, such blockwise pattern in $\boldsymbol{E} := \boldsymbol{W}^{E\top}\boldsymbol{W}^{E}$ tends to exist in a trained model under both the squared loss and the cross-entropy loss, and regardless of whether we (i) train all layers or (ii) only train the embedding layer while freezing all other layers. Moreover, the loss achieved in case (ii) is only slightly worse than in case (i). Finally, we also show (Figure 3) that on real data, words that are unambiguous (e.g. "calculus", "Mozart") exhibit a similar pattern as Theorem 1 states: same-topic words have more similar embeddings, and therefore larger embedding dot products, than different-topic words. Quantitatively, if we only restrict ourselves to words that are unambigious (i.e. likely to be emitted only under few topics), a similar phenomenon can be observed (see Table 5).

5. TOPIC STRUCTURE CAN BE ENCODED IN SELF-ATTENTION

Whereas the previous section showed that the token embedding layer can in principle perform the heavy-lifting in learning the topic-modeling distribution, we further show that self-attention *also* can encode the topic structures, when we disallow training the embedding layer. That is, we freeze the token embeddings to be one-hot.

5.1. The two-stage optimization process of self-attention

While inspecting the training dynamics of this one-layer transformer on the topic modeling data distribution, we observed a roughly *two-stage* process (illustrated by Figure 4): with certain initialization and learning rate settings, in **Stage 1**, the key matrix (W^K) and the query matrix (W^Q) stay close to 0, i.e. each position pays a near-uniform attention to all positions in the document, while the norm of the value matrix (W^V) increases significantly. In **Stage 2**, the norm of the the value matrix (W^V) already plateaus, and only after that, do the key and query matrices (W^K and W^Q) start to move.

Thus, while reasoning about the training process of transformers in our data distribution, we take motivation from the above empirical observation of such two-stage process, and consider a corresponding simplification: in Stage 1, the attention is frozen to be uniform, and only $\boldsymbol{W}^V$ is trained; in Stage 2, $\boldsymbol{W}^V$ is frozen, while $\boldsymbol{W}^K$ and $\boldsymbol{W}^Q$ are trained. This simplification is a reasonable proxy for standard training, and we furthermore validate that our theoretical characterizations are robust to standard training, both using SGD and Adam. We provide more discussion on the two-stage optimization process in Section 8.

5.2. Optimal W^V given uniform attention

The Stage 1 of optimization process is convex (but not strongly convex) in $\boldsymbol{W}^V$, and we show that the set of minima consist of exactly the set of $\boldsymbol{W}^V$ that exhibits a *blockwise* pattern:

Theorem 2 (Optimal $\boldsymbol{W}^V$ with mild L_2-regularization when freezing uniform attention)**.** *Suppose the data distribution follows the topic modeling assumption in Section 3.1 and Assumption 1. Suppose we train a single layer transformer given by equation 7 with $\boldsymbol{W}^K = 0, \boldsymbol{W}^Q = 0, \boldsymbol{b}^{pred} = 0$, under the L_2-regularized masked language modeling objective (equation 2) with the squared loss (equation 3). Then, $\lim_{\lambda\to 0} \operatorname{argmin} L_{l2reg}(\boldsymbol{W}^V) = \{\boldsymbol{W}^{V*}\}$ in which $\boldsymbol{W}^{V*} \in \mathbb{R}^{(Tv+1)\times(Tv+1)}$ satisfies:*

1. *The 0-th row of $\boldsymbol{W}^{V*}$:*
   - *(a)* $\forall j \in \{0, \cdots, Tv\}, \boldsymbol{W}^{V*}_{0j} = 0$
2. *The 0-th column of $\boldsymbol{W}^{V*}$:*
   - *(a)* $\forall i \in \{1, \cdots, Tv\}, \boldsymbol{W}^{V*}_{i0} = \frac{c_2c_3 - c_1Tv}{c_2^2 + Tv}$
3. $\boldsymbol{W}^{V*}_{ij}$ *($\forall i, j \in \{1, \cdots, Tv\}$):*
   - *(a)* $\forall l \notin \texttt{topic}(i),\ \boldsymbol{W}^{V*}_{il} = \boldsymbol{W}^{V*}_{\textit{diff-topic}} := -\frac{c_1c_2 + c_3}{c_2^2 + Tv}$
   - *(b)* $\forall l \in \texttt{topic}(i), \boldsymbol{W}^{V*}_{il} = \boldsymbol{W}^{V*}_{\textit{same-topic}} := \boldsymbol{W}^{V*}_{\textit{diff-topic}} + \frac{c_3}{v}$

in which the constants are:

- $c_1 = \frac{p_r}{(1-p_c-p_r)(1-(1-p_c)p_m)Tv} \in (0, 1)$
- $c_2 = \frac{1}{(1-p_c-p_r)p_m} - 1 \in (0, +\infty)$
- $c_3 = \frac{1}{1-(1-p_c)p_m} \in (1, +\infty)$

Empirically, the loss achieved by freezing $\boldsymbol{W}^K = \boldsymbol{W}^Q = 0$ and only training $\boldsymbol{W}^V$ is only slightly greater than the loss achieved by training all of them jointly, see Appendix G.

Intuitively, this block-wise $\boldsymbol{W}^V$ shows that, while inferring about the words at the masked positions: the model looks at unmasked positions in the document, each unmasked word only contributes to predicting words of the *same topic*, each unmasked word does not contribute to predicting words of *different topics*, and the model implicitly aggregates the topic distribution among the unmasked words, to infer the token distribution in the original document prior to masking.

The proof of this Theorem 2 is deferred to Appendix E. Proving a comparable result under the cross-entropy loss equation 4 is more challenging due to the same reasons outlined in Remark 1. However, empirically such block-wise $\boldsymbol{W}^V$ shows up for both the cross-entropy loss and the squared loss, as we show in Section 6.

5.3. Optimal attention weights

In our analysis on the stage 2 optimization process, we freeze the $\boldsymbol{W}^V$ to be some representative optima from stage 1 (Theorem 2), and characterize the optimal attention weights by comparing the following three types of attention weights: among the *same words* at different positions, among different words of the *same topic*, and among words of *different topics*.

We mainly consider the type of optimal $\boldsymbol{W}^V$ characterized in Theorem 2: $\boldsymbol{W}^V$ with uniform blocks (see Figure 2). Empirically, the model often approximately converges to these type of pattern (Section 6).

To formally reason about the behavior of average attention weights, we consider a simplified setting:

Assumption 2 (Attention pattern)**.** *Following the notation in equation 5, assume that for any masked document $\tilde{\boldsymbol{w}}$ with embedding $\tilde{\boldsymbol{X}}$,*

$$A(\tilde{\boldsymbol{X}})_{ij} = \begin{cases} c_1, \text{if } \tilde{w}_i = \tilde{w}_j \\ c_2, \text{if } \tilde{w}_i \neq \tilde{w}_j \text{ but } \texttt{topic}(\tilde{w}_i) = \texttt{topic}(\tilde{w}_j) \\ c_3, \text{if } \texttt{topic}(\tilde{w}_i) \neq \texttt{topic}(\tilde{w}_j) \end{cases}$$

in which $c_2 = \alpha c_3$ and $c_1 = \beta c_3$.

We note that this family of attention weights is *realizable*, and by symmetricity (among different topics and among the words in the same topic) and convexity (in $A(\tilde{\boldsymbol{X}})$), it is simple to prove that the attention pattern outlined in Assumption 2 is among the optimal attention patterns.

We will characterize the setting of α and β that minimizes the loss, under the following assumptions:

Assumption 3. *We consider these asymptotic settings:*

- *$T \to \infty$, i.e. the total number of topics grows to infinity.*
- ***(Sparse documents):** $\tau \to \infty, \tau = o(T)$, i.e. the number of topics in each document also grows to infinity, but much smaller than the total number of topics. (This is a common parameter regime: we typically think of each document as a sparse combination of topics.)*
- ***(No sparsely supported topics):** $v > (\frac{1}{1-(1-p_c)p_m}+1)^2 + 1$ (v is the number of tokens in each topic. $v \geq 10$ suffices under Assumption 4. This is also a common regime, where we assume no topic consists only of a small number of words.)*

Assumption 4. *In the training objective (Section 3.2), we consider the case $p_m < \frac{1}{2},\ p_c = p_r \in (0, \frac{1}{2})$.* [4]

Theorem 3 (Optimal attention weights)**.** *Suppose the data distribution follows the topic modeling assumption in Section 3.1 and Assumption 1. Suppose we train a single layer transformer given by equation 7 with $\boldsymbol{b}^{pred} = 0$ and $\boldsymbol{W}^V$ frozen to the optima in Theorem 2, under masked language modeling objective (equation 1) with the squared loss (equation 3), under Assumption 2, Assumption 3, and Assumption 4. Then, the optimal (α, β) satisfy*

$$\frac{v-1}{v}\alpha + \frac{1}{v}\beta \in (\lambda_1(\tau-1), \lambda_2 T)$$

in which $\lambda_1 := \frac{(1-(1-p_c)p_m+p_m p_r)(1+(1-p_c)p_m)}{2(1-(1-p_c)p_m)}$ and $\lambda_2 := 100(\frac{1-(1-p_c)p_m}{p_m p_r}+1)$.

Remark 4. *In particular, Theorem 3 implies that if we choose τ, T such that the lower bound exceeds 1, we expect the attention between same-topic words to be* on average *larger than that between different-topic words.*

Remark 5. *Note that when $\boldsymbol{W}^V$ is block-diagonal with uniform blocks, it is impossible to meaningfully bound α or β individually; instead, only their weighted average ($\frac{v-1}{v}\alpha + \frac{1}{v}\beta$) matters. In other words, different (α, β) will incur the same loss, as long as the above weighted average remains the same. Intuitively, this is because such block-diagonal $\boldsymbol{W}^V$ with uniform blocks* sums up *the attention on all words in each topic, and make predictions* solely based on the sums. *The proof of Theorem 3 is deferred to Appendix F.3.*

Remark 6. *When there is no L_2-regularization, the first-stage optima of $\boldsymbol{W}^V$ is not unique. We include additional analysis for representative cases of $\boldsymbol{W}^V$ in Appendix F.4.*

Remark 7. *When T, τ are finite, the loss expression turns out to be too complicated to characterize in closed form (because all the $o(1)$ terms need to be expanded). So we instead numerically compute the loss landscape as a function of α and β. See Appendix F.5.*

6. EXPERIMENTS

We analyze properties of the training dynamics via extensive experimental analysis. We will describe both the setup for synthetic (LDA-generated) data, and for Wikipedia data.

6.1. Results on synthetic LDA-generated data

Experimental setup In our experiments, we generate data following Section 3.1 with $T = 10, v = 10$, N uniformly randomly chosen from $[100, 150]$, except that Step 1 is changed to sampling the topic distribution according

[4] This setting is consistent with the masking scheme proposed in Devlin et al. (2019).

to the Dirichlet distribution (consistent with LDA, Blei et al., 2003) with $\alpha = 0.1$. Most sentences contain 2 to 4 topics. Our training objective follows Section 3.2 with $p_m = 0.15, p_c = 0.1, p_r = 0.1$ following Devlin et al. (2019). We use the model architecture following Section 3.3 but add back the bias terms $\boldsymbol{b}^K, \boldsymbol{b}^Q, \boldsymbol{b}^V$, following standard implementation in Wolf et al. (2020).

Trained token embeddings In Figure 1, we show that for a model in which all components are trained, the learned embedding weight $\boldsymbol{W}^E$ is such that ${\boldsymbol{W}^E}^\top \boldsymbol{W}^E$ displays a block-wise pattern. In particular, a diagonal pattern is a special case. These results show that our theory in Section 4 characterizes the optima of embedding layer which can be found by using either cross-entropy or squared losses, either SGD or Adam optimizers, and even when the other layers in the model are trained instead of frozen.

Learned value matrix W^V We show that when the word embeddings are *frozen to one-hot* and the attention weights are uniform (by setting $\boldsymbol{W}^K = 0, \boldsymbol{W}^Q = 0$), the trained $\boldsymbol{W}^V$ has a block-wise pattern, corresponding to the topical structure (see Figure 2).

We show (in Figure 10 in Appendix G.1) that even when the attention weights $\boldsymbol{W}^K, \boldsymbol{W}^Q$ are jointly trained with $\boldsymbol{W}^V$, the model would still approximately converge to the type of block-wise $\boldsymbol{W}^V$ described in our analyses in Section 5.2.

Convergence point of trained attention weights We show that, our conclusion in Theorem 3 holds not just when $\boldsymbol{W}^V$ is *frozen* to a block-wise pattern, but also when it is *trained* and naturally converges to such pattern. And we show (in Table 3 in Appendix G.2) that on average, each word pays more attention to words of *the same topic* than to words of *different topics*.

6.2. Results on natural language data

For a set of pre-trained transformer-based models (and their corresponding tokenizers) downloaded from Huggingface (Wolf et al., 2020), we compare the embedding similarity and attention weights between same-topic tokens and different-topic tokens. The topics are determined by fitting an LDA model with 100 topics on a sample of Wikipedia corpus (WikimediaFoundation, 2023) tokenized by the above tokenizers. We filter stop words. For each topic, we only keep a fraction of tokens that LDA assigns the highest likelihood in this topic. Consistent with our theoretical setting, we restrict to keeping only one topic for each word. In Table 1, we provide the results after such pre-processing. We provide additional details about the experimental setup and additional results (including when the last restriction of "one topic per word" is removed) in Appendix G.3.

7. RELATED WORKS

Recent works start to combine theoretical constructions and controlled experiments to justify the expressive power of transformers through the lens of Turing completeness (Bhattamishra et al., 2020b), function approximation (Yun et al., 2020), representing formal languages (Bhattamishra et al., 2020a; Ebrahimi et al., 2020; Yao et al., 2021; Liu et al., 2023), learning abstract algebraic operations (Zhang et al., 2022a), and statistical sample complexity (Wei et al., 2021; Edelman et al., 2022). Methodologically, we join a long line of works that characterize the capacity of neural network models by assessing their abilities in learning some simple models of the data (Siegelmann & Sontag, 1992; Gers & Schmidhuber, 2001; Weiss et al., 2018; Suzgun et al., 2019; Merrill, 2019; Hewitt et al., 2020; Li & Risteski, 2021; Yao et al., 2021; Zhang et al., 2022a; Liu et al., 2023). Our work extends this line of works, and in particular, our results indicate that there may be multiple reasonable *representational* optima, which calls for formally analyzing the training dynamics (Snell et al., 2021; Jelassi et al., 2022) to gain deeper understanding of what the model actually learns from such data distributions. Our result also formalizes some relevant intuitions given in Elhage et al. (2021), such as embedding layer capturing some bigram statistics. In topic modeling distribution, such "bigram statistics" translates to co-occurrence in a document. We include additional in-depth discussions of related works in Appendix B.

8. DISCUSSION: TWO-STAGE TRAINING

This two-stage optimization process (Section 5.1 and Figure 4) can be thought of as one iteration of the alternating optimization procedure. That is, we first train $\boldsymbol{W}^V$ while freezing $(\boldsymbol{W}^K, \boldsymbol{W}^Q)$, and then freeze $\boldsymbol{W}^V$ while training $(\boldsymbol{W}^K, \boldsymbol{W}^Q)$, and repeat this process.

In practice, $\boldsymbol{W}^K, \boldsymbol{W}^Q, \boldsymbol{W}^V$ in transformers are typically trained jointly instead of alternatingly. However, our empirical results show that, the conclusions drawn from the two-stage optimization analysis carry over even when they are trained jointly. Moreover, we don't find any qualitative aspects of normal training that are not captured by this two-stage approximation.

Intuitively, such two-stage phenomena occurs because if $\boldsymbol{W}^K, \boldsymbol{W}^Q, \boldsymbol{W}^V$ are initialized to random matrices near zero, and simultaneously trained, then in the initial steps, $\nabla_{\boldsymbol{W}^K} L$ contains the term $\boldsymbol{W}^Q$ (see equation 5), which is close to 0. By contrast, $\nabla_{\boldsymbol{W}^V} L$ contains the softmax-normalized attention weights $A(\tilde{\boldsymbol{X}})$ (see equation 7). Comparing these two, we shall see that $\nabla_{\boldsymbol{W}^V} L$ tends to be of larger in magnitude than $\nabla_{\boldsymbol{W}^K} L$, because each column of $\boldsymbol{W}^Q$ sums up to approximately 0, whereas each column of $A(\tilde{\boldsymbol{X}})$ sums up to exactly 1.

Model	Ambiguity Threshold	Avg embedding Cosine Similarity (Same-topic/Diff-topic)	Avg embedding Dot Product (Same-topic/Diff-topic)	Avg attn weight (Same-topic /Diff-topic)
Bert	0.0005	1.21	1.19	1.32
	0.001	1.13	1.15	1.28
	0.002	1.11	1.13	1.22
Albert	0.0005	5.64	6.29	1.33
	0.001	4.18	3.74	1.28
	0.002	3.24	2.93	1.22
Bart	0.0005	2.80	2.67	1.35
	0.001	1.95	1.92	1.31
	0.002	1.63	1.62	1.23
Electra	0.0005	5.98	5.37	2.14
	0.001	7.70	7.35	2.09
	0.002	7.46	8.08	1.95
Roberta	0.0005	6.44	6.81	1.40
	0.001	5.73	6.31	1.31
	0.002	5.24	5.30	1.22
Bert	0.0005	1.00080	1.00063	0.99943
(randomly	0.001	0.99974	1.00036	0.99996
initialized)	0.002	1.00016	1.00027	1.00007

Table 1. For models pretained on Wikipedia dataset, their token embeddings and attention weights encode topic structure. The different columns are: (1) The "ambiguity threshold", i.e. the number of words per topic, divided by the vocabulary size; **each word is only assigned one topic.** (2) The average embedding cosine similarity between different words of the *same topic*, divided by that between words of *different topics*. (3) The average embedding dot product between different words of the *same topic*, divided by that between words of *different topics*. (4) The average attention weight between different words of the *same topic*, divided by that between words of *different topics*. (The attention weights are normalized for debiasing, see discussion in Appendix G.3 for more details). Different rows represent different evaluation settings, controlled by "ambiguity threshold". Note that the avg same-topic embedding similarity and attention weight are consistently greater than the avg diff-topic counterparts, verifying our conclusions in Theorem 1 and Theorem 3.

Therefore, in the initial steps (i.e. Stage 1), $\boldsymbol{W}^V$ intuitively grows much faster than $\boldsymbol{W}^K$. For the same reason (note the symmetry between $\boldsymbol{W}^K$ and $\boldsymbol{W}^Q$, see equation 5), $\boldsymbol{W}^V$ intuitively grows much faster than $\boldsymbol{W}^Q$, too.

In Stage 2, it is less intuitively clear why $\|\boldsymbol{W}^V\|_F$ tends to plateau. Note that empirically, even when $\|\boldsymbol{W}^V\|_F$ plateaus, the $\boldsymbol{W}^V$ matrix itself still fluctuates with non-vanishing step-by-step changes. (That is, in each step, $\boldsymbol{W}^V$ "locally rotates" with an approximately constant norm.) Hence we refer to our Stage 2 analysis (which freezes $\boldsymbol{W}^V$ itself) as a simplification. However, the final empirical convergence point of $\boldsymbol{W}^V$ matches our theoretical analysis.

We show in Figure 5 (in Appendix A) that an approximate version of this multi-stage phenomenon can be observed on multi-layer transformers trained on Wikipedia as well.

Finally, this two-stage phenomenon is sensitive to hyperparameters like initialization and learning rate. The training process is not usually visibly two-stage using the common default hyperparameters. We leave it as an interesting future work to theoretically analyze the training dynamics when the two-stage phenomenon is not present. We include other discussions in Appendix A.

9. CONCLUSION

We initiated the study of understanding training dynamics of transformers in the presence of semantic structure captured by a topic model. Interesting directions of future work includes extending the analysis to data distributions that captures "syntactic" structure, e.g. through simple sandboxes like PCFGs. When both the model and the data distributions are complex, it remains a daunting challenge to "disentangle" how the many different aspects of the data (e.g. semantic and syntactic elements) are learned through the different parts of model architecture (e.g. attention, positional encodings, and embeddings).

Acknowledgements We thank Bingbin Liu, Yusha Liu, and Tanya Marwah for proofreading and providing constructive comments, Yewen Fan for helpful suggestions on empirically obtaining the two-stage optimization process, and Emmy Liu and Graham Neubig for insightful discussions on the connections with empirical observations.

Andrej Risteski and Yuchen Li acknowledge support by NSF awards IIS-2211907 and CCF-2238523. Andrej Risteski also acknowledges support by Amazon Research Award "Causal + Deep Out-of-Distribution Learning".

References

Arora, S., Ge, R., Koehler, F., Ma, T., and Moitra, A. Provable algorithms for inference in topic models. In Balcan, M. F. and Weinberger, K. Q. (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2859–2867, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/arorab16.html.

Awasthi, P. and Risteski, A. On some provably correct cases of variational inference for topic models. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/68a83eeb494a308fe5295da69428a507-Paper.pdf.

Belinkov, Y. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, March 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7.

Bhattamishra, S., Ahuja, K., and Goyal, N. On the Ability and Limitations of Transformers to Recognize Formal Languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7096–7116, Online, November 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.576. URL https://aclanthology.org/2020.emnlp-main.576.

Bhattamishra, S., Patel, A., and Goyal, N. On the computational power of transformers and its implications in sequence modeling. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pp. 455–475, Online, November 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.conll-1.37. URL https://aclanthology.org/2020.conll-1.37.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022, mar 2003. ISSN 1532-4435.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Clark, K., Khandelwal, U., Levy, O., and Manning, C. D. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 276–286, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4828. URL https://aclanthology.org/W19-4828.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Ebrahimi, J., Gelda, D., and Zhang, W. How can self-attention networks recognize Dyck-n languages? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4301–4306, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.384. URL https://aclanthology.org/2020.findings-emnlp.384.

Edelman, B. L., Goel, S., Kakade, S., and Zhang, C. Inductive biases and variable creation in self-attention mechanisms. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5793–5831. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/edelman22a.html.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt,

L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Gers, F. and Schmidhuber, J. Lstm recurrent networks learn simple context-free and context-sensitive languages. *IEEE transactions on neural networks*, 12 6:1333–40, 2001.

Hewitt, J. and Liang, P. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275.

Hewitt, J. and Manning, C. D. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419. URL https://www.aclweb.org/anthology/N19-1419.

Hewitt, J., Hahn, M., Ganguli, S., Liang, P., and Manning, C. D. RNNs can generate bounded hierarchical languages with optimal memory. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1978–2010, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.156. URL https://www.aclweb.org/anthology/2020.emnlp-main.156.

Htut, P. M., Phang, J., Bordia, S., and Bowman, S. R. Do attention heads in bert track syntactic dependencies? *ArXiv*, abs/1911.12246, 2019.

Jelassi, S., Sander, M. E., and Li, Y. Vision transformers provably learn spatial structure. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=eMW9AkXaREI.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Kovaleva, O., Romanov, A., Rogers, A., and Rumshisky, A. Revealing the dark secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4365–4374, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1445. URL https://aclanthology.org/D19-1445.

Li, X. and Gong, H. Robust optimization for multilingual translation with imbalanced data. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 25086–25099. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/d324a0cc02881779dcda44a675fdcaaa-Paper.pdf.

Li, Y. and Risteski, A. The limitations of limited context for constituency parsing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 2675–2687, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.208. URL https://aclanthology.org/2021.acl-long.208.

Liu, B., Hsu, D., Ravikumar, P. K., and Risteski, A. Masked prediction: A parameter identifiability view. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=Hbvlb4D1aFC.

Liu, B., Ash, J. T., Goel, S., Krishnamurthy, A., and Zhang, C. Transformers learn shortcuts to automata. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=De4FYqjFueZ.

Liu, L., Liu, X., Gao, J., Chen, W., and Han, J. Understanding the difficulty of training transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5747–5763, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.463. URL https://aclanthology.org/2020.emnlp-main.463.

Luo, Z., Weng, C., Wu, S., Zhou, M., and Ge, R. One objective for all models–self-supervised learning for topic models. *arXiv preprint arXiv:2203.03539*, 2022.

Meng, Y., Zhang, Y., Huang, J., Zhang, Y., and Han, J. Topic discovery via latent space clustering of pretrained language model representations. In *Proceedings of the ACM Web Conference 2022*, WWW '22, pp. 3143–3152, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450390965. doi: 10.1145/3485447.3512034. URL https://doi.org/10.1145/3485447.3512034.

Merrill, W. Sequential neural networks as automata. In *Proceedings of the Workshop on Deep Learning and Formal Languages: Building Bridges*, pp. 1–13, Florence, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3901. URL https://www.aclweb.org/anthology/W19-3901.

Nguyen, T. Q. and Salazar, J. Transformers without tears: Improving the normalization of self-attention. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong, November 2-3 2019. Association for Computational Linguistics. URL https://aclanthology.org/2019.iwslt-1.17.

Press, O. and Wolf, L. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 157–163, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-2025.

Sia, S., Dalmia, A., and Mielke, S. J. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1728–1736, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.135. URL https://aclanthology.org/2020.emnlp-main.135.

Siegelmann, H. T. and Sontag, E. D. On the computational power of neural nets. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pp. 440–449, New York, NY, USA, 1992. Association for Computing Machinery. ISBN 089791497X. doi: 10.1145/130385.130432. URL https://doi.org/10.1145/130385.130432.

Snell, C., Zhong, R., Klein, D., and Steinhardt, J. Approximating how single head attention learns, 2021.

Sontag, D. and Roy, D. Complexity of inference in latent dirichlet allocation. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper/2011/file/3871bd64012152bfb53fdf04b401193f-Paper.pdf.

Sun, K. and Marasović, A. Effective attention sheds light on interpretability. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 4126–4135, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.361. URL https://aclanthology.org/2021.findings-acl.361.

Sun, X. and Lu, W. Understanding attention for text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3418–3428, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.312. URL https://aclanthology.org/2020.acl-main.312.

Suzgun, M., Belinkov, Y., Shieber, S., and Gehrmann, S. LSTM networks can perform dynamic counting. In *Proceedings of the Workshop on Deep Learning and Formal Languages: Building Bridges*, pp. 44–54, Florence, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3905. URL https://www.aclweb.org/anthology/W19-3905.

Talebpour, M., García Seco de Herrera, A., and Jameel, S. Topics in contextualised attention embeddings. In Kamps, J., Goeuriot, L., Crestani, F., Maistro, M., Joho, H., Davis, B., Gurrin, C., Kruschwitz, U., and Caputo, A. (eds.), *Advances in Information Retrieval*, pp. 221–238, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-28238-6.

Tenney, I., Das, D., and Pavlick, E. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1452. URL https://aclanthology.org/P19-1452.

Thompson, L. and Mimno, D. Topic modeling with contextualized word representation clusters, 2020.

Tosh, C., Krishnamurthy, A., and Hsu, D. Contrastive estimation reveals topic posterior information to linear models. *J. Mach. Learn. Res.*, 22(1), jan 2021. ISSN 1532-4435.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural*

Information Processing Systems, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Vig, J. and Belinkov, Y. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 63–76, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4808. URL https://aclanthology.org/W19-4808.

Wei, C., Chen, Y., and Ma, T. Statistically meaningful approximation: a case study on approximating turing machines with transformers, 2021. URL https://arxiv.org/abs/2107.13163.

Weiss, G., Goldberg, Y., and Yahav, E. On the practical computational power of finite precision RNNs for language recognition. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 740–745, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2117. URL https://www.aclweb.org/anthology/P18-2117.

WikimediaFoundation. Wikimedia downloads. *Wikimedia Downloads*, 2023. URL https://dumps.wikimedia.org.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., and Liu, T.-Y. On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Yao, S., Peng, B., Papadimitriou, C., and Narasimhan, K. Self-attention networks can process bounded hierarchical languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3770–3785, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.292. URL https://aclanthology.org/2021.acl-long.292.

Yun, C., Bhojanapalli, S., Rawat, A. S., Reddi, S., and Kumar, S. Are transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ByxRM0Ntvr.

Zhang, J., Karimireddy, S. P., Veit, A., Kim, S., Reddi, S., Kumar, S., and Sra, S. Why are adaptive methods good for attention models? In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 15383–15393. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/b05b57f6add810d3b7490866d74c0053-Paper.pdf.

Zhang, Y., Backurs, A., Bubeck, S., Eldan, R., Gunasekar, S., and Wagner, T. Unveiling transformers with lego: a synthetic reasoning task, 2022a. URL https://arxiv.org/abs/2206.04301.

Zhang, Z., Fang, M., Chen, L., and Namazi Rad, M. R. Is neural topic modelling better than clustering? an empirical study on clustering with contextual embeddings for topics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3886–3893, Seattle, United States, July 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.285. URL https://aclanthology.org/2022.naacl-main.285.

A. ADDITIONAL DISCUSSIONS

A.1. The two-stage optimization process (continued)

We complement the discussion (in Section 8) on the intuitions and limitations of the two-stage simplification of the optimization process (Section 5.1 and Figure 4).

Here, we show in Figure 5 that an approximate version of this multi-stage phenomenon can be observed on multi-layer transformers trained on Wikipedia as well.

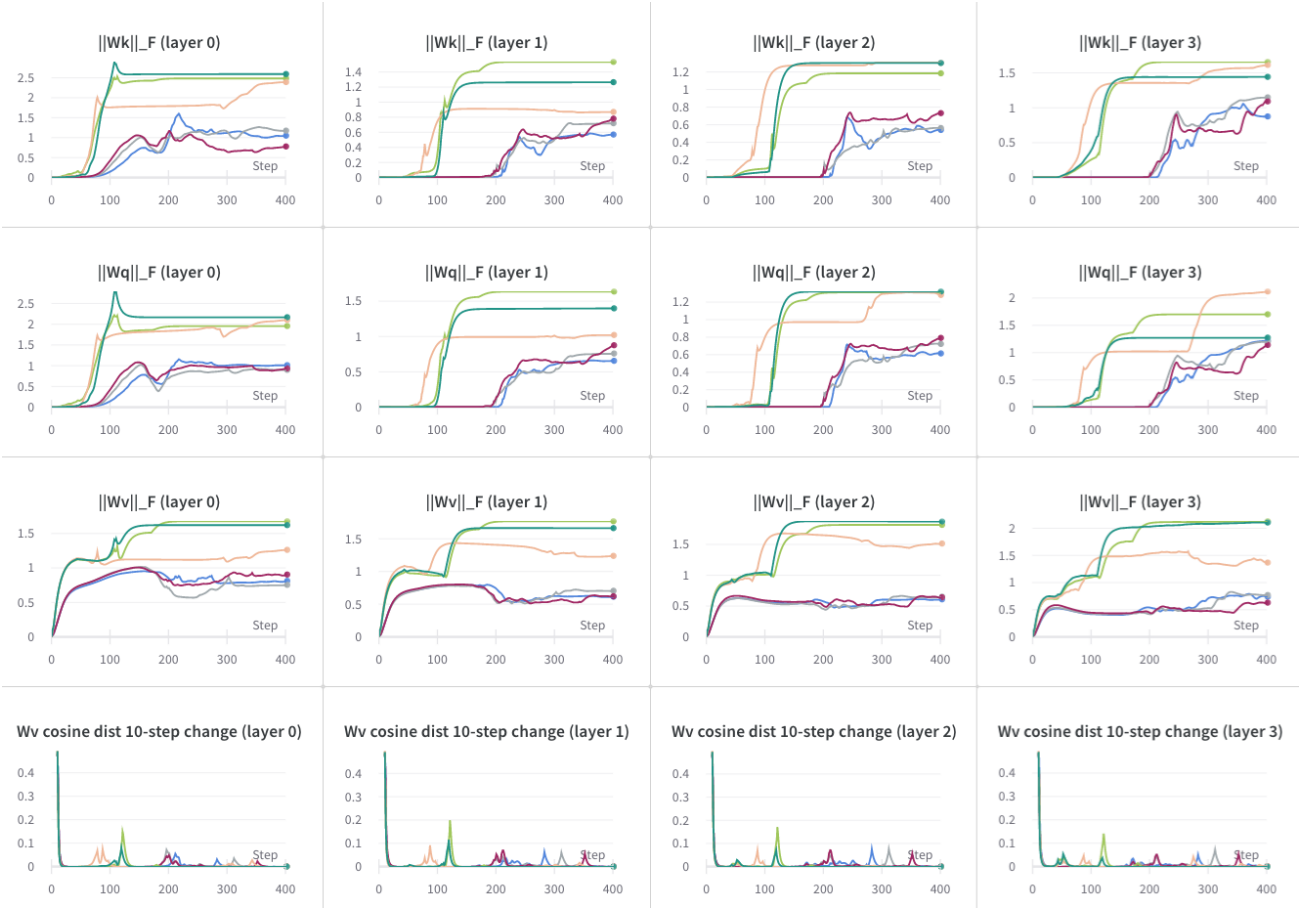

Figure 5. Two-stage learning dynamics of a 4-layer, 4-head-per-layer transformer trained on Wikipedia data. All weight matrices (key $\boldsymbol{W}^K$, query $\boldsymbol{W}^Q$, value $\boldsymbol{W}^V$ in each layer) are initialized to random matrices near zero, and *simultaneously trained*. Each column corresponds to one layer. The top 3 rows plot the trajectories of the Frobenius norms of $\boldsymbol{W}^K$, $\boldsymbol{W}^Q$, and $\boldsymbol{W}^V$ (weights from all heads in the same layer are concatenated together) after each gradient step. The bottom row measures the rotation of $\boldsymbol{W}^V$, i.e. the cosine distance between $\boldsymbol{W}^V$ in step t and $\boldsymbol{W}^V$ in step $(t-10)$. Cosine distance is defined as $\frac{1-cs}{2} \in [0,1]$, in which cs is the classic cosine similarity. The initial 400 steps of the learning dynamics naturally exhibit an *approximately two-stage* phenomenon: in **Stage 1** (roughly steps 0-100), for all 4 layers, the norms of $\boldsymbol{W}^K$ and $\boldsymbol{W}^Q$ stay close to 0, while the norm of $\boldsymbol{W}^V$ increases significantly and the orientation of $\boldsymbol{W}^V$ changes rapidly. In **Stage 2** (roughly steps 100-400), the norms of $\boldsymbol{W}^K$'s and $\boldsymbol{W}^Q$'s start increasing significantly, much later than $\boldsymbol{W}^V$ matrices do. Different curves in the figure correspond to different settings of the hyperparameters as well as different runs in each setting.

A.2. Do topic-wise behaviors perfectly correlate with co-occurrence counts?

Additionally, we note that fitting a topic model is closely related to word co-occurrence statistics, which raises the following question: should those empirical phenomenon (i.e. higher same-topic attention and more similar same-topic embeddings, shown in Table 5) be more fundamentally attributed to larger co-occurrence counts?

In the following, we also compare them with some preliminary empirical results on the behavior of embedding and attention, from both topic modeling and co-occurrence perspectives. Specifically, we compare the average attention weights and average embedding dot products, between same-topic word pairs and the N pairs of words that co-occur the most frequently in a sample of the Wikipedia corpus. The cutoff N is determined so that the number of "top co-occurring word pairs" is the same as the number of word pairs in each topic (controlled by the ambiguity threshold). The results are summarized in Table 2.

Based on those results, we conjecture that the topic-wise behavior of token embeddings and attention weights cannot be fully explained by simple co-occurrence counts.

Reasoning about their connections more formally would require analyzing some data distributions that better decouple these factors. We think that would be an interesting direction of future work.

# Word Pairs	Avg Attn Weight (Same-Topic)	Avg Attn Weight (Top Co-occur.)	Avg Embedding Cosine Similarity (Same-Topic)	Avg Embedding Cosine Similarity (Top Co-occur.)
105	0.00659	0.00751	0.468	0.316
435	0.00621	0.00695	0.461	0.311
1711	0.00597	0.00677	0.425	0.323

Table 2. For a BERT model pretained on Wikipedia dataset, the topic-wise behavior of its token embeddings and attention weights (shown in Table 1) cannot be fully explained by co-occurrence. The different columns are: (1) The number of pairs of tokens that have the highest co-occurrence counts (with stop tokens removed). The cutoffs are selected so that each row contains the same number of words pairs as one topic, corresponding to the rows in Table 1; (2) The average attention weights between same-topic words; (3) The average attention weights between tokens that co-occur the most; (4) The average embedding cosine similarity between different words of the *same topic*. (5) The average embedding cosine similarity between between tokens that co-occur the most. Note that for all "# word pairs" cutoffs considered, same-topic tokens have smaller average attention weight, but larger average embedding cosine similarity.

B. ADDITIONAL DISCUSSIONS ON RELATED WORKS

One line of prior works explain the success of transformers by empirically showing that the components (e.g. attention heads) of a trained model (e.g. BERT Devlin et al., 2019), contain abundant information for solving a wide range of "probing" tasks, across syntax and semantics (Hewitt & Manning, 2019; Clark et al., 2019; Tenney et al., 2019; Hewitt & Liang, 2019; Kovaleva et al., 2019; Belinkov, 2022), or through other approaches involving the attention weights (Vig & Belinkov, 2019; Htut et al., 2019; Sun & Marasović, 2021).

On the optimization side, Nguyen & Salazar (2019); Xiong et al. (2020); Liu et al. (2020); Zhang et al. (2020); Li & Gong (2021) propose algorithmic improvements (often with theoretical motivations) to help stabilize the training process of transformers. Towards explaining the training process of attention-based neural networks, Sun & Lu (2020) analyzes the trends of two quantities that are relevant to model performance and interpretability in text classification setting.

More relevant to our work, Snell et al. (2021) considers cross-attention in LSTM Seq2Seq models trained on machine-translation-inspired settings. Although the setting is different from ours (i.e. self-attention in transformers trained on topic modeling data distribution), we think some insights can transfer. In particular, it is worth noting that they proposed an intuitive simplifying assumption of the two-stage learning process of attention heads (but without theoretical or empirical validation). Our work uses a similar assumption [5] (Section 5.1). In our work, we validate our version of the two-stage assumption by providing a particular way to initialize the attention weight matrices, along with theoretical intuitions (Section 8) and empirical validation on synthetic data (Figure 4) as well as real data (Figure 5), showing that this two-stage process can be a reasonable approximation to the early steps of the real training dynamics of attention-based models under the settings that we analyze.

Recent work by Jelassi et al. (2022) theoretically shows how transformers learn the spatial structure of image-type datasets through gradient-descent-based optimization algorithms. In particular, their attention weights depend on the positional encodings only. Different from their work, our result (motivated by studying the semantics in language) focuses on topic modeling distribution that actually ignores the position information, so the attention weights only depend on the "bag of words" (i.e. the contents). In that sense, Jelassi et al. (2022) and our work complement each other, since real-world data distribution usually involves a combination of position-dependent and position-independent factors. An interesting future work would be studying how these factors interact during the training process.

Regarding the type of data distribution that we consider, we join a series of works that theoretically reason about the ability of learning under topic-modeling-based distributions (Sontag & Roy, 2011; Awasthi & Risteski, 2015; Arora et al., 2016; Tosh et al., 2021; Luo et al., 2022). In particular, Luo et al. (2022) shows that if a model can achieve low loss on contrastive or mask-prediction objectives, then it can recover topic posterior. However, these prior works do not theoretically analyze the optimization process of the transformer architecture. In fact, model architecture can indeed critically influence the resulting model obtained by masked-prediction-type tasks (see Liu et al. (2022) who highlight the subtlety of the interaction between the particular form of the task and the model specification). Hence, our analysis extends beyond the scope of these prior works by incorporating the theoretical analysis on the optimization process of the transformer architecture on topic modeling data distribution. Empirically, Sia et al. (2020); Thompson & Mimno (2020); Meng et al. (2022); Zhang et al. (2022b); Talebpour et al. (2023) analyze topic discovery via clustering the *contextualized* representations produced by pretrained language models. Different from these works, our theory and experiments on token embeddings focus on the convergence of embedding layer *parameters*.

[5] During our study, we independently re-proposed the two-stage training of attention heads, and later found that (Snell et al., 2021) already used this very similar assumption much earlier than us. Hence we added the comparison with the above prior work during an update of our paper. Moreover, while Snell et al. (2021) is the earliest paper we are aware of that explicitly assumes a two-stage training process specifically for attention heads, we note that similar approaches (more generally, alternating optimization) commonly appear in the optimization literature under a broader variety of settings.

C. ADDITIONAL INFORMATION ON THE SETUP

The positional encoding at the input is also removed, because the position information of a word in a document is irrelevant to the topic model defined in Section 3.1.

We also use a single-head attention.

C.1. Lemma on the optimal linear transform when freezing uniform attention

Under our setting, we first prove the following useful Lemma 1. Intuitively, it states that, when freezing uniform attention, the output of self-attention weights essentially counts the *unmasked* tokens in the document (as a result of the masking process described in Section 3.2). Given those counts, the best way to predict a token at the masked positions in the *original* document (i.e. prior to the masking process) is to:

1. First, aggregate the counts of the unmasked words within each topic, to infer the topic distribution in the observed document. In this, we further have the restriction that:
   - Each unmasked word only contributes to predicting words of the *same topic*
   - Each unmasked word does not contribute to predicting words of *different topics*
   - Never predict the mask token (`[MASK]`), because the original document does not contain any `[MASK]`
2. Second, we "denoise" the topic distribution, i.e. we subtract the probability caused by filling in random words in the masking process (described in Section 3.2).

In line with our single layer transformer architecture (Section 3.3, equation 7), we consider a special case in which the attention is *uniform*, i.e. $\forall i, j \in \{1, \cdots, N\}, A(\tilde{\boldsymbol{X}})_{ij} = \frac{1}{N}$, denoted by $A(\tilde{\boldsymbol{X}}) = \left[\frac{1}{N}\right]_{N\times N}$. (This can be achieved by setting $\boldsymbol{W}^K = 0, \boldsymbol{W}^Q = 0$.)

$$f(\tilde{\boldsymbol{X}}) = \boldsymbol{W}\tilde{\boldsymbol{X}}\left[\frac{1}{N}\right]_{N\times N} \tag{C.8}$$

which applies self-attention (equation 5) on the one-hot representation of the masked document $\tilde{\boldsymbol{X}} \in \{0,1\}^{(Tv+1)\times N}$.

Lemma 1 (optimal linear transform when freezing uniform attention)**.** *Consider the simplified transformer architecture given by equation C.8 with , as well as the masked language modeling objective (equation 1) with squared loss (equation 3). Then the set of minimizers* $\operatorname{argmin} L(\boldsymbol{W})$ *consists of all* $\boldsymbol{W} \in \mathbb{R}^{(Tv+1)\times(Tv+1)}$ *that satisfy: there exist constants* $u_0, \cdots, u_{Tv} \in \mathbb{R}$ *such that*

1. *The 0-th row of* $\boldsymbol{W}$*:*
   (a) $\boldsymbol{W}_{00} = -\left(\frac{1}{p_m(1-p_c-p_r)} - 1\right)\cdot u_0$
   (b) $\forall t \in [T], \sum_{l\in t} \boldsymbol{W}_{0l} = u_0 v$
2. *The 0-th column of* $\boldsymbol{W}$*:*
   (a) $\forall i \in \{1, \cdots, Tv\}, \boldsymbol{W}_{i0} = -\frac{p_r}{(1-p_c-p_r)(1-(1-p_c)p_m)Tv} - \left(\frac{1}{(1-p_c-p_r)p_m} - 1\right) u_i$
3. $\boldsymbol{W}_{ij}$ *(*$\forall i, j \in \{1, \cdots, Tv\}$*):*
   (a) $\sum_{l\in \texttt{topic}(i)} \boldsymbol{W}_{il} = \frac{1}{1-(1-p_c)p_m} + u_i v$
   (b) $\forall t \in [T]$ *such that* $\texttt{topic}(i) \neq t$, $\sum_{l\in t} \boldsymbol{W}_{il} = u_i v$

Remark 8. *At the first glance, it might seem that the objective has a unique optima because it involves a squared loss, which is strongly convex. However, such uniqueness is undermined by the uniform attention condition:* $\boldsymbol{W}$ *is multiplied with a rank-1 matrix* $A(\tilde{\boldsymbol{X}}) = \left[\frac{1}{N}\right]_{N\times N}$*. This* $A(\tilde{\boldsymbol{X}})$ *will appear as a matrix multiplier in the Hessian of the objective with respect to* $\boldsymbol{W}$*, and so the Hessian is of rank 1, and therefore cannot have a positive minimum eigenvalue, implying that the objective is in fact* not *strongly convex.*

In fact, this optimization objective becomes strongly convex with an L_2 *regularization for some* $\lambda > 0$*.*

$$\operatorname*{argmin}_{\boldsymbol{W}^V} L_{MLM}(\boldsymbol{W}^V) + \lambda\|\boldsymbol{W}^V\|_F$$

Proof. For document $\boldsymbol{w}$ and the corresponding (masked) one-hot embedding $\tilde{\boldsymbol{X}}$:

$$\begin{aligned}
&\left[\tilde{\boldsymbol{X}} A(\tilde{\boldsymbol{X}})\right]_{ij} \\
&= \frac{1}{N}\sum_{l=1}^{N} \tilde{\boldsymbol{X}}_{il} \quad \text{(i.e. independent of } j\text{)} \\
&= \frac{1}{N}\sum_{l=1}^{N} \mathbf{1}_{\tilde{\boldsymbol{X}}_{il}=1} \quad \text{(since } \tilde{\boldsymbol{X}} \text{ is one-hot)} \\
&= \begin{cases} p_m(1-p_c-p_r) & \text{if } i=0 \\ P_{\boldsymbol{w}}(i)(1-(1-p_c)p_m) + \frac{p_m p_r}{vT} & \text{if } i \in \{1,\cdots,Tv\} \end{cases} \quad \text{(by equation F.17)}
\end{aligned}$$

Thus, the model prediction $\boldsymbol{W}\tilde{\boldsymbol{X}}A(\tilde{\boldsymbol{X}})$ satisfies

$$\begin{aligned}
&(\boldsymbol{W}\tilde{\boldsymbol{X}}A(\tilde{\boldsymbol{X}}))_{ij} = \boldsymbol{W}_{i0}p_m(1-p_c-p_r) + \sum_{l=1}^{Tv}\boldsymbol{W}_{il}\left(P_{\boldsymbol{w}}(l)(1-(1-p_c)p_m) + \frac{p_m p_r}{vT}\right) \\
&= \boldsymbol{W}_{i0}p_m(1-p_c-p_r) + (1-(1-p_c)p_m)\cdot\sum_{l=1}^{Tv}\boldsymbol{W}_{il}P_{\boldsymbol{w}}(l) + \frac{p_m p_r}{vT}\cdot\sum_{l=1}^{Tv}\boldsymbol{W}_{il} \\
&= \boldsymbol{W}_{i0}p_m(1-p_c-p_r) + (1-(1-p_c)p_m)\cdot\left(\sum_{l\in\texttt{topic}(i)}\boldsymbol{W}_{il}P_{\boldsymbol{w}}(i) + \sum_{l\notin\texttt{topic}(i)}\boldsymbol{W}_{il}P_{\boldsymbol{w}}(l)\right) + \frac{p_m p_r}{vT}\cdot\sum_{l=1}^{Tv}\boldsymbol{W}_{il}
\end{aligned} \tag{C.9}$$

and the last step follows since $\forall l \in \texttt{topic}(i), P_{\boldsymbol{w}}(l) = P_{\boldsymbol{w}}(i)$ under our setting in Section 3.1.

Recall that the loss is

$$L(\boldsymbol{W}) = \mathbb{E}_{\boldsymbol{X}\sim\mathcal{D}_{\boldsymbol{X}}}\mathbb{E}_M \frac{1}{|M|}\sum_{j\in M}\|(\boldsymbol{W}\tilde{\boldsymbol{X}}A(\tilde{\boldsymbol{X}}))_{:j} - \boldsymbol{X}_{:j}\|_2^2$$

We will show that the average taken over $j \in M$ is the same as the average taken over all positions $j \in [N]$, by Assumption 1 and because M is uniformly randomly sampled from $[N]$. Moreover, note that $A(\tilde{\boldsymbol{X}}) = \left[\frac{1}{N}\right]_{N\times N}$, so $(\boldsymbol{W}\tilde{\boldsymbol{X}}A(\tilde{\boldsymbol{X}}))_{:j}$ is independent of j. The above observations imply that the loss can be simplified to

$$L(\boldsymbol{W}) = \mathbb{E}_{\boldsymbol{X}\sim\mathcal{D}_{\boldsymbol{X}}}\frac{1}{N}\sum_{j=1}^{N}\|(\boldsymbol{W}\tilde{\boldsymbol{X}}A(\tilde{\boldsymbol{X}}))_{:j} - \boldsymbol{X}_{:j}\|_2^2$$

and so $L(\boldsymbol{W})$ is minimized when $\forall \boldsymbol{X}$,

$$(\boldsymbol{W}\tilde{\boldsymbol{X}}A(\tilde{\boldsymbol{X}}))_{:j} = \frac{1}{N}\sum_{l=1}^{N}\boldsymbol{X}_{:l}$$

which requires $\forall i \in \{0,\cdots,Tv+1\}$,

$$\begin{aligned}
&(\boldsymbol{W}\tilde{\boldsymbol{X}}A(\tilde{\boldsymbol{X}}))_{0j} = 0 \\
&(\boldsymbol{W}\tilde{\boldsymbol{X}}A(\tilde{\boldsymbol{X}}))_{ij} = P_{\boldsymbol{w}}(i), \quad \forall i \in \{1,\cdots,Tv\}
\end{aligned} \tag{C.10}$$

From equation C.9 and equation C.10 we get:

$$\begin{aligned}
0 &= \boldsymbol{W}_{00}p_m(1-p_c-p_r) + (1-(1-p_c)p_m)\cdot\sum_{l=1}^{Tv}\boldsymbol{W}_{0l}P_{\boldsymbol{w}}(l) + \frac{p_m p_r}{vT}\cdot\sum_{l=1}^{Tv}\boldsymbol{W}_{0l} \\
P_{\boldsymbol{w}}(i) &= \boldsymbol{W}_{i0}p_m(1-p_c-p_r) + (1-(1-p_c)p_m)\cdot\left(\sum_{l\in\texttt{topic}(i)}\boldsymbol{W}_{il}P_{\boldsymbol{w}}(i) + \sum_{l\notin\texttt{topic}(i)}\boldsymbol{W}_{il}P_{\boldsymbol{w}}(l)\right) + \frac{p_m p_r}{vT}\cdot\sum_{l=1}^{Tv}\boldsymbol{W}_{il}
\end{aligned} \tag{C.11}$$

Note that under the topic modeling distribution in Section 3.1, for any topic $t \in [T]$,

$$P_{\boldsymbol{w}}((t-1)v+1) = P_{\boldsymbol{w}}((t-1)v+2) = \cdots P_{\boldsymbol{w}}(tv)$$

Hence we simplify equation C.11 by considering the proportions of the "representative" tokens for each topic:

$$\{P_{\boldsymbol{w}}(tv) : t \in [T]\}$$

We obtain: for all sets of $\{P_{\boldsymbol{w}}(i) : i \in [Tv]\}$ satisfying our distribution in Section 3.1

$$0 = \boldsymbol{W}_{00} p_m (1-p_c-p_r) + (1-(1-p_c)p_m) \cdot \sum_{t=1}^{T} \sum_{l \in t} \boldsymbol{W}_{0l} P_{\boldsymbol{w}}(tv) + \frac{p_m p_r}{vT} \cdot \sum_{t=1}^{T} \sum_{l \in t} \boldsymbol{W}_{0l} \tag{C.12}$$

and $\forall i \in \{1, \cdots, Tv\}$

$$P_{\boldsymbol{w}}(i) = \boldsymbol{W}_{i0} p_m (1-p_c-p_r) + (1-(1-p_c)p_m) \cdot \left(\sum_{l \in \texttt{topic}(i)} \boldsymbol{W}_{il} P_{\boldsymbol{w}}(i) + \sum_{t \neq \texttt{topic}(i)} \sum_{l \in t} \boldsymbol{W}_{il} P_{\boldsymbol{w}}(tv) \right) + \frac{p_m p_r}{vT} \cdot \sum_{l=1}^{Tv} \boldsymbol{W}_{il} \tag{C.13}$$

Claim 1. *$\forall i \in \{1, \cdots, Tv\}, \exists u_i \in \mathbb{R}$ such that $\forall t \neq \texttt{topic}(i), \sum_{l \in t} \boldsymbol{W}_{il} = u_i v$. When $i = 0, \exists u_0 \in \mathbb{R}$ such that $\forall t \in [T], \sum_{l \in t} \boldsymbol{W}_{0l} = u_0 v$.*

Proof. $\forall i \in \{1, \cdots, Tv\}, \exists u_i \in \mathbb{R}$, suppose towards contradiction that $\exists t_1, t_2 \neq \texttt{topic}(i)$ such that $\sum_{l \in t_1} \boldsymbol{W}_{il} > \sum_{l \in t_2} \boldsymbol{W}_{il}$. We will show that equation C.13 cannot hold for all sets of $\{P_{\boldsymbol{w}}(i) : i \in [Tv]\}$ satisfying our distribution in Section 3.1.

Specifically, fix $P_{\boldsymbol{w}}(i) = \frac{1}{2v}$ and consider the following settings of $\{P_{\boldsymbol{w}}(j) : j \notin \texttt{topic}(i)\}$:

- $P_{\boldsymbol{w}}(j) = \frac{1}{2v}$ if $\texttt{topic}(j) = t_1$ and 0 otherwise. Then equation C.13 becomes

$$\frac{1}{2v} = \boldsymbol{W}_{i0} p_m (1-p_c-p_r) + (1-(1-p_c)p_m) \cdot \left(\sum_{l \in \texttt{topic}(i)} \boldsymbol{W}_{il} \frac{1}{2v} + \sum_{l \in t_1} \boldsymbol{W}_{il} \frac{1}{2v} \right) + \frac{p_m p_r}{vT} \cdot \sum_{l=1}^{Tv} \boldsymbol{W}_{il}$$

- $P_{\boldsymbol{w}}(j) = \frac{1}{2v}$ if $\texttt{topic}(j) = t_2$ and 0 otherwise. Then equation C.13 becomes

$$\frac{1}{2v} = \boldsymbol{W}_{i0} p_m (1-p_c-p_r) + (1-(1-p_c)p_m) \cdot \left(\sum_{l \in \texttt{topic}(i)} \boldsymbol{W}_{il} \frac{1}{2v} + \sum_{l \in t_2} \boldsymbol{W}_{il} \frac{1}{2v} \right) + \frac{p_m p_r}{vT} \cdot \sum_{l=1}^{Tv} \boldsymbol{W}_{il}$$

Clearly the above two equations cannot both hold, because $\sum_{l \in t_1} \boldsymbol{W}_{il} > \sum_{l \in t_2} \boldsymbol{W}_{il}$.

Hence we proved by contradiction that $\forall t_1, t_2 \neq \texttt{topic}(i), \sum_{l \in t_1} \boldsymbol{W}_{il} = \sum_{l \in t_2} \boldsymbol{W}_{il}$. Likewise, when $i = 0$, $\forall t_1, t_2 in [T], \sum_{l \in t_1} \boldsymbol{W}_{0l} = \sum_{l \in t_2} \boldsymbol{W}_{0l}$.

□

By Claim 1, equation C.12 becomes

$$\begin{aligned}
0 &= \boldsymbol{W}_{00} p_m (1-p_c-p_r) + (1-(1-p_c)p_m) \cdot \sum_{t=1}^{T} u_0 v P_{\boldsymbol{w}}(tv) + \frac{p_m p_r}{vT} \cdot \sum_{t=1}^{T} u_0 v \\
&= \boldsymbol{W}_{00} p_m (1-p_c-p_r) + (1-(1-p_c)p_m) \cdot u_0 + \frac{p_m p_r}{vT} \cdot T u_0 v \\
&= \boldsymbol{W}_{00} p_m (1-p_c-p_r) + (1-(1-p_c)p_m) \cdot u_0 + p_m p_r u_0 \\
&= \boldsymbol{W}_{00} p_m (1-p_c-p_r) + (1-(1-p_c-p_r)p_m) \cdot u_0
\end{aligned}$$

Therefore

$$\boldsymbol{W}_{00} = -\frac{(1-(1-p_c-p_r)p_m)\cdot u_0}{p_m(1-p_c-p_r)} = -\left(\frac{1}{p_m(1-p_c-p_r)} - 1\right)\cdot u_0$$

By Claim 1, equation C.13 becomes

$$\begin{aligned}
P_{\boldsymbol{w}}(i) &= \boldsymbol{W}_{i0}p_m(1-p_c-p_r) + (1-(1-p_c)p_m)\cdot\left(\sum_{l\in\texttt{topic}(i)}\boldsymbol{W}_{il}P_{\boldsymbol{w}}(i) + \sum_{t\neq\texttt{topic}(i)} u_i v P_{\boldsymbol{w}}(tv)\right)\\
&\quad + \frac{p_m p_r}{vT}\cdot\Big(\sum_{l\in\texttt{topic}(i)}\boldsymbol{W}_{il} + (T-1)u_i v\Big)\\
&= \boldsymbol{W}_{i0}p_m(1-p_c-p_r) + (1-(1-p_c)p_m)\cdot\left(\sum_{l\in\texttt{topic}(i)}\boldsymbol{W}_{il}P_{\boldsymbol{w}}(i) + u_i(1-vP_{\boldsymbol{w}}(i))\right)\\
&\quad + \frac{p_m p_r}{vT}\cdot\Big(\sum_{l\in\texttt{topic}(i)}\boldsymbol{W}_{il} + (T-1)u_i v\Big)\\
&= (1-(1-p_c)p_m)\Big(\sum_{l\in\texttt{topic}(i)}\boldsymbol{W}_{il} - u_i v\Big)P_{\boldsymbol{w}}(i) + \boldsymbol{W}_{i0}p_m(1-p_c-p_r) + (1-(1-p_c)p_m)\,u_i\\
&\quad + \frac{p_m p_r}{vT}\cdot\Big(\sum_{l\in\texttt{topic}(i)}\boldsymbol{W}_{il} + (T-1)u_i v\Big)
\end{aligned}$$

Since this has to hold for all $P_{\boldsymbol{w}}(i) \in [0, \frac{1}{v}]$, the coefficients must match, i.e.

$$(1-(1-p_c)p_m)\Big(\sum_{l\in\texttt{topic}(i)}\boldsymbol{W}_{il} - u_i v\Big) = 1 \tag{C.14}$$

$$\boldsymbol{W}_{i0}p_m(1-p_c-p_r) + (1-(1-p_c)p_m)\,u_i + \frac{p_m p_r}{vT}\cdot\Big(\sum_{l\in\texttt{topic}(i)}\boldsymbol{W}_{il} + (T-1)u_i v\Big) = 0 \tag{C.15}$$

By equation C.14,

$$\sum_{l\in\texttt{topic}(i)}\boldsymbol{W}_{il} = u_i v + \frac{1}{1-(1-p_c)p_m}$$

Plugging into equation C.15,

$$\begin{aligned}
\boldsymbol{W}_{i0} &= -\frac{(1-(1-p_c)p_m)\,u_i + \frac{p_m p_r}{vT}\cdot(u_i v + \frac{1}{1-(1-p_c)p_m} + (T-1)u_i v)}{p_m(1-p_c-p_r)}\\
&= -\frac{(1-(1-p_c)p_m)\,u_i + \frac{p_m p_r}{vT}\cdot(\frac{1}{1-(1-p_c)p_m} + Tu_i v)}{p_m(1-p_c-p_r)}\\
&= -\frac{(1-(1-p_c)p_m)\,u_i + \frac{p_m p_r}{vT(1-(1-p_c)p_m)} + p_m p_r u_i}{p_m(1-p_c-p_r)}\\
&= -\frac{p_r}{(1-p_c-p_r)vT(1-(1-p_c)p_m)} - \frac{(1-(1-p_c-p_r)p_m)}{p_m(1-p_c-p_r)}u_i\\
&= -\frac{p_r}{(1-p_c-p_r)(1-(1-p_c)p_m)Tv} - \left(\frac{1}{p_m(1-p_c-p_r)} - 1\right)u_i
\end{aligned}$$

□

D. PROOF OF THEOREM 1: OPTIMAL TOKEN EMBEDDING

Theorem (optimal token embedding, Theorem 1 restated)**.** *Consider training a transformer given by equation 6 with* $\boldsymbol{W}^K = 0, \boldsymbol{W}^Q = 0, \boldsymbol{W}^V = I$ *and* $\forall i \in \{1, \cdots, Tv\}, \boldsymbol{b}_i^{pred} = -\frac{p_m p_r}{(1-(1-p_c)p_m)Tv}$ *on data coming from the topic model described in Section 3, with the masked language modeling objective (equation 1) with squared loss (equation 3).*

Then, the optimal word embeddings $\boldsymbol{W}^E$ *are such that* $\boldsymbol{E} := \boldsymbol{W}^{E\top}\boldsymbol{W}^E$ *satisfies: there exist constants* $u_0, \cdots, u_{Tv} \in \mathbb{R}$ *such that*

1. *The 0-th row of* $\boldsymbol{E}$*:*
   (a) $\boldsymbol{E}_{00} = -\left(\frac{1}{p_m(1-p_c-p_r)} - 1\right) \cdot u_0$
   (b) $\forall t \in [T], \sum_{l \in t} \boldsymbol{E}_{0l} = u_0 v$
2. *The 0-th column of* $\boldsymbol{E}$*:*
   (a) $\forall i \in \{1, \cdots, Tv\}, \boldsymbol{E}_{i0} = -\left(\frac{1}{(1-p_c-p_r)p_m} - 1\right) u_i$
3. $\boldsymbol{E}_{ij}$ *(*$\forall i, j \in \{1, \cdots, Tv\}$*):*
   (a) $\sum_{l \in \texttt{topic}(i)} \boldsymbol{E}_{il} = \frac{1}{1-(1-p_c)p_m} + u_i v$
   (b) $\forall t \in [T]$ *such that* $\texttt{topic}(i) \neq t$*,* $\sum_{l \in t} \boldsymbol{E}_{il} = u_i v$

Proof. Under this setting, the model output is

$$\begin{aligned} f(\tilde{\boldsymbol{X}}) &= \boldsymbol{W}^{E\top}\boldsymbol{W}^E \tilde{\boldsymbol{X}} A(\boldsymbol{W}^E \tilde{\boldsymbol{X}}) + \boldsymbol{b}^{\text{pred}} \\ &= \boldsymbol{E}\tilde{\boldsymbol{X}} \frac{1}{N} \mathbf{1}_{N\times N} + \boldsymbol{b}^{\text{pred}} \\ &= \boldsymbol{E}'\tilde{\boldsymbol{X}} \frac{1}{N} \mathbf{1}_{N\times N} \end{aligned} \tag{D.16}$$

in which $\mathbf{1}$ refers to the all-one matrix, and $\boldsymbol{E}' \in \mathbb{R}^{(Tv+1)\times(Tv+1)}$ is defined such that

$$\boldsymbol{E}'_{ij} = \begin{cases} \boldsymbol{E}_{ij} - \frac{p_r}{(1-p_r-p_c)(1-(1-p_c)p_m)Tv}, & \text{if } i \in \{1, \cdots, Tv\}, j = 0 \\ \boldsymbol{E}_{ij}, & \text{otherwise} \end{cases}$$

and the last step is because by equation F.17,

$$\left(\tilde{\boldsymbol{X}} \frac{1}{N} \mathbf{1}_{N\times N}\right)_{0j} = p_m(1-p_c-p_r) \quad \forall j$$

and $\forall i \in \{1, \cdots, Tv\}$,

$$\begin{aligned} &\left(\boldsymbol{E}'\tilde{\boldsymbol{X}} \frac{1}{N} \mathbf{1}_{N\times N}\right)_{ij} \\ &= \left(\boldsymbol{E}\tilde{\boldsymbol{X}} \frac{1}{N} \mathbf{1}_{N\times N}\right)_{ij} - \frac{p_r}{(1-p_r-p_c)\,(1-(1-p_c)p_m)\,Tv} \cdot p_m(1-p_c-p_r) \\ &= \left(\boldsymbol{E}\tilde{\boldsymbol{X}} \frac{1}{N} \mathbf{1}_{N\times N}\right)_{ij} - \frac{p_m p_r}{(1-(1-p_c)p_m)\,Tv} \\ &= \left(\boldsymbol{E}\tilde{\boldsymbol{X}} \frac{1}{N} \mathbf{1}_{N\times N}\right)_{ij} + \boldsymbol{b}_i^{\text{pred}} \end{aligned}$$

Let $\boldsymbol{E}'^*$ denote any matrix in

$$\operatorname*{argmin}_{\boldsymbol{E}'} \mathbb{E}_{\boldsymbol{X}\sim\mathcal{D}_{\boldsymbol{X}}} \mathbb{E}_M \frac{1}{|M|} \sum_{j\in M} \|(\boldsymbol{E}'\tilde{\boldsymbol{X}} \frac{1}{N} \mathbf{1}_{N\times N})_{:j} - \boldsymbol{X}_{:j}\|_2^2$$

then by Lemma 1, there exist constants $u_0, \cdots, u_{Tv} \in \mathbb{R}$ such that

1. The 0-th row of $\boldsymbol{E}'^*$:
   (a) $\boldsymbol{E}'^*_{00} = -\left(\frac{1}{p_m(1-p_c-p_r)} - 1\right) \cdot u_0$
   (b) $\forall t \in [T], \sum_{l \in t} \boldsymbol{E}'^*_{0l} = u_0 v$
2. The 0-th column of $\boldsymbol{E}'^*$:
   (a) $\forall i \in \{1, \cdots, Tv\}, \boldsymbol{E}'^*_{i0} = -\frac{p_r}{(1-p_c-p_r)(1-(1-p_c)p_m)Tv} - \left(\frac{1}{(1-p_c-p_r)p_m} - 1\right) u_i$
3. $\boldsymbol{E}'^*_{ij}$ ($\forall i, j \in \{1, \cdots, Tv\}$):
   (a) $\sum_{l \in \texttt{topic}(i)} \boldsymbol{E}'^*_{il} = \frac{1}{1-(1-p_c)p_m} + u_i v$
   (b) $\forall t \in [T]$ such that $\texttt{topic}(i) \neq t$, $\sum_{l \in t} \boldsymbol{E}'^*_{il} = u_i v$

Therefore, by equation D.16, let $\boldsymbol{E}^*$ denote any matrix in

$$\operatorname*{argmin}_{\boldsymbol{E}} \mathbb{E}_{\boldsymbol{X} \sim \mathcal{D}_{\boldsymbol{X}}} \mathbb{E}_M \frac{1}{|M|} \sum_{j \in M} \|(\boldsymbol{E}\tilde{\boldsymbol{X}} \frac{1}{N} \mathbf{1}_{N \times N})_{:j} + \boldsymbol{b}^{\text{pred}} - \boldsymbol{X}_{:j}\|_2^2$$

then there exist constants $u_0, \cdots, u_{Tv} \in \mathbb{R}$ such that

1. The 0-th row of $\boldsymbol{E}^*$:
   (a) $\boldsymbol{E}^*_{00} = -\left(\frac{1}{p_m(1-p_c-p_r)} - 1\right) \cdot u_0$
   (b) $\forall t \in [T], \sum_{l \in t} \boldsymbol{E}^*_{0l} = u_0 v$
2. The 0-th column of $\boldsymbol{E}^*$:
   (a) $\forall i \in \{1, \cdots, Tv\}, \boldsymbol{E}^*_{i0} = -\left(\frac{1}{(1-p_c-p_r)p_m} - 1\right) u_i$
3. $\boldsymbol{E}^*_{ij}$ ($\forall i, j \in \{1, \cdots, Tv\}$):
   (a) $\sum_{l \in \texttt{topic}(i)} \boldsymbol{E}^*_{il} = \frac{1}{1-(1-p_c)p_m} + u_i v$
   (b) $\forall t \in [T]$ such that $\texttt{topic}(i) \neq t$, $\sum_{l \in t} \boldsymbol{E}^*_{il} = u_i v$

Finally, note that a subset of this family of optima is *realizable*, in the sense that there exists such $\boldsymbol{E}^*$ and $u_0, \cdots, u_{Tv} \in \mathbb{R}$ s.t. there exists $\boldsymbol{W}^E \in \mathbb{R}^{d \times (Tv+1)}$ s.t. $\boldsymbol{E}^* = \boldsymbol{W}^{E\top} \boldsymbol{W}^E$. The simplest example is

$$\begin{aligned} u_0, \cdots, u_{Tv} &= 0 \\ d &= Tv + 1 \\ \boldsymbol{E}^* &= \frac{1}{1-(1-p_c)p_m} I \\ \boldsymbol{W}^E &= \frac{1}{\sqrt{1-(1-p_c)p_m}} I \end{aligned}$$

□

E. PROVING OPTIMAL W^V IN SELF-ATTENTION

E.1. Optimal W^V when freezing uniform attention without regularization

Theorem 4 (optimal $\boldsymbol{W}^V$ when freezing uniform attention)**.** *On the topic modeling data distribution described in Section 3.1, with the* `topic` *relation defined in Definition 1, under Assumption 1, with a single layer transformer given by equation 7 whose* $\boldsymbol{W}^K = 0, \boldsymbol{W}^Q = 0, \boldsymbol{b}^{pred} = 0$*, under masked language modeling objective (equation 1) with the squared loss (equation 3),* $\operatorname{argmin} L(\boldsymbol{W}^V)$ *consists of all* $\boldsymbol{W}^V \in \mathbb{R}^{(Tv+1)\times(Tv+1)}$ *that satisfy: there exist constants* $u_0, \cdots, u_{Tv} \in \mathbb{R}$ *such that*

1. *The 0-th row of* $\boldsymbol{W}^V$*:*
   - *(a)* $\boldsymbol{W}^V_{00} = -\left(\frac{1}{p_m(1-p_c-p_r)} - 1\right) \cdot u_0$
   - *(b)* $\forall t \in [T], \sum_{l\in t} \boldsymbol{W}^V_{0l} = u_0 v$
2. *The 0-th column of* $\boldsymbol{W}^V$*:*
   - *(a)* $\forall i \in \{1, \cdots, Tv\}, \boldsymbol{W}^V_{i0} = -\frac{p_r}{(1-p_c-p_r)(1-(1-p_c)p_m)Tv} - \left(\frac{1}{(1-p_c-p_r)p_m} - 1\right) u_i$
3. $\boldsymbol{W}^V_{ij}$ *(*$\forall i, j \in \{1, \cdots, Tv\}$*):*
   - *(a)* $\sum_{l\in \texttt{topic}(i)} \boldsymbol{W}^V_{il} = \frac{1}{1-(1-p_c)p_m} + u_i v$
   - *(b)* $\forall t \in [T]$ *such that* $\texttt{topic}(i) \neq t$*,* $\sum_{l\in t} \boldsymbol{W}^V_{il} = u_i v$

Proof. Note that this is exactly the statement of Lemma 1 (proved in Appendix C.1) in the case of $\boldsymbol{W} := \boldsymbol{W}^V$. □

E.2. Proof of Theorem 2: case when adding L_2 regularization

Theorem (optimal $\boldsymbol{W}^V$ with mild L_2-regularization when freezing uniform attention, restated)**.** *On the topic modeling data distribution described in Section 3.1, with the* `topic` *relation defined in Definition 1, under Assumption 1, with a single layer transformer given by equation 7 whose* $\boldsymbol{W}^K = 0, \boldsymbol{W}^Q = 0, \boldsymbol{b}^{pred} = 0$*, under the* L_2*-regularized masked language modeling objective (equation 2) with the squared loss (equation 3),* $\lim_{\lambda\to 0} \operatorname{argmin} L_{l2reg}(\boldsymbol{W}^V) = \{\boldsymbol{W}^{V*}\}$ *in which* $\boldsymbol{W}^{V*} \in \mathbb{R}^{(Tv+1)\times(Tv+1)}$ *satisfies:*

1. *The 0-th row of* $\boldsymbol{W}^{V*}$*:*
   - *(a)* $\forall j \in \{0, \cdots, Tv\}, \boldsymbol{W}^{V*}_{0j} = 0$
2. *The 0-th column of* $\boldsymbol{W}^{V*}$*:*
   - *(a)* $\forall i \in \{1, \cdots, Tv\}, \boldsymbol{W}^{V*}_{i0} = \frac{c_2c_3 - c_1Tv}{c_2^2 + Tv}$
3. $\boldsymbol{W}^{V*}_{ij}$ *(*$\forall i, j \in \{1, \cdots, Tv\}$*):*
   - *(a)* $\forall l \notin \texttt{topic}(i),\ \boldsymbol{W}^{V*}_{il} = \boldsymbol{W}^{V*}_{\textit{diff-topic}} := -\frac{c_1c_2 + c_3}{c_2^2 + Tv}$
   - *(b)* $\forall l \in \texttt{topic}(i),\ \boldsymbol{W}^{V*}_{il} = \boldsymbol{W}^{V*}_{\textit{same-topic}} := \boldsymbol{W}^{V*}_{\textit{diff-topic}} + \frac{c_3}{v}$

in which the constants

- $c_1 = \frac{p_r}{(1-p_c-p_r)(1-(1-p_c)p_m)Tv} \in (0, 1)$
- $c_2 = \frac{1}{(1-p_c-p_r)p_m} - 1 \in (0, +\infty)$
- $c_3 = \frac{1}{1-(1-p_c)p_m} \in (1, +\infty)$

Proof. We proceed in the following two steps.

Step 1: the optima converges to one outlined in Lemma 1

Let S denote the set of optima outlined in Lemma 1. Suppose towards contradiction that $\exists \boldsymbol{W}^{V*} \notin S$ such that $\boldsymbol{W}^{V*} \in \lim_{\lambda \to 0} \operatorname{argmin} L_{\text{l2reg}}(\boldsymbol{W}^V)$.

In comparison, $\forall \boldsymbol{W} \in S$, by Lemma 1, since $\boldsymbol{W}^{V*} \notin S$,

$$L(\boldsymbol{W}) < L(\boldsymbol{W}^{V*})$$

Moreover, note that since $\|\boldsymbol{W}\|_F$ is finite,

$$\lim_{\lambda \to 0} \lambda \|\boldsymbol{W}\|_F^2 = 0 \leq \lim_{\lambda \to 0} \lambda \|\boldsymbol{W}^{V*}\|_F^2$$

Combining the above two observations gives

$$\lim_{\lambda \to 0} L_{\text{l2reg}}(\boldsymbol{W}) = L(\boldsymbol{W}) + \lim_{\lambda \to 0} \lambda \|\boldsymbol{W}\|_F^2 < L(\boldsymbol{W}^{V*}) + \lim_{\lambda \to 0} \lambda \|\boldsymbol{W}^{V*}\|_F^2 = \lim_{\lambda \to 0} L_{\text{l2reg}}(\boldsymbol{W}^{V*})$$

which contradicts $\boldsymbol{W}^{V*} \in \lim_{\lambda \to 0} \operatorname{argmin} L_{\text{l2reg}}(\boldsymbol{W}^V)$.

Therefore, we have proved by contradiction that

$$\forall \boldsymbol{W}^V \in \lim_{\lambda \to 0} \operatorname{argmin} L_{\text{l2reg}}(\boldsymbol{W}^V), \quad \boldsymbol{W}^V \in S$$

Step 2: solve for the coefficients that minimize the L_2 penalty

By Step 1,

$$\begin{aligned}
\lim_{\lambda \to 0} \operatorname{argmin} L_{\text{l2reg}}(\boldsymbol{W}^V) &= \lim_{\lambda \to 0} \operatorname*{argmin}_{\boldsymbol{W}^V \in S} L_{\text{l2reg}}(\boldsymbol{W}^V) \\
&= \lim_{\lambda \to 0} \operatorname*{argmin}_{\boldsymbol{W}^V \in S} L(\boldsymbol{W}^V) + \lambda \|\boldsymbol{W}^V\|_F^2 \\
&= \lim_{\lambda \to 0} \operatorname*{argmin}_{\boldsymbol{W}^V \in S} \min L(\boldsymbol{W}^V) + \lambda \|\boldsymbol{W}^V\|_F^2 \\
&= \lim_{\lambda \to 0} \operatorname*{argmin}_{\boldsymbol{W}^V \in S} \lambda \|\boldsymbol{W}^V\|_F^2
\end{aligned}$$

in which the last step is because $\forall \boldsymbol{W}^V \in S$, $L(\boldsymbol{W}^V) = \min L(\boldsymbol{W}^V)$, which is a constant independent of $\boldsymbol{W}^V$.

Then it suffices to find the constants $u_0, \cdots, u_{Tv} \in \mathbb{R}$ that minimizes $\|\boldsymbol{W}^V\|_F$.

□

F. ADDITIONAL RESULTS ON ATTENTION WEIGHTS

F.1. Helping lemmas on masking probabilities

In this section, we will calculate a few expressions for the masking probabilities, which will be useful for the proofs later on. We will also introduce a few constants for brevity of notation.

A straightforward calculation shows that the probabilities after the masking process satisfy:

Proposition 1 (Probabilities after masking)**.** *After the masking process as in Section 3.1 is applied to a document $\boldsymbol{w}$, the distribution for the new document $\tilde{\boldsymbol{w}}$ satisfies*

$$P_{\tilde{\boldsymbol{w}}}(i) = \begin{cases} \frac{1}{v\tau}(1-(1-p_c)p_m) + \frac{p_m p_r}{vT}, & \text{if } \texttt{topic}(i) \in \{t_1, \cdots, t_\tau\} \\ p_m(1-p_c-p_r), & \text{if } i = \texttt{[MASK]} := 0 \\ \frac{p_m p_r}{vT}, & \text{otherwise} \end{cases} \tag{F.17}$$

For convenience, we will introduce the notation

$$p_1 := \frac{1}{v\tau}(1-(1-p_c)p_m) + \frac{p_m p_r}{vT} \tag{F.18}$$

$$p_2 := \frac{p_m p_r}{vT} \tag{F.19}$$

Another straightforward calculation can be used to express the relationship between the constant c_3 in Assumption 2 and the α, β. Namely, we have:

Proposition 2 (Expressing c_3 in terms of α, β)**.** *The constant c_3 in Assumption 2 satisfies:*

$$c_3 = \begin{cases} \frac{1}{(\beta p_1 + \alpha p_1(v-1) + p_1 v(\tau-1) + p_2 v(T-\tau))N}, & \text{if } \tilde{w}_j \in \{t_i\}_{i\in[\tau]} \\ \frac{1}{(\beta p_2 + \alpha p_2(v-1) + p_1 v\tau + p_2 v(T-\tau-1))N}, & \text{if } \tilde{w}_j \in [T]\backslash\{t_i\}_{i\in[\tau]} \end{cases}$$

Again, for notational convenience, we will introduce z_1, z_2, s.t.

$$z_1 := \beta p_1 + \alpha p_1(v-1) + p_1 v(\tau-1) + p_2 v(T-\tau)$$
$$z_2 := \beta p_2 + \alpha p_2(v-1) + p_1 v\tau + p_2 v(T-\tau-1)$$

Proof of Proposition 2. We will get these equalities by considering the marginalization constraints, depending on the topic of $\tilde{w}_j$. Consider first a j, such that $\tilde{w}_j \in \{t_i\}_{i\in[\tau]}$:

- Note, for every position i, with probability p_1, we have $\tilde{w}_i = \tilde{w}_j$, so $A(\tilde{\boldsymbol{X}})_{ij} = \beta c_3$ by Assumption 2.
- Note also, for every position i, with probabilitiy $p_1(v-1)$, we have $\tilde{w}_i \neq \tilde{w}_j$ but $\texttt{topic}(\tilde{w}_i) = \texttt{topic}(\tilde{w}_j)$, and so $A(\tilde{\boldsymbol{X}})_{ij} = \alpha c_3$.
- Finally, note that for every position i, with probability $(p_1 v(\tau-1) + p_2 v(T-\tau))$ we have $\texttt{topic}(\tilde{w}_i) \neq \texttt{topic}(\tilde{w}_j)$, so $A(\tilde{\boldsymbol{X}})_{ij} = c_3$.

Since $\sum_{i=1}^N A(\tilde{\boldsymbol{X}})_{ij} = 1$, we obtain $c_3 = \frac{1}{(\beta p_1 + \alpha p_1(v-1) + p_1 v(\tau-1) + p_2 v(T-\tau))N}$.

Consider next a j, s.t. $\tilde{w}_j \in [T]\backslash\{t_i\}_{i\in[\tau]}$. By similar considerations as before,

- With probability p_2, a position i in $\tilde{\boldsymbol{w}}$ satisfies $\tilde{w}_i = \tilde{w}_j$, so $A(\tilde{\boldsymbol{X}})_{ij} = \beta c_3$.
- With probability $p_2(v-1)$, a position i satisfies $\tilde{w}_i \neq \tilde{w}_j$ but $\texttt{topic}(\tilde{w}_i) = \texttt{topic}(\tilde{w}_j)$, so $A(\tilde{\boldsymbol{X}})_{ij} = \alpha c_3$.
- Finally, with probability $p_1 v\tau + p_2 v(T-\tau-1)$, a position i in satisfies $\texttt{topic}(\tilde{w}_i) \neq \texttt{topic}(\tilde{w}_j)$, so $A(\tilde{\boldsymbol{X}})_{ij} = c_3$.

Since $\sum_{i=1}^N A(\tilde{\boldsymbol{X}})_{ij} = 1$, we obtain $c_3 = \frac{1}{(\beta p_2 + \alpha p_2(v-1) + p_1 v\tau + p_2 v(T-\tau-1))N}$. The proposition thus follows.

□

F.2. Implication of topic-wise attention assumption on model output

In this section we calculate the part $\tilde{\boldsymbol{X}}A(\tilde{\boldsymbol{X}})$ using the results of Appendix F.1:

Proposition 3. *Using the calculation and notations of z_1 and z_2 in Appendix F.1:*

$$\left[\tilde{\boldsymbol{X}}A(\tilde{\boldsymbol{X}})\right]_{ij} = \sum_{l=1}^{N} \tilde{\boldsymbol{X}}_{il} A(\tilde{\boldsymbol{X}})_{lj} = \sum_{l=1}^{N} \mathbf{1}_{\tilde{\boldsymbol{X}}_{il}=1} A(\tilde{\boldsymbol{X}})_{lj} = P(\tilde{\boldsymbol{X}}_{il}=1)\cdot N \cdot A(\tilde{\boldsymbol{X}})_{lj}$$

$$= \begin{cases} p_1 N \frac{\beta}{z_1 N} = \frac{p_1\beta}{z_1}, & \textit{if } i=j, \texttt{topic}(j) \in \{t_1,\cdots,t_\tau\} \textbf{\textit{(Same token)}} \\ p_1 N \frac{\alpha}{z_1 N} = \frac{p_1\alpha}{z_1}, & \textit{if } i\neq j, \texttt{topic}(i) = \texttt{topic}(j) \in \{t_1,\cdots,t_\tau\} \textbf{\textit{(Different token, same topic)}} \\ p_1 N \frac{1}{z_1 N} = \frac{p_1}{z_1}, & \textit{if } \texttt{topic}(i) \neq \texttt{topic}(j), \texttt{topic}(i) \in \{t_1,\cdots,t_\tau\}, \texttt{topic}(j) \in \{t_1,\cdots,t_\tau\} \\ p_2 N \frac{1}{z_1 N} = \frac{p_2}{z_1}, & \textit{if } \texttt{topic}(i) \neq \texttt{topic}(j), \texttt{topic}(i) \notin \{t_1,\cdots,t_\tau\}, \texttt{topic}(j) \in \{t_1,\cdots,t_\tau\} \\ p_2 N \frac{\beta}{z_2 N} = \frac{p_2\beta}{z_2}, & \textit{if } i=j, \texttt{topic}(j) \notin \{t_1,\cdots,t_\tau\} \\ p_2 N \frac{\alpha}{z_2 N} = \frac{p_2\alpha}{z_2}, & \textit{if } i\neq j, \texttt{topic}(i) = \texttt{topic}(j) \notin \{t_1,\cdots,t_\tau\} \\ p_1 N \frac{1}{z_2 N} = \frac{p_1}{z_2}, & \textit{if } \texttt{topic}(i) \neq \texttt{topic}(j), \texttt{topic}(i) \in \{t_1,\cdots,t_\tau\}, \texttt{topic}(j) \notin \{t_1,\cdots,t_\tau\} \\ p_2 N \frac{1}{z_2 N} = \frac{p_2}{z_2}, & \textit{if } \texttt{topic}(i) \neq \texttt{topic}(j), \texttt{topic}(i) \notin \{t_1,\cdots,t_\tau\}, \texttt{topic}(j) \notin \{t_1,\cdots,t_\tau\} \end{cases} \tag{F.20}$$

F.3. Proof of Theorem 3 (optimal attention when freezing $\boldsymbol{W}^V$ to uniform blocks)

Theorem (optimal attention weights when freezing block-wise $\boldsymbol{W}^V$, Theorem 3 restated)**.** *Suppose the data distribution follows the topic modeling assumption in Section 3.1 and Assumption 1. Suppose we train a single layer transformer given by equation 7 with $\boldsymbol{b}^{pred} = 0$ and $\boldsymbol{W}^V$ frozen to the optima in Theorem 2, under masked language modeling objective (equation 1) with the squared loss (equation 3), under Assumption 2, Assumption 3, and Assumption 4. Then, the optimal (α, β) satisfy*

$$\frac{v-1}{v}\alpha + \frac{1}{v}\beta \in (\lambda_1(\tau-1), \lambda_2 T)$$

in which the constants $\lambda_1 := \frac{(1-(1-p_c)p_m+p_m p_r)(1+(1-p_c)p_m)}{2(1-(1-p_c)p_m)}$ and $\lambda_2 := 100(\frac{1-(1-p_c)p_m}{p_m p_r}+1)$.

Proof. Define $\gamma := \frac{v-1}{v}\alpha + \frac{1}{v}\beta$.

Recall the architecture under consideration, i.e.

$$\hat{\boldsymbol{X}} := \boldsymbol{W}^V \tilde{\boldsymbol{X}} A(\tilde{\boldsymbol{X}})$$

The squared loss (equation 3) is

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}\sim\mathcal{D}_{\boldsymbol{X}}}\mathbb{E}_M\left[\frac{1}{|M|}\sum_{j\in M} l(f(\tilde{\boldsymbol{X}})_{:j}, \boldsymbol{X}_{:j})\right]_{ij} \\
&= \frac{1}{p_m N}\mathbb{E}_{\boldsymbol{X}\sim\mathcal{D}_{\boldsymbol{X}}}\mathbb{E}_M\left[\sum_{j:\tilde{w}_j=\texttt{[MASK]}} l(f(\tilde{\boldsymbol{X}})_{:j}, \boldsymbol{X}_{:j}) + \sum_{j\in M, \tilde{w}_j\neq\texttt{[MASK]}} l(f(\tilde{\boldsymbol{X}})_{:j}, \boldsymbol{X}_{:j})\right] \\
&= \frac{1}{p_m N}\mathbb{E}_{\boldsymbol{X}\sim\mathcal{D}_{\boldsymbol{X}}}\mathbb{E}_M\left[\sum_{j:\tilde{w}_j=\texttt{[MASK]}} \|(\boldsymbol{W}^V\tilde{\boldsymbol{X}}A(\tilde{\boldsymbol{X}}))_{:j} - \boldsymbol{X}_{:j}\|_2^2 + \sum_{j\in M, \tilde{w}_j\neq\texttt{[MASK]}} \|(\boldsymbol{W}^V\tilde{\boldsymbol{X}}A(\tilde{\boldsymbol{X}}))_{:j} - \boldsymbol{X}_{:j}\|_2^2\right] \\
&= \frac{1}{p_m N}\mathbb{E}_{\boldsymbol{X}\sim\mathcal{D}_{\boldsymbol{X}}}\mathbb{E}_M\left[\sum_{j:\tilde{w}_j=\texttt{[MASK]}} \|\boldsymbol{W}^V\tilde{\boldsymbol{X}}A(\tilde{\boldsymbol{X}})_{:j} - \boldsymbol{X}_{:j}\|_2^2 + \sum_{j\in M, \tilde{w}_j\neq\texttt{[MASK]}} \|\boldsymbol{W}^V\tilde{\boldsymbol{X}}A(\tilde{\boldsymbol{X}})_{:j} - \boldsymbol{X}_{:j}\|_2^2\right]
\end{aligned}$$

Note that when $\tilde{w}_j = \texttt{[MASK]}$, $A(\tilde{\boldsymbol{X}})_{:j}$ is the attention from [MASK] to other tokens, and therefore is independent of the setting of α and β in Assumption 2. Thus, in the following, we only consider the case in which $j \in M, \tilde{w}_j \neq \texttt{[MASK]}$, namely, w_j is masked, but $\tilde{w}_j$ is chosen to be either the correct token or the random token. Hence define:

$$L(\gamma) := \frac{1}{p_m N}\mathbb{E}_{\boldsymbol{X}\sim\mathcal{D}_{\boldsymbol{X}}}\mathbb{E}_M\left[\sum_{j\in M, \tilde{w}_j\neq\texttt{[MASK]}} \|\boldsymbol{W}^V\tilde{\boldsymbol{X}}A(\tilde{\boldsymbol{X}})_{:j} - \boldsymbol{X}_{:j}\|_2^2\right] \tag{F.21}$$

Note that $\forall \boldsymbol{y} \in \mathbb{R}^{Tv+1}$

$$(\boldsymbol{W}^V\boldsymbol{y})_i = \begin{cases} 0, & i = 0 \\ q(\frac{1}{v}\sum_{l\in\text{topic}(i)} y_l), & i \in \{1, \cdots, Tv+1\} \end{cases}$$

in which

$$q(x) := \frac{1}{1-(1-p_c)p_m}x - \frac{p_m p_r}{(1-(1-p_c)p_m)Tv}$$

In our context, we will consider $\boldsymbol{y} = \tilde{\boldsymbol{X}}A(\tilde{\boldsymbol{X}})_{:j}$ in $L(\gamma)$ above.

For a document $\boldsymbol{w}$ which contains topics $t_1, \cdots, t_\tau \in [T]$, there are the following cases:

Case 1: $\texttt{topic}(\tilde{w}_j) = \texttt{topic}(w_j)$ When $\tilde{w}_j$ after masking belongs to the same topic as the correct token w_j. (This happens with probability $p_c + \frac{p_r}{T}$)

By equation F.20,

$$\frac{1}{v}\sum_{l\in\text{topic}(i)}\left[\tilde{\boldsymbol{X}}A(\tilde{\boldsymbol{X}})\right]_{lj} = \begin{cases} \frac{1}{v}\left(\frac{p_1\beta}{z_1} + \sum_{l\in\text{topic}(i),l\neq\tilde{w}_j}\frac{p_1\alpha}{z_1}\right) = \frac{1}{v}\left(\frac{p_1\beta+(v-1)p_1\alpha}{z_1}\right) = \frac{p_1\gamma}{z_1}, & \text{if } \texttt{topic}(i) = \texttt{topic}(\tilde{w}_j) \\ \frac{1}{v}\left(\sum_{l\in\text{topic}(i)}\frac{p_1}{z_1}\right) = \frac{p_1}{z_1}, & \text{if } \texttt{topic}(i) \neq \texttt{topic}(\tilde{w}_j), \texttt{topic}(i) \in \{t_1,\cdots,t_\tau\} \\ \frac{1}{v}\left(\sum_{l\in\text{topic}(i)}\frac{p_2}{z_1}\right) = \frac{p_2}{z_1}, & \text{if } \texttt{topic}(i) \neq \texttt{topic}(\tilde{w}_j), \texttt{topic}(i) \notin \{t_1,\cdots,t_\tau\} \end{cases} \tag{F.22}$$

Recall that the label is

$$\boldsymbol{X}_{:j} = \begin{cases} 1, & i = w_j \\ 0, & i \in \{0,\cdots,Tv\}\backslash w_j \end{cases}$$

Hence the contribution to the loss from token $\tilde{w}_j$ is

$$\begin{aligned}
&(p_c + \frac{p_r}{T})[\left(1 - q\left(\frac{p_1\gamma}{z_1}\right)\right)^2 + q(\frac{p_1\gamma}{z_1})^2\cdot(v-1) + q(\frac{p_1}{z_1})^2\cdot v(\tau-1) + q(\frac{p_2}{z_1})^2\cdot v(T-\tau)] \\
&= (p_c + \frac{p_r}{T})[(1 - q(\frac{p_1\gamma}{p_1\gamma v + p_1 v(\tau-1) + p_2 v(T-\tau)}))^2 \\
&+ q(\frac{p_1\gamma}{p_1\gamma v + p_1 v(\tau-1) + p_2 v(T-\tau)})^2(v-1) + q(\frac{p_1}{p_1\gamma v + p_1 v(\tau-1) + p_2 v(T-\tau)})^2 v(\tau-1) \\
&+ q(\frac{p_2}{p_1\gamma v + p_1 v(\tau-1) + p_2 v(T-\tau)})^2 v(T-\tau)]
\end{aligned}$$

Plugging in the asymptotics from Assumption 3, the above becomes

$$p_c[(1 - q(\frac{p_1\gamma}{p_1\gamma v + p_1 v(\tau-1) + p_2 v(T-\tau)}))^2 + q(\frac{p_1\gamma}{p_1\gamma v + p_1 v(\tau-1) + p_2 v(T-\tau)})^2(v-1)] \pm O(\frac{1}{T}) \tag{F.23}$$

Case 2: $\texttt{topic}(\tilde{w}_j) \in \{t_1,\cdots,t_\tau\}\backslash\{\texttt{topic}(w_j)\}$ When $\tilde{w}_j$ after masking belongs to a different topic from that of the correct token w_j, but still a topic existing in $\boldsymbol{w}$. (This happens with probability $\frac{p_r(\tau-1)}{T}$)

$\left[\tilde{\boldsymbol{X}}A(\tilde{\boldsymbol{X}})\right]_{:j}$ is the same as equation F.22.

Hence the loss is

$$\begin{aligned}
&p_r\frac{\tau-1}{T}[(1 - q(\frac{p_1}{z_1}))^2 + q(\frac{p_1\gamma}{z_1})^2\cdot v + q(\frac{p_1}{z_1})^2\cdot(v(\tau-1)-1) + q(\frac{p_2}{z_1})^2\cdot v(T-\tau)] \\
&= p_r\frac{\tau-1}{T}[(1 - q(\frac{p_1}{p_1\gamma v + p_1 v(\tau-1) + p_2 v(T-\tau)})^2 + q(\frac{p_1\gamma}{p_1\gamma v + p_1 v(\tau-1) + p_2 v(T-\tau)})^2\cdot v \\
&\quad + q(\frac{p_1}{p_1\gamma v + p_1 v(\tau-1) + p_2 v(T-\tau)})^2(v(\tau-1)-1) \\
&+ q(\frac{p_2}{p_1\gamma v + p_1 v(\tau-1) + p_2 v(T-\tau)})^2 v(T-\tau)]
\end{aligned}$$

Plugging in the asymptotics from Assumption 3, the above terms vanish.

Case 3: $\texttt{topic}(\tilde{w}_j) \in [T]\backslash\{t_1,\cdots,t_\tau\}$ When $\tilde{w}_j$ after masking belongs to a topic that does not exist in $\boldsymbol{w}$. (This happens with probability $p_r(1-\frac{\tau}{T})$)

By equation F.20,

$$\left[\tilde{\boldsymbol{X}} A(\tilde{\boldsymbol{X}})\right]_{ij} = \begin{cases} \frac{p_2\beta}{z_2}, & \text{if } i = \tilde{w}_j \\ \frac{p_2\alpha}{z_2}, & \text{if } i \neq \tilde{w}_j, \texttt{topic}(i) = \texttt{topic}(\tilde{w}_j) \\ \frac{p_1}{z_2}, & \text{if } \texttt{topic}(i) \neq \texttt{topic}(\tilde{w}_j), \texttt{topic}(i) \in \{t_1, \cdots, t_\tau\} \\ \frac{p_2}{z_2}, & \text{if } \texttt{topic}(i) \neq \texttt{topic}(\tilde{w}_j), \texttt{topic}(i) \notin \{t_1, \cdots, t_\tau\} \end{cases} \tag{F.24}$$

$$\frac{1}{v}\sum_{l \in \text{topic}(i)} \left[\tilde{\boldsymbol{X}} A(\tilde{\boldsymbol{X}})\right]_{lj} = \begin{cases} \frac{1}{v}\left(\frac{p_2\beta}{z_2} + \sum_{l\in\text{topic}(i), l\neq\tilde{w}_j} \frac{p_2\alpha}{z_2}\right) = \frac{1}{v}\left(\frac{p_2\beta + (v-1)p_2\alpha}{z_2}\right) = \frac{p_2\gamma}{z_2}, & \text{if } \texttt{topic}(i) = \texttt{topic}(\tilde{w}_j) \\ \frac{1}{v}\left(\sum_{l\in\text{topic}(i)} \frac{p_1}{z_2}\right) = \frac{p_1}{z_2}, & \text{if } \texttt{topic}(i) \neq \texttt{topic}(\tilde{w}_j), \texttt{topic}(i) \in \{t_1, \cdots, t_\tau\} \\ \frac{1}{v}\left(\sum_{l\in\text{topic}(i)} \frac{p_2}{z_2}\right) = \frac{p_2}{z_2}, & \text{if } \texttt{topic}(i) \neq \texttt{topic}(\tilde{w}_j), \texttt{topic}(i) \notin \{t_1, \cdots, t_\tau\} \end{cases} \tag{F.25}$$

Hence the loss is

$$\begin{aligned} & p_r(1-\frac{\tau}{T})[(1-q(\frac{p_1}{z_2}))^2 + q(\frac{p_1}{z_2})^2 \cdot (v\tau - 1) + q(\frac{p_2\gamma}{z_2})^2 \cdot v + q(\frac{p_2}{z_2})^2 \cdot v(T-\tau-1)] \\ &= p_r(1-\frac{\tau}{T})[(1-q(\frac{p_1}{p_2\gamma v + p_1 v\tau + p_2 v(T-\tau-1)}))^2 + q(\frac{p_1}{p_2\gamma v + p_1 v\tau + p_2 v(T-\tau-1)})^2(v\tau-1) \\ &+ q(\frac{p_2\gamma}{p_2\gamma v + p_1 v\tau + p_2 v(T-\tau-1)})^2 v + q(\frac{p_2}{p_2\gamma v + p_1 v\tau + p_2 v(T-\tau-1)})^2 v(T-\tau-1)] \end{aligned}$$

Plugging in the asymptotics from Assumption 3, the above becomes

$$p_r(1 + q(\frac{p_2\gamma}{p_2\gamma v + p_1 v\tau + p_2 v(T-\tau-1)})^2) \tag{F.26}$$

Combining the above cases Adding equation F.23 and equation F.26, we can see in the asymptotic regime of interest, we have:

$$\begin{aligned} L(\gamma) &= p_c[(1-q(\frac{p_1\gamma}{p_1\gamma v + p_1 v(\tau-1) + p_2 v(T-\tau)}))^2 + q(\frac{p_1\gamma}{p_1\gamma v + p_1 v(\tau-1) + p_2 v(T-\tau)})^2(v-1)] \\ &\quad + p_r(1 + q(\frac{p_2\gamma}{p_1\gamma v + p_1 v\tau + p_2 v(T-\tau-1)})^2) \pm O(\frac{1}{T}) \\ &= p_c[(1-\frac{c_4 p_1\gamma}{p_1\gamma v + p_1 v(\tau-1) + p_2 v(T-\tau)})^2 + (\frac{c_4 p_1\gamma}{p_1\gamma v + p_1 v(\tau-1) + p_2 v(T-\tau)})^2(v-1)] \\ &\quad + p_r + p_r(\frac{c_4 p_2\gamma}{p_2\gamma v + p_1 v\tau + p_2 v(T-\tau-1)})^2 \pm O(\frac{1}{T}) \end{aligned}$$

in which the constant c_4 is defined as

- $c_4 \coloneqq \frac{1}{1-(1-p_c)p_m} \in (1, 2)$

Plugging in the definition of p_1, p_2 in equation F.18, equation F.19

$$\begin{aligned} L(\gamma) &= p_c[(1-\frac{\frac{1}{v\tau}\gamma}{\frac{1}{c_4\tau}\gamma + \frac{1}{c_4\tau}(\tau-1) + \frac{p_m p_r}{T}(T-\tau)})^2 + (\frac{\frac{1}{v\tau}\gamma}{\frac{1}{c_4\tau}\gamma + \frac{1}{c_4\tau}(\tau-1) + \frac{p_m p_r}{T}(T-\tau)})^2(v-1)] \\ &\quad + p_r + p_r(\frac{c_4\frac{p_m p_r}{vT}\gamma}{\frac{p_m p_r}{vT}\gamma v + \frac{1}{c_4} + \frac{p_m p_r}{T}(T-\tau-1)})^2 \pm O(\frac{1}{T}) \\ &= p_c\left[(1-\frac{c_4\gamma}{v\gamma + v(\tau-1) + c_4 p_m p_r v\tau})^2 + (\frac{c_4\gamma}{v\gamma + v(\tau-1) + c_4 p_m p_r v\tau})^2(v-1)\right] \\ &\quad + p_r + p_r(\frac{c_4 p_m p_r\gamma}{p_m p_r\gamma v + (\frac{1}{c_4} + p_m p_r)vT})^2 \pm O(\frac{1}{T}) \end{aligned} \tag{F.27}$$

We will again consider several possible cases for γ in equation F.27.

Case 1: When $\gamma \leq \frac{(1+c_4 p_m p_r)(2-c_4)}{2c_4}(\tau - 1)$.

Let c_5 denote the constant:

$$c_5 := \frac{(1 + c_4 p_m p_r)(2 - c_4)}{2c_4}$$

then focusing on this term in the loss equation F.27:

$$\begin{aligned}
&\frac{c_4\gamma}{v\gamma + v(\tau-1) + c_4 p_m p_r v\tau} \\
&< \frac{c_4\gamma}{v\gamma + v\frac{\gamma}{c_5} + c_4 p_m p_r v\frac{\gamma}{c_5}} \\
&= \frac{c_4}{v + v\frac{1}{c_5} + c_4 p_m p_r v\frac{1}{c_5}} \\
&= \frac{c_4}{v(1 + \frac{1+c_4 p_m p_r}{c_5})}
\end{aligned}$$

and so

$$L(\gamma) > p_c \left(1 - \frac{c_4}{v(1 + \frac{1+c_4 p_m p_r}{c_5})}\right)^2 + p_r \pm o(1) \tag{F.28}$$

Case 2: When $\gamma \geq 100\frac{\frac{1}{c_4}+p_m p_r}{p_m p_r}T$.

then since $\tau = o(T)$ by Assumption 3:

$$\begin{aligned}
\frac{c_4\gamma}{v\gamma + v(\tau-1) + c_4 p_m p_r v\tau} &= \frac{c_4\gamma}{v\gamma} + o(1) = \frac{c_4}{v} + o(1) \\
\frac{c_4\gamma}{v\gamma + v(\tau-1) + c_4 p_m p_r v\tau} &= \frac{c_4\gamma}{v\gamma} + o(1) = \frac{c_4}{v} + o(1) \\
\frac{c_4 p_m p_r \gamma}{p_m p_r \gamma v + (\frac{1}{c_4} + p_m p_r)vT} &\geq \frac{c_4 p_m p_r \gamma}{p_m p_r \gamma v + \frac{1}{100} p_m p_r v\gamma} = \frac{100c_4}{101v}
\end{aligned}$$

and therefore plugging into equation F.27:

$$\begin{aligned}
L(\gamma) &\geq p_c[(1 - \frac{c_4}{v} \pm o(1))^2 + (\frac{c_4}{v} \pm o(1))^2(v-1)] + p_r + p_r(\frac{100c_4}{101v})^2 \pm o(1) \\
&= p_c[1 - \frac{2c_4}{v} + \frac{c_4^2}{v^2} + \frac{c_4^2}{v^2}(v-1)] + p_r + p_r(\frac{100c_4}{101v})^2 \pm o(1) \\
&= p_c[1 - \frac{2c_4}{v} + \frac{c_4^2}{v}] + p_r + p_r(\frac{100c_4}{101v})^2 \pm o(1) \\
&= p_c[1 - \frac{c_4(2-c_4)}{v}] + p_r + p_r(\frac{100c_4}{101v})^2 \pm o(1)
\end{aligned} \tag{F.29}$$

Case 3: When $\frac{(1+c_4 p_m p_r)(2-c_4)}{2c_4}(\tau - 1) < \gamma < 100\frac{\frac{1}{c_4}+p_m p_r}{p_m p_r}T$.

Note that this case is the complement of Case 1 and Case 2 above, and so we have considered all possibilities. We will show that there exists γ in this case such that $L(\gamma)$ is smaller (by an $\Omega(1)$ constant difference) than the lower bound of

$L(\gamma)$ proven in Case 1 and Case 2 above, based on which we know $\operatorname{argmin} L(\gamma)$ cannot lie in Case 1 or Case 2, and thus conclude that $\operatorname{argmin} L(\gamma)$ is within this case. Specifically, let:

$$\gamma = \sqrt{\tau T}$$

then similar to Case 2, since $\tau = o(T)$ by Assumption 3:

$$\begin{aligned}
\frac{c_4\gamma}{v\gamma + v(\tau-1) + c_4 p_m p_r v\tau} &= \frac{c_4\gamma}{v\gamma} + o(1) = \frac{c_4}{v} + o(1) \\
\frac{c_4\gamma}{v\gamma + v(\tau-1) + c_4 p_m p_r v\tau} &= \frac{c_4\gamma}{v\gamma} + o(1) = \frac{c_4}{v} + o(1) \\
\frac{c_4 p_m p_r \gamma}{p_m p_r \gamma v + (\frac{1}{c_4} + p_m p_r) vT} &= o(1)
\end{aligned}$$

and therefore plugging into equation F.27:

$$\begin{aligned}
L(\gamma) &= p_c[(1 - \frac{c_4\gamma}{v\gamma} \pm o(1))^2 + (\frac{c_4\gamma}{v\gamma} \pm o(1))^2 (v-1)] + p_r \pm o(1) \\
&= p_c[(1 - \frac{c_4}{v} \pm o(1))^2 + (\frac{c_4}{v} \pm o(1))^2 (v-1)] + p_r \pm o(1) \\
&= p_c[1 - \frac{2c_4}{v} + \frac{c_4^2}{v^2} + \frac{c_4^2}{v^2}(v-1)] + p_r \pm o(1) \\
&= p_c[1 - \frac{2c_4}{v} + \frac{c_4^2}{v}] + p_r \pm o(1) \\
&= p_c[1 - \frac{c_4(2-c_4)}{v}] + p_r \pm o(1)
\end{aligned} \tag{F.30}$$

Comparing the above cases

Note that $L(\gamma)$ in Case 3 is strictly smaller than $L(\gamma)$ in Case 1 and Case 2, because:

- Comparing equation F.28 and equation F.30: $(1 - \frac{c_4}{v(1+\frac{1+c_4 p_m p_r}{c_5})})^2 > 1 - \frac{c_4(2-c_4)}{v}$ because $c_5 \in (0, \frac{(1+c_4 p_m p_r)(2-c_4)}{c_4})$
- Comparing equation F.29 and equation F.30: in the former, the term $p_r(\frac{100c_4}{101v})^2 > 0$ is the extra constant (of scale $\Omega(1)$, i.e. non-vanishing even under our asymptotic assumptions Assumption 3) compared with the latter.

Therefore we conclude that

$$\operatorname{argmin} L(\gamma) \subseteq (\frac{(1+c_4 p_m p_r)(2-c_4)}{2c_4}(\tau-1), 100\frac{\frac{1}{c_4} + p_m p_r}{p_m p_r} T)$$

□

Remark 9. *In Theorem 3, we specify some necessary conditions that the optimal γ must satisfy. It is challenging to precisely characterize the optima (to within $o(1)$ error), because doing so may require explicitly writing those smaller scale terms hidden (in $\pm o(1)$) by our asymptotic setting (Assumption 3). Those smaller scale terms, however, do not affect our analysis, because these $\pm o(1)$ terms cannot reverse the $\Omega(1)$ constant separation between the loss in the above different cases.*

F.4. Optimal attention weights (when freezing diagonal $\boldsymbol{W}^V$)

Our Stage-2 analysis on the optimal attention weights (equation 5) is based on freezing $\boldsymbol{W}^V$ to be the Stage-1 optima characterized in Theorem 2. Notably, in Theorem 2, the uniqueness of the optima (i.e. a clean block-wise pattern) crucially depends on the L_2 regularization. Indeed, as we prove in Theorem 4 (in Appendix E), without the regularization, there is a family of optima (depending on a series of free constants) all of which can encode the topic structure.

Among these alternative optima, we are particularly interested in a special case — one that has a *diagonal* pattern. This type of diagonally structured $\boldsymbol{W}^V$ often occurs when we train the single-layered transformer model without L_2 regularization.

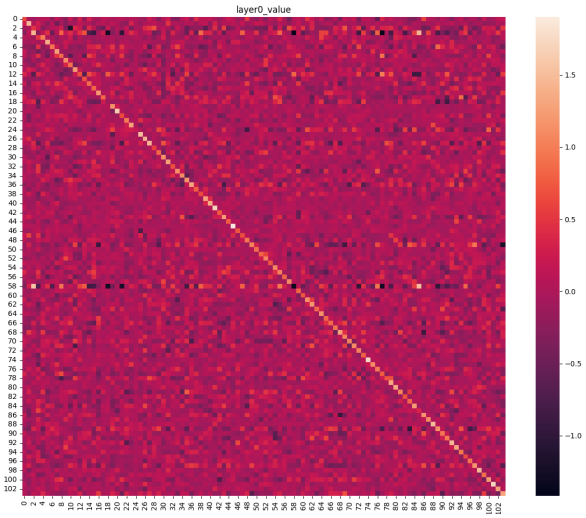

Figure 6. Without L_2 regularization, the trained $\boldsymbol{W}^V$ sometimes shows a *diagonal* pattern, which is a special case of the family of optima characterized in Theorem 4 (in Appendix E).

Motivated by this empirical observation, we formally define the particular optima from Theorem 4 (in Appendix E) that is a diagonal pattern.

Definition 2 (diagonal $\boldsymbol{W}^V$)**.** *The diagonal optima of $\boldsymbol{W}^V$, denoted as $\boldsymbol{D}^V$, is the only matrix in $\mathbb{R}^{(Tv+1)\times(Tv+1)}$ that satisfies both $\boldsymbol{D}^V \in \operatorname{argmin} L(\boldsymbol{W}^V)$ (in Theorem 4) and*

$$\forall i, j \in \{1, \cdots, Tv\},\ \boldsymbol{W}^V_{ij} = 0 \text{ if } i \neq j$$

Corresponding to this case, we provide an analysis on the Stage-2 optimal attention weights, which shows a very interesting different behavior from the result in Theorem 3 (for block-wise $\boldsymbol{W}^V$).

Theorem 5 (optimal attention weights when freezing diagonal $\boldsymbol{W}^V$)**.** *Suppose the data distribution follows the topic modeling assumption in Section 3.1 and Assumption 1. Suppose we train a single layer transformer given by equation 7 with $\boldsymbol{b}^{pred} = 0$ and $\boldsymbol{W}^V$ frozen to $\boldsymbol{D}^V$ in Definition 2, under masked language modeling objective (equation 1) with the squared loss (equation 3), under Assumption 2, Assumption 3, and Assumption 4. Then, the optimal (α, β) satisfy:*

$$\lambda_3\tau < \beta < \lambda_4 T$$
$$\alpha < \lambda_5 \beta$$

in which the constants

$$\lambda_3 := \frac{1-(1-p_c)p_m + p_m p_r}{100} \cdot v$$
$$\lambda_4 := \frac{1-(1-p_c)p_m}{\sqrt{v-1}-2+(1-p_c)p_m} \cdot \frac{1-(1-p_c-p_r)p_m}{p_m p_r} \cdot v$$
$$\lambda_5 := \frac{1}{(v-1)(1-(1-p_c)p_m)}$$

Proof. Following the same steps leading to equation F.21, define:

$$L(\alpha,\beta) := \frac{1}{p_m N}\mathbb{E}_{\boldsymbol{X}\sim\mathcal{D}_{\boldsymbol{X}}}\mathbb{E}_M\left[\sum_{j\in M, \tilde{w}_j \neq \texttt{[MASK]}} \|\boldsymbol{D}^V\tilde{\boldsymbol{X}}A(\tilde{\boldsymbol{X}})_{:j} - \boldsymbol{X}_{:j}\|_2^2\right] \tag{F.31}$$

Note that $\forall \boldsymbol{y} \in \mathbb{R}^{Tv+1}$

$$(\boldsymbol{D}^V\boldsymbol{y})_i = \begin{cases} 0, & i = 0 \\ q(y_i), & i \in \{1,\cdots,Tv+1\} \end{cases}$$

in which $q(x) := \frac{1}{1-(1-p_c)p_m}x - \frac{p_m p_r}{(1-(1-p_c)p_m)Tv}$

In our context, we will consider $\boldsymbol{y} = \tilde{\boldsymbol{X}}A(\tilde{\boldsymbol{X}})_{:j}$ in $L(\alpha,\beta)$ above.

For a document $\boldsymbol{w}$ which contains topics $t_1,\cdots,t_\tau \in [T]$, there are the following cases:

Case 1: $\tilde{w}_j = w_j$ When $\tilde{w}_j$ after masking is the correct token w_j. (This happens with probability $p_c + \frac{p_r}{vT}$)

By equation F.20,

$$\left[\tilde{\boldsymbol{X}}A(\tilde{\boldsymbol{X}})\right]_{ij} = \begin{cases} \frac{p_1\beta}{z_1}, & \text{if } i = \tilde{w}_j \\ \frac{p_1\alpha}{z_1}, & \text{if } i \neq \tilde{w}_j, \texttt{topic}(i) = \texttt{topic}(\tilde{w}_j) \\ \frac{p_1}{z_1}, & \text{if } \texttt{topic}(i) \neq \texttt{topic}(\tilde{w}_j), \texttt{topic}(i) \in \{t_1,\cdots,t_\tau\} \\ \frac{p_2}{z_1}, & \text{if } \texttt{topic}(i) \neq \texttt{topic}(\tilde{w}_j), \texttt{topic}(i) \notin \{t_1,\cdots,t_\tau\} \end{cases} \tag{F.32}$$

Recall that the label is $\boldsymbol{X}_{:j} = \begin{cases} 1, & i = w_j \\ 0, & i \in \{0,\cdots,Tv\}\backslash w_j \end{cases}$

Hence the contribution to the loss from token $\tilde{w}_j$ is

$$\begin{aligned}
&(p_c + \frac{p_r}{vT})[\left(1 - q\left(\frac{p_1\beta}{z_1}\right)\right)^2 + q(\frac{p_1\alpha}{z_1})^2\cdot(v-1) + q(\frac{p_1}{z_1})^2\cdot v(\tau-1) + q(\frac{p_2}{z_1})^2\cdot v(T-\tau)] \\
&= (p_c + \frac{p_r}{vT})[(1 - q(\frac{p_1\beta}{\beta p_1 + \alpha p_1(v-1) + p_1 v(\tau-1) + p_2 v(T-\tau)}))^2 \\
&+ q(\frac{p_1\alpha}{\beta p_1 + \alpha p_1(v-1) + p_1 v(\tau-1) + p_2 v(T-\tau)})^2(v-1) \\
&+ q(\frac{p_1}{\beta p_1 + \alpha p_1(v-1) + p_1 v(\tau-1) + p_2 v(T-\tau)})^2 v(\tau-1) \\
&+ q(\frac{p_2}{\beta p_1 + \alpha p_1(v-1) + p_1 v(\tau-1) + p_2 v(T-\tau)})^2 v(T-\tau)]
\end{aligned}$$

Plugging in the asymptotics from Assumption 3, the above becomes

$$\begin{aligned} p_c[(1-q(\frac{p_1\beta}{\beta p_1+\alpha p_1(v-1)+p_1v(\tau-1)+p_2v(T-\tau)}))^2 \\ +q(\frac{p_1\alpha}{\beta p_1+\alpha p_1(v-1)+p_1v(\tau-1)+p_2v(T-\tau)})^2(v-1)\pm O(\frac{1}{\tau})] \end{aligned} \tag{F.33}$$

Case 2: $\tilde{w}_j \neq w_j, \texttt{topic}(\tilde{w}_j) = \texttt{topic}(w_j)$ When $\tilde{w}_j$ after masking is not the correct token but belongs to the same topic as the correct token w_j. (This happens with probability $\frac{p_r}{T}(1-\frac{1}{v})$)

$\left[\tilde{\boldsymbol{X}}A(\tilde{\boldsymbol{X}})\right]_{:j}$ is the same as equation F.32.

Hence the loss is

$$\begin{aligned} &\frac{p_r}{T}(1-\frac{1}{v})[(1-q(\frac{p_1\alpha}{z_1}))^2+q(\frac{p_1\beta}{z_1}))^2+q(\frac{p_1\alpha}{z_1})^2\cdot(v-2)+q(\frac{p_1}{z_1})^2\cdot v(\tau-1)+q(\frac{p_2}{z_1})^2\cdot v(T-\tau)] \\ &=\frac{p_r}{T}(1-\frac{1}{v})[(1-q(\frac{p_1\alpha}{\beta p_1+\alpha p_1(v-1)+p_1v(\tau-1)+p_2v(T-\tau)}))^2 \\ &+q(\frac{p_1\beta}{\beta p_1+\alpha p_1(v-1)+p_1v(\tau-1)+p_2v(T-\tau)})^2 \\ &+q(\frac{p_1\alpha}{\beta p_1+\alpha p_1(v-1)+p_1v(\tau-1)+p_2v(T-\tau)})^2(v-2) \\ &+q(\frac{p_1}{\beta p_1+\alpha p_1(v-1)+p_1v(\tau-1)+p_2v(T-\tau)})^2v(\tau-1) \\ &+q(\frac{p_2}{\beta p_1+\alpha p_1(v-1)+p_1v(\tau-1)+p_2v(T-\tau)})^2v(T-\tau)] \end{aligned}$$

Plugging in the asymptotics from Assumption 3, the above terms vanish.

Case 3: $\texttt{topic}(\tilde{w}_j) \in \{t_1,\cdots,t_\tau\}\backslash\{\texttt{topic}(w_j)\}$ When $\tilde{w}_j$ after masking belongs to a different topic from that of the correct token w_j, but still a topic existing in $\boldsymbol{w}$. (This happens with probability $\frac{p_r(\tau-1)}{T}$)

$\left[\tilde{\boldsymbol{X}}A(\tilde{\boldsymbol{X}})\right]_{:j}$ is the same as equation F.32.

Hence the loss is

$$\begin{aligned} &p_r\frac{\tau-1}{T}[(1-q(\frac{p_1}{z_1}))^2+q(\frac{p_1\beta}{z_1}))^2+q(\frac{p_1\alpha}{z_1})^2\cdot(v-1)+q(\frac{p_1}{z_1})^2\cdot(v(\tau-1)-1)+q(\frac{p_2}{z_1})^2\cdot v(T-\tau)] \\ &=p_r\frac{\tau-1}{T}[(1-q(\frac{p_1}{\beta p_1+\alpha p_1(v-1)+p_1v(\tau-1)+p_2v(T-\tau)})^2 \\ &+q(\frac{p_1\beta}{\beta p_1+\alpha p_1(v-1)+p_1v(\tau-1)+p_2v(T-\tau)})^2 \\ &+q(\frac{p_1\alpha}{\beta p_1+\alpha p_1(v-1)+p_1v(\tau-1)+p_2v(T-\tau)})^2(v-1) \\ &+q(\frac{p_1}{\beta p_1+\alpha p_1(v-1)+p_1v(\tau-1)+p_2v(T-\tau)})^2(v(\tau-1)-1) \\ &+q(\frac{p_2}{\beta p_1+\alpha p_1(v-1)+p_1v(\tau-1)+p_2v(T-\tau)})^2v(T-\tau)] \end{aligned}$$

Plugging in the asymptotics from Assumption 3, the above terms vanish.

Case 4: $\texttt{topic}(\tilde{w}_j) \in [T]\backslash\{t_1,\cdots,t_\tau\}$ When $\tilde{w}_j$ after masking belongs to a topic that does not exist in $\boldsymbol{w}$. (This happens with probability $p_r(1-\frac{\tau}{T})$)

By equation F.20,

$$\left[\tilde{\boldsymbol{X}} A(\tilde{\boldsymbol{X}})\right]_{ij} = \begin{cases} \frac{p_2\beta}{z_2}, & \text{if } i = \tilde{w}_j \\ \frac{p_2\alpha}{z_2}, & \text{if } i \neq \tilde{w}_j, \texttt{topic}(i) = \texttt{topic}(\tilde{w}_j) \\ \frac{p_1}{z_2}, & \text{if } \texttt{topic}(i) \neq \texttt{topic}(\tilde{w}_j), \texttt{topic}(i) \in \{t_1, \cdots, t_\tau\} \\ \frac{p_2}{z_2}, & \text{if } \texttt{topic}(i) \neq \texttt{topic}(\tilde{w}_j), \texttt{topic}(i) \notin \{t_1, \cdots, t_\tau\} \end{cases} \tag{F.34}$$

Hence the loss is

$$\begin{aligned}
&p_r(1-\frac{\tau}{T})[(1-q(\frac{p_1}{z_2}))^2 + q(\frac{p_1}{z_2})^2 \cdot (v\tau - 1) + q(\frac{p_2\beta}{z_2})^2 + q(\frac{p_2\alpha}{z_2})^2 \cdot (v-1) + q(\frac{p_2}{z_2})^2 \cdot v(T-\tau-1)] \\
&= p_r(1-\frac{\tau}{T})[(1-q(\frac{p_1}{\beta p_2 + \alpha p_2(v-1) + p_1 v\tau + p_2 v(T-\tau-1)}))^2 \\
&+ q(\frac{p_1}{\beta p_2 + \alpha p_2(v-1) + p_1 v\tau + p_2 v(T-\tau-1)})^2 (v\tau - 1) \\
&+ q(\frac{p_2\beta}{\beta p_2 + \alpha p_2(v-1) + p_1 v\tau + p_2 v(T-\tau-1)})^2 + q(\frac{p_2\alpha}{\beta p_2 + \alpha p_2(v-1) + p_1 v\tau + p_2 v(T-\tau-1)})^2 (v-1) \\
&+ q(\frac{p_2}{\beta p_2 + \alpha p_2(v-1) + p_1 v\tau + p_2 v(T-\tau-1)})^2 v(T-\tau-1)]
\end{aligned}$$

Plugging in the asymptotics from Assumption 3, the above becomes

$$\begin{aligned}
&p_r[(1 + q(\frac{p_2\beta}{\beta p_2 + \alpha p_2(v-1) + p_1 v\tau + p_2 v(T-\tau-1)})^2) \\
&\quad + q(\frac{p_2\alpha}{\beta p_2 + \alpha p_2(v-1) + p_1 v\tau + p_2 v(T-\tau-1)})^2 (v-1)] \pm o(1)
\end{aligned} \tag{F.35}$$

Combining the above cases Adding equation F.33 and equation F.35, we can see in the asymptotic regime of interest:

$$\begin{aligned}
L(\alpha,\beta) &= p_c[(1-q(\frac{p_1\beta}{\beta p_1 + \alpha p_1(v-1) + p_1 v(\tau-1) + p_2 v(T-\tau)}))^2 \\
&\quad + q(\frac{p_1\alpha}{\beta p_1 + \alpha p_1(v-1) + p_1 v(\tau-1) + p_2 v(T-\tau)})^2 (v-1)] \\
&\quad + p_r[(1 + q(\frac{p_2\beta}{\beta p_2 + \alpha p_2(v-1) + p_1 v\tau + p_2 v(T-\tau-1)})^2) \\
&\quad + q(\frac{p_2\alpha}{\beta p_2 + \alpha p_2(v-1) + p_1 v\tau + p_2 v(T-\tau-1)})^2 (v-1)] \pm o(1) \\
&= p_c[(1 - \frac{\frac{1}{v\tau}\beta}{\beta\frac{1}{c_4 v\tau} + \alpha\frac{v-1}{c_4 v\tau} + \frac{1}{c_4} + p_m p_r})^2 + (\frac{\frac{1}{v\tau}\alpha}{\beta\frac{1}{c_4 v\tau} + \alpha\frac{v-1}{c_4 v\tau} + \frac{1}{c_4} + p_m p_r})^2 (v-1)] \\
&\quad + p_r[(1 + (\frac{c_4\frac{p_m p_r}{vT}\beta}{\beta\frac{p_m p_r}{vT} + \alpha\frac{p_m p_r(v-1)}{vT} + \frac{1}{c_4} + p_m p_r})^2) + (\frac{c_4\frac{p_m p_r}{vT}\alpha}{\beta\frac{p_m p_r}{vT} + \alpha\frac{p_m p_r(v-1)}{vT} + \frac{1}{c_4} + p_m p_r})^2 (v-1)] \pm o(1)
\end{aligned}$$

in which the constant c_4 is defined as

- $c_4 \coloneqq \frac{1}{1-(1-p_c)p_m} \in (1,2)$ by Assumption 4.

$$\begin{aligned}
L(\alpha,\beta) &= p_c[(1 - \frac{c_4\beta}{\beta + (v-1)\alpha + (1+c_4 p_m p_r)v\tau})^2 + (\frac{c_4\alpha}{\beta + (v-1)\alpha + (1+c_4 p_m p_r)v\tau})^2 (v-1)] \\
&\quad + p_r[(1 + (\frac{c_4\beta}{\beta + (v-1)\alpha + (\frac{1}{c_4 p_m p_r} + 1)vT})^2) + (\frac{c_4\alpha}{\beta + (v-1)\alpha + (\frac{1}{c_4 p_m p_r} + 1)vT})^2 (v-1)] \pm o(1)
\end{aligned} \tag{F.36}$$

We will again consider several possible cases for α, β in equation F.36.

- **Case 1,** $\beta \leq \frac{\mathbf{1+c_4p_mp_r}}{\mathbf{100c_4}}\mathbf{v}\tau$: then $\frac{c_4\beta}{\beta+(v-1)\alpha+(1+c_4p_mp_r)v\tau} \leq \frac{c_4\beta}{\beta+100c_4\beta} < \frac{1}{100}$, and hence

$$L(\alpha,\beta) \geq p_c(1-\frac{1}{100})^2 + p_r \pm o(1)$$

- **Case 2,** $\beta > \frac{\mathbf{1+c_4p_mp_r}}{\mathbf{100c_4}}\mathbf{v}\tau$: we have the following subcases:
  - If $\alpha \geq \frac{c_4}{v-1}\beta$, then $\frac{c_4\beta}{\beta+(v-1)\alpha+(1+c_4p_mp_r)v\tau} < \frac{c_4\beta}{\beta+(v-1)\alpha} \leq \frac{c_4\beta}{\beta+c_4\beta} < \frac{c_4}{1+c_4}$, and hence by equation F.36 $L(\alpha,\beta) \geq p_c(1-\frac{c_4}{1+c_4})^2 + p_r \pm o(1)$.
  - If $\alpha < \frac{c_4}{v-1}\beta$, then $\frac{c_4\beta}{\beta+(v-1)\alpha+(\frac{1}{c_4p_mp_r}+1)vT} > \frac{c_4\beta}{\beta+c_4\beta+(\frac{1}{c_4p_mp_r}+1)vT}$
    * If $\beta \geq c_7(\frac{1}{c_4p_mp_r}+1)vT$ (for some constant $c_7 := \frac{1}{c_4(\sqrt{v-1}-\frac{1}{c_4}-1)}$), then $\frac{c_4\beta}{\beta+(v-1)\alpha+(\frac{1}{c_4p_mp_r}+1)vT} > \frac{c_4\beta}{\beta+c_4\beta+(\frac{1}{c_4p_mp_r}+1)vT} \geq \frac{c_4\beta}{\beta+c_4\beta+\frac{1}{c_7}\beta} = \frac{c_4}{1+c_4+\frac{1}{c_7}}$, and hence $L(\alpha,\beta) > p_r[1+(\frac{c_4}{1+c_4+\frac{1}{c_7}})^2] \pm o(1)$
    * If $\beta < c_7(\frac{1}{c_4p_mp_r}+1)vT$: note that this case is the complement of all cases (and subcases) above, and so we have considered all possibilities. We will show that there exists (α,β) in this case such that $L(\alpha,\beta)$ is smaller (by an $\Omega(1)$ constant difference) than the lower bound of $L(\alpha,\beta)$ proven in all cases above, based on which we know $\operatorname{argmin} L(\alpha,\beta)$ cannot lie in any of the above cases, and thus conclude that $\operatorname{argmin} L(\alpha,\beta)$ is within this case.
      Specifically: let $\alpha = \sqrt{\tau T}$ and $\beta = \frac{v-1}{c_4-1}\alpha = \frac{v-1}{c_4-1}\sqrt{\tau T}$, then

$$\begin{aligned}
\frac{c_4\beta}{\beta+(v-1)\alpha+(1+c_4p_mp_r)v\tau} &= \frac{c_4\frac{v-1}{c_4-1}}{\frac{v-1}{c_4-1}+(v-1)} \pm o(1) = 1 \pm o(1)\\
\frac{c_4\alpha}{\beta+(v-1)\alpha+(1+c_4p_mp_r)v\tau} &= \frac{c_4}{\frac{v-1}{c_4-1}+(v-1)} \pm o(1) = \frac{c_4-1}{v-1} \pm o(1)\\
\frac{c_4\beta}{\beta+(v-1)\alpha+(\frac{1}{c_4p_mp_r}+1)vT} &= o(1)\\
\frac{c_4\alpha}{\beta+(v-1)\alpha+(\frac{1}{c_4p_mp_r}+1)vT} &= o(1)
\end{aligned}$$

      Plugging into equation F.36:

$$\begin{aligned}
L(\alpha,\beta) &= p_c[(1-(1\pm o(1)))^2 + (\frac{c_4-1}{v-1} \pm o(1))^2(v-1)] + p_r[(1+(o(1))^2) + (o(1))^2(v-1)] \pm o(1)\\
&= p_c[\frac{(c_4-1)^2}{v-1}] + p_r \pm o(1)\\
&< p_c\frac{1}{v-1} + p_r \pm o(1)
\end{aligned}$$

    Note that this is smaller than all previous cases, because
    * $\frac{1}{v-1} < (1-\frac{1}{100})^2$ since v is a large finite constant (see Assumption 3 and Assumption 4).
    * $\frac{1}{v-1} < (1-\frac{c_4}{1+c_4})^2$ since v is a large finite constant (see Assumption 3 and Assumption 4).
    * $\frac{1}{v-1} < (\frac{c_4}{1+c_4+\frac{1}{c_7}})^2$ by the definition of c_7 above.

Therefore, we conclude that all $\alpha,\beta > 0$ that minimize $L(\alpha,\beta)$ must satisfy

$$\begin{aligned}
\frac{1+c_4p_mp_r}{100c_4}v\tau < &\beta < c_7(\frac{1}{c_4p_mp_r}+1)vT\\
&\alpha < \frac{c_4}{v-1}\beta
\end{aligned}$$

□

Remark 10. *Remark 9 applies to this proof too.*

F.5. Loss landscape with respect to attention weights in the non-asymptotic setting

When T, τ are finite, the loss expression turns out to be too complicated to characterize in closed form (because all the $o(1)$ terms need to be expanded). So we instead numerically compute the loss landscape as a function of α and β.

We set $T = 100$ following our experimental setup on Wikipedia dataset (in Section 6), and $v = 300$ (so total vocabulary size $Tv = 30000$) following the pre-trained BERT tokenizer in Huggingface implementation (Wolf et al., 2020). We will vary $\tau \in \{20, 40, 60, 80\}$.

Diagonal $\boldsymbol{W}^V$ First, when $\boldsymbol{W}^V$ is fixed to a diagonal structure (Definition 2), Theorem 5 predicts that the loss is lowest when β is within an interval (boundaries controlled by τ and T), and α is less than a constant multiple of β. Both constraints are visible in the non-asymptotic setting, as we show in the following:

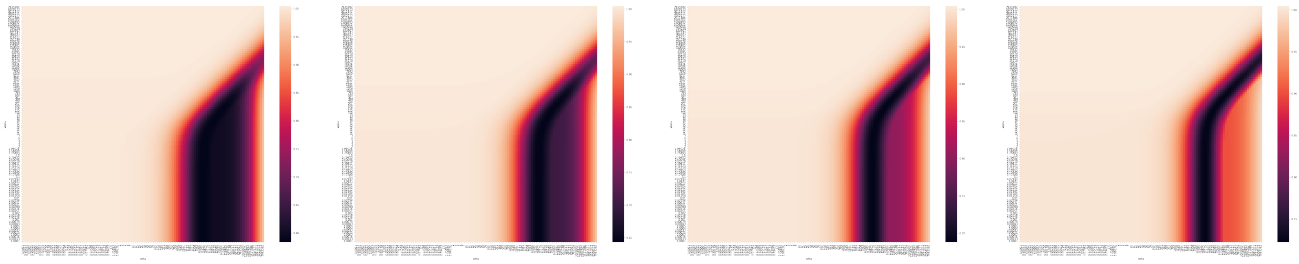

Figure 7. Landscape of squared loss under diagonal $\boldsymbol{W}^V$ (Definition 2), $T = 100, v = 300$. (left-to-right) $\tau = 20$, $\tau = 40$, $\tau = 60$, $\tau = 80$. In each plot, we perform a grid search over $\alpha, \beta \in [10^{-4}, 10^7]$ (both axes use log-scale). Darker color represents lower loss. Across a wide range of τ (compared to T), the loss is lowest when β is within an interval (lower bound growing with τ), and the optimal α is less than a constant multiple of β.

$\boldsymbol{W}^V$ with uniform blocks On the other hand, when $\boldsymbol{W}^V$ is fixed to a block-wise structure with uniform blocks (i.e. optima in Theorem 2), Theorem 3 predicts that the loss is lowest when a convex combination of α and β is within an interval (boundaries controlled by τ and T). As we show in the following, a variant of this constraint visibly holds in the non-asymptotic setting.

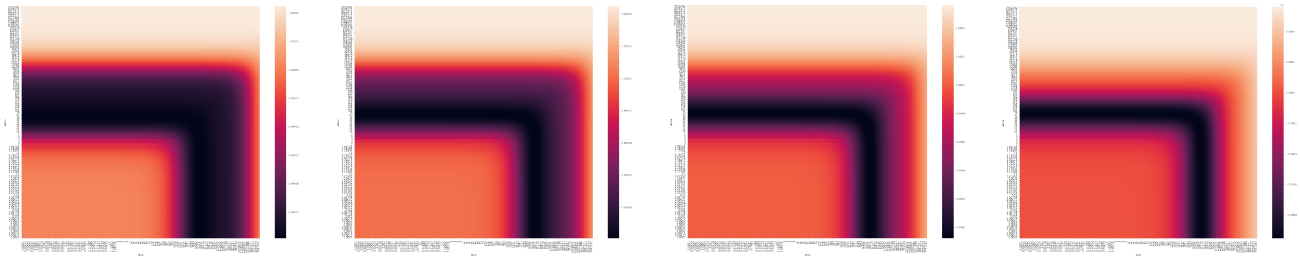

Figure 8. Landscape of squared loss for block-wise $\boldsymbol{W}^V$ with uniform blocks (i.e. optima in Theorem 2), $T = 100, v = 300$. (left-to-right) $\tau = 20$, $\tau = 40$, $\tau = 60$, $\tau = 80$. In each plot, we perform a grid search over $\alpha, \beta \in [10^{-4}, 10^7]$ (both axes use log-scale). Darker color represents lower loss. Across a wide range of τ (compared to T), the loss is lowest when (α, β) is in some corner-shaped region (both α and β are within some intervals whose lower bounds grow with τ).

G. ADDITIONAL EMPIRICAL RESULTS

G.1. Additional results on learned value matrix $\boldsymbol{W}^V$

In Theorem 2 and Figure 2 we have shown that when freezing uniform attention weights and one-hot word embedding, under L_2-regularization, training a single layer transformer on our synthetic topic modeling distribution (Section 3.1) would make its $\boldsymbol{W}^V$ converge to a block-wise pattern that encodes the topic structure.

In the following Figure 9, we additionally show empirical results *without* L_2-regularization, matching our theory in Theorem 4.

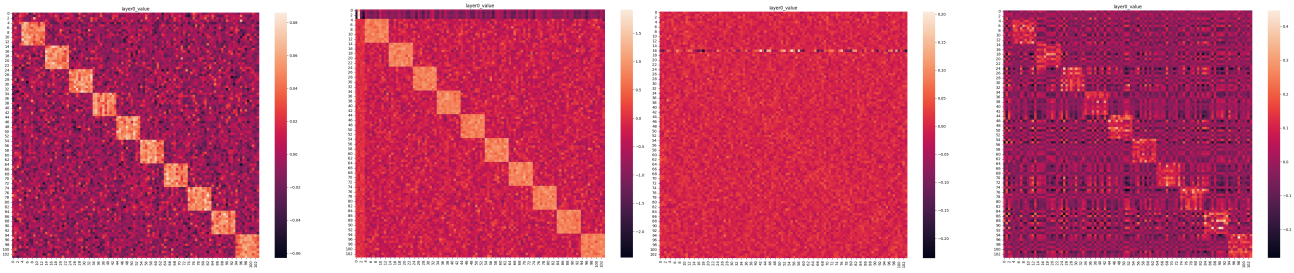

Figure 9. Convergence point of trained $\boldsymbol{W}^V$ (no L_2-regularization) when freezing uniform attention weights and one-hot word embedding. The four plots correspond to different combinations of loss function and optimizer. Top: with cross-entropy loss, Bottom: with squared loss. Left: using SGD optimizer, Right: using Adam optimizer, both with learning rate 0.01. The block-wise pattern verifies our theory in Section 5.2. The 10 blocks correspond to the 10 topics in the data distribution. In particular, in the third figure, the blocks are very weak and not easily visible, but we checked that the mean of the 1000 entries corresponding to the block positions is 0.00552563, which is over 10x the magnitude of the mean of a random subset of 1000 non-block entries (mean -0.00015675332, stdev 0.00060286524).

Complementing our experimental results in Section 6, Figure 10 shows that even when the attention weights $\boldsymbol{W}^K, \boldsymbol{W}^Q$ are *jointly trained* with $\boldsymbol{W}^V$, the model would still approximately converge to the type of block-wise $\boldsymbol{W}^V$ described in our analyses in Section 5.2.

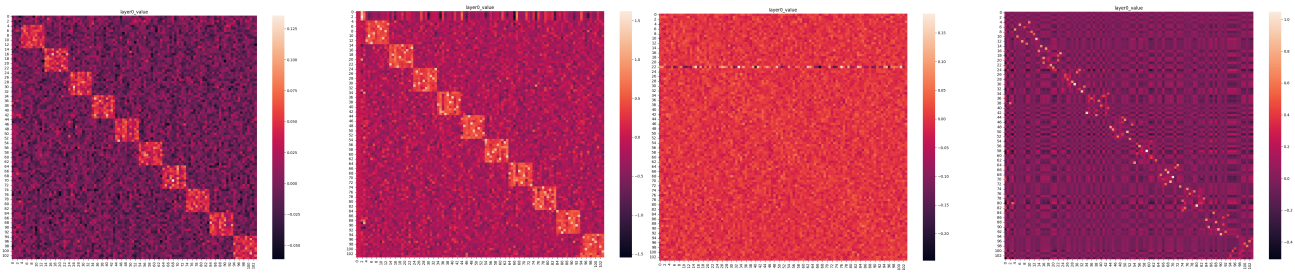

Figure 10. Convergence point of trained $\boldsymbol{W}^V$ when freezing one-hot word embedding but training attention weights. Top: with cross-entropy loss, Bottom: with squared loss. Left: using SGD optimizer, Right: using Adam optimizer, both with learning rate 0.01. The block-wise pattern shows that our analysis in Section 5.2 closely approximates the empirical training dynamics when $\boldsymbol{W}^K, \boldsymbol{W}^Q, \boldsymbol{W}^V$ are trained jointly. The 10 blocks correspond to the 10 topics in the data distribution. In particular, in the third figure, the blocks are very weak and not easily visible, but we checked that the mean of the 1000 entries corresponding to the block positions is 0.006545205, which is over 10x the magnitude of the mean of a random subset of 1000 non-block entries (mean -0.0006503917, stdev 0.0006370574).

G.2. Additional results on learned attention weights

Complementing our experimental results in Section 6, Table 3 shows that when the trained $\boldsymbol{W}^V$ is closer to **uniform within each block**, i.e. on average, each word pays more attention to *different words of the same topic* than to words of *different topics*.

Optimizer and Learning Rate	Avg Same-Word Attention	Avg Same-Topic--Different-Word Attention	Avg Different-Topic Attention
Adam 0.003	0.00759 ± 0.00171	0.0108 ± 0.000657	0.00689 ± 0.000160
Adam 0.01	0.00811 ± 0.000705	0.010 ± 0.000392	0.00707 ± 0.000178
Adam 0.03	0.00453 ± 0.000346	0.0116 ± 0.000460	0.00665 ± 0.000200
SGD 0.01	0.0105	0.0106	0.00673
SGD 0.03	0.0140 ± 0.00158	0.0103 ± 0.000357	0.00641 ± 0.0000239

Table 3. Average attention weights when the model (with one-hot word embeddings) is trained under the cross-entropy loss and $\boldsymbol{W}^V$ converges to a block-wise pattern with closer to uniform blocks. We report mean $\pm$ std. deviation over 3 runs. The row "SGD 0.01" only contains 1 run, and the row "SGD 0.003" is removed, because these models had much higher final train and dev losses than others. For these failed runs, all three types of attention weights have similar averages, a sign that $\boldsymbol{W}^K$ and $\boldsymbol{W}^Q$ did not learn meaningful topical structures. Note that under most settings, *same-word* attention is larger than *same-topic-different-word* attention, which is larger than *different-topic* attention, verifying our conclusion in Theorem 3. The models trained using "Adam 0.03" has larger *same-topic-different-word* attention, which possibly made it unnecessary to rely on *same-word* attention to achieve a low loss, though our theory suggests that increasing *same-word* attention could further reduce the loss.

On the other hand, when the trained $\boldsymbol{W}^V$ is closer to **a diagonal pattern**, the above ordering is partially reversed, Table 4 shows that on average, each word pays the most attention to the *same word* in the document, followed by words of *different topics*, and the least attention to *different words of the same topic*.

Learning Rate	Avg Same-Word Attention	Avg Same-Topic--Different-Word Attention	Avg Different-Topic Attention
0.003	0.0916 ± 0.000901	0.00185 ± 0.000170	0.00256 ± 0.0000332
0.01	0.0918 ± 0.00244	0.00182 ± 0.000474	0.00256 ± 0.000109

Table 4. Average attention weights when the model is trained under the cross-entropy loss with the Adam optimizer and $\boldsymbol{W}^V$ converges to a diagonal pattern. We report mean $\pm$ std. deviation over 7 runs, selected out of 10, by removing the runs in which the diagonal pattern in $\boldsymbol{W}^V$ is not visible or weak. Note that on average, *same-word* attention is larger than *different-topic* attention, which is larger than *same-topic-different-word* attention, verifying our conclusion in Theorem 5.

G.3. Additional details and results on natural language data

In particular, for fair comparison, we should focus on the embedding similarity and attention weights between *different words of the same topic* and *different words of different topics*. (This is because those metrics are less meaningful for a pair of two *same words*, since their embeddings dot product is expected to be larger, which further biases the attention score comparisons.)

Ambiguity filter We also note that, for each word, an LDA model assigns some probability distribution of its topics. To determine whether two words are of the same topic, it is more meaningful if they share a topic in which both words have high likelihood. (By contrast, if two words each has some rarely-used topic that happens to overlap, we intuitively think of them as having different topics.)

To formalize such intuition, we filter out stop tokens, and other tokens that are not central to any topic (determined by the LDA). That is, for each topic t, LDA assigns to it a likelihood p_i for each word w_i in the vocabulary (of size n). We sort these (word, likelihood) pairs by decreasing likelihood:

$$(w_1, p_1), \cdots, (w_n, p_n)$$

then for a pre-defined threshold parameter $\theta \in (0, 1)$ controlling the proportion of words to be assigned to each topic, we only consider the topic t to contain the following words

$$\{w_i : i \leq \theta n\}$$

Debiasing average attention weight Moreover, we note that sentence length may cause a bias in attention weights calculation: intuitively, the average attention weight is the inverse of sentence length, but longer sentences usually contain more topics (and hence a larger proportion of different-topic word pairs). Thus, we expect that the average attention weight between different-topic word pairs are smaller than that between same-topic word pairs, *even for a transformer with random parameters*. (Empirically this bias indeed exists robustly, both on synthetic data and on Wikipedia data.) Therefore, we debias the effect of sentence length on attention weights: for each sentence, while computing the pairwise attention weights among its words, we "normalize the sentence length to 100", that is, we multiply the raw attention weights by sentence length, and then divide the result by 100. In this way, the average attention weight in each sentence is always $\frac{1}{100}$, regardless of the proportion of same-topic and different-topic word pairs. Indeed, as Table 1 and Table 5 show, for a randomly initialized BERT model, after our debiasing, the average same-topic and different-topic attention weights are roughly equal.

Results For a set of pre-trained transformer-based models downloaded from Huggingface (Wolf et al., 2020), we compare the embedding similarity and attention weights between same-topic tokens and different-topic tokens. The topics are determined by fitting an LDA model with 100 topics on a sample of tokenized Wikipedia corpus. We apply the above-mentioned ambiguity filter and debiasing.

- When we further restrict to keeping only one topic for each word (to be consistent with the setting in our theoretical analysis): see Table 1.
- Without the last restriction above: see the following Table 5.

Model	Ambiguity Threshold	Avg embedding Cosine Similarity (Same-topic/Diff-topic)	Avg embedding Dot Product (Same-topic/Diff-topic)	Avg attn weight (Same-topic /Diff-topic)
Bert	0.0005	1.14	1.04	1.23
	0.001	0.97	1.05	1.17
	0.002	0.99	0.93	1.13
Albert	0.0005	4.15	3.06	1.23
	0.001	3.09	3.04	1.17
	0.002	1.54	1.44	1.11
Bart	0.0005	2.51	1.76	1.27
	0.001	1.63	1.12	1.20
	0.002	1.06	0.85	1.11
Electra	0.0005	5.28	3.99	1.70
	0.001	5.56	5.57	1.58
	0.002	6.39	5.61	1.48
Roberta	0.0005	4.39	5.01	1.19
	0.001	5.20	4.25	1.15
	0.002	4.71	4.15	1.12
Bert (randomly initialized)	0.0005	0.99814	0.99957	1.00009
	0.001	0.99820	1.00167	1.00013
	0.002	0.99964	0.99928	0.99978

Table 5. For models pretained on Wikipedia dataset, their token embeddings and attention weights encode topic structure. The different columns are: (1) The "ambiguity threshold", i.e. the number of words per topic, divided by the vocabulary size; each word is only assigned one **or more** topic(s) (2) The average embedding cosine similarity between different words of the *same topic*, divided by that between words of *different topics*. (3) The average embedding dot product between different words of the *same topic*, divided by that between words of *different topics*. (4) The average attention weight between different words of the *same topic*, divided by that between words of *different topics*. (The attention weights are normalized for debiasing, see Appendix G.3). Different rows represent different evaluation settings, controlled by "ambiguity threshold". Note that the avg same-topic embedding similarity and attention weight are mostly greater than the avg diff-topic counterparts (with some exceptions). Allowing multiple topics per word is different from our theoretical setup, so our conclusions in Theorem 1 and Theorem 3 do not cover this setting, though we conjecture that some variants of these theoretical results can be proven using similar approaches to ours.