From Self-Attention to Markov Models: Unveiling the Dynamics of Generative Transformers

M. Emrullah Ildiz [1] Yixiao Huang [1] Yingcong Li [1] Ankit Singh Rawat [2] Samet Oymak [1]

Abstract

Modern language models rely on the transformer architecture and attention mechanism to perform language understanding and text generation. In this work, we study learning a 1-layer self-attention model from a set of prompts and the associated outputs sampled from the model. We first establish a formal link between the self-attention mechanism and Markov models under suitable conditions: Inputting a prompt to the self-attention model samples the output token according to a *context-conditioned Markov Chain* (CCMC). *CCMC* is obtained by weighing the transition matrix of a standard Markov chain according to the sufficient statistics of the prompt/context. Building on this formalism, we develop identifiability/coverage conditions for the data distribution that guarantee consistent estimation of the latent model under a teacher-student setting and establish sample complexity guarantees under IID data. Finally, we study the problem of learning from a single output trajectory generated in response to an initial prompt. We characterize a *winner-takes-all* phenomenon where the generative process of self-attention evolves to sampling from a small set of *winner tokens* that dominate the context window. This provides a mathematical explanation to the tendency of modern LLMs to generate repetitive text.

1. Introduction

The attention mechanism (Vaswani et al., 2017) is a key component of the canonical transformer architecture which underlies the recent advances in language modeling (Radford et al., 2018; 2019; Brown et al., 2020; Chowdhery et al., 2022; Touvron et al., 2023). The self-attention layer allows all tokens within an input sequence to interact with each other. Through these interactions, the transformer assesses the similarities of each token to a given query and composes their value embedding in a non-local fashion.

In this work, we study the mathematical properties of the one-layer self-attention model where the model is trained to predict the next token of an input sequence. The token generation process via the self-attention mechanism is non-trivial because the generation depends on *entire* input sequence. This aspect is crucial for the capabilities of modern *large language models* (LLMs) where the response is conditioned on the user prompt. It is also unlike well-understood topics such as Markov Chains where the model generates the next state based on the current one. Theoretical analysis is further complicated by the fact that the optimization landscape is typically nonconvex. This motivates us to ask:

> **Q:** Can self-attention be formally related to fundamental models such as Markov chains? Can this allow us to study its optimization, approximation, and generalization properties?

Our main contribution is addressing this question by formally mapping the generative process of one self-attention layer to, what we call, *Context-Conditioned Markov Chains (CCMC)* under suitable conditions. In essence, CCMC modifies the transition probabilities of a *base Markov chain* according to the sequence of tokens/states observed so far. Thus, learning a self-attention layer from the (prompt, output) pairs generated by it can be interpreted as learning a Markov chain from its *context-conditioned* transitions.

Concretely, we make the following contributions:

- **CCMC ⇔ Self-attention (Sec 2).** We introduce CCMC and show that it can precisely represent the transition dynamics of self-attention under suitable conditions. Importantly, the optimization of self-attention weights becomes convex, hence tractable via gradient descent, under maximum likelihood estimation i.e., $-\log$ loss.
- **Consistency of learning (Sec 3).** We study the learnability of a self-attention layer where we observe its outputs

[1] University of Michigan, Ann Arbor, USA [2] Google Research, NYC, USA. Correspondence to: M. Emrullah Ildiz <eildiz@umich.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

for a set of input prompts. In practice, this is motivated by the question: *Can we distill the generative capabilities of a language model by collecting its outputs on a set of instructions/prompts?* Through the CCMC connection, we identify necessary and sufficient *coverage conditions* on the prompt distribution that ensures consistent estimation of the underlying model.

- **Sample complexity (Sec 4).** Integrating consistency guarantees with finite sample analysis, we develop generalization guarantees for learning a ground-truth self-attention model from its IID (prompt, output) pairs. We establish a fast statistical rate of $\mathcal{O}(K^2/n)$ where K is the size of the token vocabulary and n is the sample size.
- **Learning from single prompt trajectory (Sec 5).** Going beyond IID samples, we provide theory and experiments on the learnability of self-attention from a *single trajectory* of its autoregressive generation. Our findings reveal a *distribution collapse* phenomenon where the transition dynamics evolve to generate only one or very few tokens while suppressing the other tokens. This also provides an explanation to why modern LLMs tend to generate repetitive sentences after a while (See et al., 2017; Holtzman et al., 2019; Xu et al., 2022). Finally, we study the characteristics of self-attention trajectory, identify novel phase transitions, and shed light on when consistent estimation succeeds or fails.
- **The role of positional encoding (App A).** We augment our theory to incorporate positional encoding (PE). We show that PE enriches CCMC to make transition dynamics adjustable by learnable positional priors.

Overall, we believe that CCMC provides a powerful and rigorous framework to study self-attention and its characteristics. Note that, the token generation process implemented by self-attention is inherently non-Markovian because the next generated token depends on the whole past input prompt/trajectory. Thus, CCMC is also a non-Markovian process, however, it admits a simple representation in terms of a *base Markov chain* and the prompt/trajectory characteristics. CCMC is illustrated in Figure 1 and formally introduced in the next section together with its connection to the self-attention mechanism.

2. Setup: Markov Chain and Self-Attention

Notation. Let $[n]$ denote the set $\{1, \cdots, n\}$ for an integer $n \geq 1$. We use lower-case and upper-case bold letters (e.g., $\boldsymbol{a}$, $\boldsymbol{A}$) to represent vectors and matrices, respectively. a_i denotes the i-th entry of the vector $\boldsymbol{a}$. Let $\boldsymbol{\Pi}_{\mathcal{S}}(\cdot)$ denote the projection operator on a set $\mathcal{S}$, $\mathbf{1}(E)$ denote the indicator function of an event E, and $\mathbb{S}(\cdot) : \mathbb{R}^L \to \mathbb{R}^L$ denote the softmax operation. We use $\lesssim$ and $\gtrsim$ for inequalities that hold up to constant/logarithmic factors.

2.1. Context-Conditioned Markov Chain (CCMC)

Let $\boldsymbol{P} = [\boldsymbol{\pi}_1 \ldots \boldsymbol{\pi}_K] \in \mathbb{R}^{K\times K}$ be the transition matrix associated with a base Markov chain where the i-th column $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iK}) \in \mathbb{R}^K$ captures the transition probabilities from state $i \in [K]$ with entries adding up to 1. Thus, given random state sequence $(x_t)_{t\geq 1}$ drawn according to $\boldsymbol{P}$, we have that $\mathbb{P}(x_{t+1} = j|x_t = i) = \pi_{ij}$.

Now consider the modified transitions for $\boldsymbol{P}$ where transition probabilities are weighted according to a vector $\boldsymbol{m} \in \mathbb{R}^K$ with non-negative entries. Concretely, we consider the following transition model

$$\mathbb{P}_{\boldsymbol{m}}(x_{t+1} = j|x_t = i) = \pi_{ij}^{\boldsymbol{m}} := \frac{m_j \cdot \pi_{ij}}{\boldsymbol{m}^\top \boldsymbol{\pi}_i}. \tag{1}$$

Note that this transition model is still a standard Markov chain with updated transition probabilities $\pi_{ij}^{\boldsymbol{m}} = \frac{m_j \cdot \pi_{ij}}{\boldsymbol{m}^\top \boldsymbol{\pi}_i}$. In contrast, we now introduce the setting where weighting $\boldsymbol{m}$ changes as a function of the input sequence.

Context-conditioned Markov Chain (CCMC). CCMC is a *non-Markovian* transition model derived from a base transition matrix $\boldsymbol{P}$. To proceed, given a state trajectory $X = (x_t)_{t=1}^L \in [K]^L$, let us define $\boldsymbol{m}(X)$ to be the empirical frequencies of individual states where

$$\boldsymbol{m}(X)_k = \frac{|\{t \in [L] \ : \ x_t = k\}|}{L}, \quad \forall k \in [K]. \tag{2}$$

CCMC is obtained by weighting the standard Markov chain transitions according to $\boldsymbol{m}(X)$ determined by X.

Definition 2.1. Let $X = (x_t)_{t=1}^L$ and $\boldsymbol{m} = \boldsymbol{m}(X)$. Given a transition matrix $\boldsymbol{P}$, the associated CCMC transition from state x_L to x_{L+1} is governed by $\boldsymbol{\pi}^X \triangleq \boldsymbol{\pi}_{x_L}^{\boldsymbol{m}(X)} \in \mathbb{R}^K$ defined as

$$\mathbb{P}_{\boldsymbol{P}}(x_{L+1} = j|X) = \pi_j^X := \frac{m_j \cdot \pi_{x_L,j}}{\boldsymbol{m}^\top \boldsymbol{\pi}_{x_L}}. \tag{3}$$

Here, note that the last element x_L of X still serves as the state of the Markov chain; however, transitions are weighted by state frequencies, which can be observed in Figure 1. These frequencies will be evolving as the model keeps generating new states. In the context of language modeling, $X = (x_t)_{t=1}^L$ corresponds to the *prompt* inputted to the model and $(x_t)_{t>L}$ is the model's response: Sections 3 and 4 will explore the learnability of underlying dynamics $\boldsymbol{P}$ from multiple diverse prompts and the corresponding model generations. On the other hand, Section 5 will study learnability from an infinite trajectory generated from a single prompt.

As we shall see, Definition 2.1 captures the dynamics of a 1-layer self-attention model when there are no positional encodings. In Appendix A, we introduce a more general setting where the transition dynamics of CCMC incorporate the positional information of the state trajectory. This

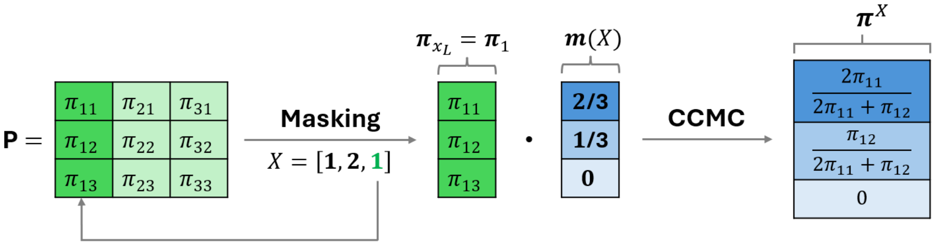

P =
π_{11} π_{21} π_{31}
π_{12} π_{22} π_{32}
π_{13} π_{23} π_{33}
Masking
$X = [1, 2, 1]$
$\pi_{x_L} = \pi_1$
$m(X)$
2/3
1/3
0
CCMC
π^X
$\frac{2\pi_{11}}{2\pi_{11}+\pi_{12}}$
$\frac{\pi_{12}}{2\pi_{11}+\pi_{12}}$
0

Figure 1. Demonstration of Definition 2.1. We provide an example where the vocabulary size $K = 3$ and the input prompt $X = [1, 2, 1]$, which results in a frequency vector $\boldsymbol{m}(X)$. $\boldsymbol{P}$ represents the transition matrix of the base Markov chain.

enriched model will similarly capture self-attention with absolute positional encoding.

2.2. Attention-based Token Generation

We consider a single attention head which admits a sequence of tokens and outputs the next token. Suppose we have a vocabulary of K tokens denoted by $[K]$, which precisely corresponds to the set of states in the (base) Markov chain. To feed tokens to the attention layer, we embed them to get an embedding matrix $\boldsymbol{E} = \begin{bmatrix}\boldsymbol{e}_1 & \dots & \boldsymbol{e}_K\end{bmatrix}^\top \in \mathbb{R}^{K\times d}$, where $\boldsymbol{e}_i \in \mathbb{R}^d$ is the embedding of the i-th token. To feed a prompt $X = (x_i)_{i=1}^L$ to the attention layer, we first obtain its embedding as follows:

$$\boldsymbol{X} = [\boldsymbol{x}_1 \dots \boldsymbol{x}_L]^\top \in \mathbb{R}^{L\times d} \quad \text{where} \quad \boldsymbol{x}_i := \boldsymbol{e}_{x_i}, \forall i \in [L].$$

Throughout, we consistently denote the embedding of a discrete sequence X and token x_i by $\boldsymbol{X}$ and $\boldsymbol{x}_i$, respectively.

Self-attention model. A single-layer self-attention head predicts the next token based on the input prompt X, with the last token x_L forming the query token in the attention layer and playing a distinct role in sampling the next token x_{L+1}. Let us denote the combined key-query weights by a trainable matrix $\boldsymbol{W} \in \mathbb{R}^{d\times d}$, and assume the value weights to be the identity matrix, i.e., $\boldsymbol{V} = \boldsymbol{I}_d$. With these, the self-attention layer $f_{\boldsymbol{W}}$ outputs

$$f_{\boldsymbol{W}}(X) = \boldsymbol{X}^\top \boldsymbol{s}_X \quad \text{where} \quad \boldsymbol{s}_X = \mathbb{S}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{x}_L). \quad \text{(SA)}$$

Here $\mathbb{S}(\cdot)$ is the softmax function and $\boldsymbol{s}_X \in \mathbb{R}^L$ is the softmax probability output associated with X. A useful observation is that $f_{\boldsymbol{W}}(X)$ produces probability-weighted combination of the input tokens. To proceed, let us define a transition matrix $\boldsymbol{P}^{\boldsymbol{W}} \in \mathbb{R}^{K\times K}$ associated with the attention model weights $\boldsymbol{W}$ as follows:

$$\boldsymbol{P}^{\boldsymbol{W}} = [\boldsymbol{\pi}_1 \ \dots \ \boldsymbol{\pi}_K] \quad \text{where} \quad \boldsymbol{\pi}_i = \mathbb{S}(\boldsymbol{E}\boldsymbol{W}\boldsymbol{e}_i). \quad (4)$$

Next, we observe the following identity on self-attention.

Lemma 2.2. *Let (X, y) be an arbitrary pair of (prompt, next token). Define $\boldsymbol{\pi}^X \in \mathbb{R}^K$ based on $\boldsymbol{P}^{\boldsymbol{W}}$ using Definition 2.1. We have that*

$$f_{\boldsymbol{W}}(X) = \boldsymbol{X}^\top \boldsymbol{s}_X = \boldsymbol{E}^\top \boldsymbol{\pi}^X.$$

Lemma 2.2 highlights a fundamental connection between self-attention and CCMC, which we leverage by defining the following sampling. The proof is provided in Appendix B.1.

Sampling-from-softmax. The idea is sampling the next token proportional to its contribution to the output of the self-attention layer. This is equivalent to sampling the next token according to its total probability within the softmax-attention map given by $\boldsymbol{\pi}^X$. Thus, *sampling-from-softmax* with weights $\boldsymbol{W} \in \mathbb{R}^{d\times d}$ is mathematically equivalent to a CCMC with transition dynamics (4).

In what follows, we will introduce and investigate an attention-based next token generation model that implements the *sampling-from-softmax* procedure. Let $\boldsymbol{C} \in \mathbb{R}^{K\times d}$ be the linear prediction head. Following attention output $f_{\boldsymbol{W}}(X)$, we sample the next token from $\boldsymbol{C}f_{\boldsymbol{W}}(X) \in \mathbb{R}^K$. We will utilize the following assumption.

Assumption 2.3. The vocabulary embeddings $(\boldsymbol{e}_k)_{k=1}^K$ are linearly independent and the classifier obeys $\boldsymbol{C}\boldsymbol{E}^\top = \boldsymbol{I}_K$.

This assumption is a slightly stronger version of the weight-tying, where the output and input embedding are the same (Press & Wolf, 2017). In addition to the weight-tying, we apply orthogonalization to the output embedding with the linearly independent conditions so that the output embedding only interacts with the corresponding token. This assumption also requires that the token embeddings are overparameterized and $d \geq K$. Under this assumption, applying Lemma 2.2, we find that $\boldsymbol{C}f_{\boldsymbol{W}}(X) = \boldsymbol{\pi}^X$. Thus, sampling from the classifier output $\boldsymbol{C}f_{\boldsymbol{W}}(X)$ becomes equivalent to *sampling-from-softmax*. While the assumption is rather strong, it will enable us to develop a thorough theoretical understanding of self-attention through the CCMC connection and convex log-likelihood formulation that facilitates consistency and sample complexity analysis.

Bijection between attention and CCMC dynamics. We will next show that, under Assumption 2.3 there is a bijection between weights $\boldsymbol{W}$ and stochastic matrices $\boldsymbol{P}$. This will be established over a $(K-1)\times K$ subspace of $d \times d$ which is the degrees of freedom of the stochastic matrices.

Definition 2.4. Let $\mathcal{S}_{\boldsymbol{E}} \subset \mathbb{R}^{d\times d}$ be the subspace spanned

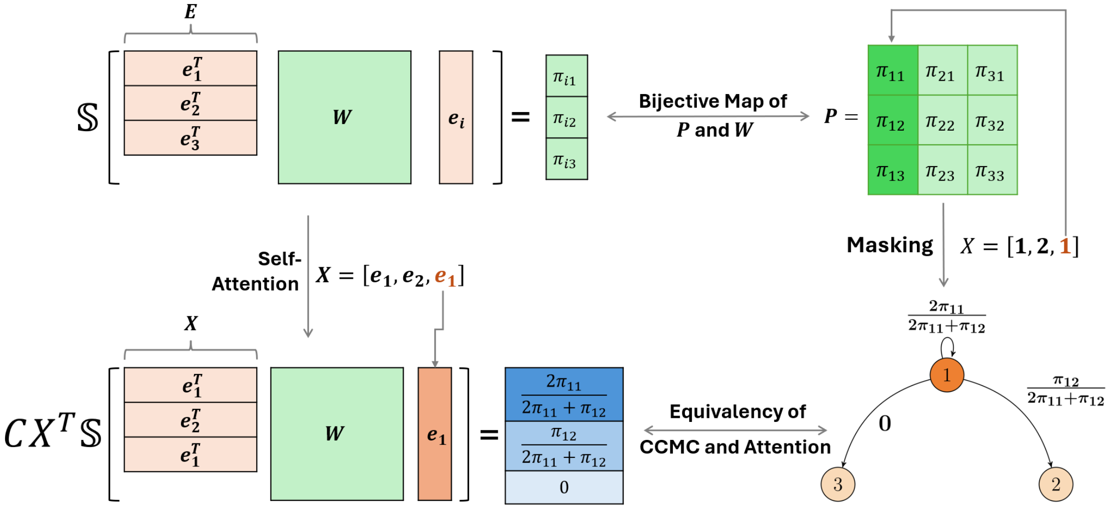

E
e_1^T
e_2^T
e_3^T
$\mathbb{S}$
W
e_i
=
π_{i1}
π_{i2}
π_{i3}
Bijective Map of
P and W
P =
π_{11} π_{21} π_{31}
π_{12} π_{22} π_{32}
π_{13} π_{23} π_{33}
Self-
Attention
$X = [e_1, e_2, e_1]$
Masking
$X = [1, 2, 1]$
X
$CX^T\mathbb{S}$
e_1^T
e_2^T
e_1^T
W
e_1
=
$\frac{2\pi_{11}}{2\pi_{11}+\pi_{12}}$
$\frac{\pi_{12}}{2\pi_{11}+\pi_{12}}$
0
Equivalency of
CCMC and Attention
$\frac{2\pi_{11}}{2\pi_{11}+\pi_{12}}$
1
$\frac{\pi_{12}}{2\pi_{11}+\pi_{12}}$
0
3
2

Figure 2. Illustration of the Equivalency between the Attention and CCMC models. We provide an example where the vocabulary size $K = 3$ and the input prompt is $X = [1, 2, 1]$. The upper figure represents how the token probabilities $\mathbb{S}(\boldsymbol{E}\boldsymbol{W}\boldsymbol{e}_i)$ can be mapped to a base transition matrix $\boldsymbol{P}$. The left-lower figure demonstrates the output of the self-attention given an input prompt $\boldsymbol{X}$. The right-lower figure derives CCMC transitions from this $\boldsymbol{P}$ given the same prompt. The resulting next token probabilities are the same for both of the models. The masking operation is demonstrated in a more detailed way in Figure 1.

by the matrices in $\{(\boldsymbol{e}_i - \boldsymbol{e}_j)\boldsymbol{e}_k^\top : i, j, k \in [K]\}$.

The next lemma states that the projection of $\boldsymbol{W}$ orthogonal to $\mathcal{S}_{\boldsymbol{E}}$ has no impact on the model output and justifies the definition of $\mathcal{S}_{\boldsymbol{E}}$.

Lemma 2.5. *For all $\boldsymbol{W} \in \mathbb{R}^{d\times d}$ and X, we have $f_{\boldsymbol{W}}(X) = f_{\boldsymbol{\Pi}_{\mathcal{S}_{\boldsymbol{E}}}(\boldsymbol{W})}(X)$.*

The proof of this lemma is provided in Appendix B.2. With this, we are ready to establish the equivalency between the CCMC and self-attention dynamics.

Theorem 2.6. *Suppose Assumption 2.3 holds. For any transition matrix $\boldsymbol{P}$ with non-zero entries, there is a unique $\boldsymbol{W} \in \mathcal{S}_{\boldsymbol{E}}$ such that for any prompt $X \in [K]^L$ and next token $y = x_{L+1} \in [K]$,*

$$\mathbb{P}_{\boldsymbol{P}}(y|X) = \boldsymbol{c}_y^\top \boldsymbol{X}^\top \mathbb{S}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{x}_L)$$

where $\mathbf{c}_y$ is the y-th row of the linear prediction head $\boldsymbol{C}$.

The proof of this theorem is provided in Appendix B.3. Note that for any $\boldsymbol{W} \in \mathbb{R}^{d\times d}$, there exists a transition matrix that satisfies the above by Lemma 2.2. As a result, Theorem 2.6 establishes the equivalence between the CCMC model and the self-attention model in the sense that $\mathbb{P}_{\boldsymbol{P}}(y|X)$ matches the output distribution of the self-attention model for any input prompt X. This equivalence can be observed in Figure 2. We will utilize this for learning latent attention models in Sections 3 and 5.

2.3. Cross-Attention Model

We also consider the cross-attention model where the query token $\boldsymbol{x}_L$ is not among the keys for the attention head. Thus, $\boldsymbol{x}_L$ becomes a free variable resulting in a more flexible transition model. The cross-attention is given by

$$\text{Key tokens: } \bar{X} = [x_1 \ \dots \ x_{L-1}]$$
$$f_{\boldsymbol{W}}^{\text{CA}}(X) = \bar{\boldsymbol{X}}^\top \mathbb{S}(\bar{\boldsymbol{X}}\boldsymbol{W}\boldsymbol{x}_L) \in \mathbb{R}^d. \qquad \text{(CA)}$$

The CCMC associated with the cross-attention only slightly differs from Definition 2.1. Now, the transition probabilities are defined with $\boldsymbol{m} = \boldsymbol{m}(\bar{X})$ because the model can only sample from the states/keys contained in $\bar{X}$. This clearly disentangles the *state* $\boldsymbol{x}_L$ of the CCMC and the transition weighting vector $\boldsymbol{m}(\bar{X})$. Specifically, unlike self-attention, this CCMC is not biased towards transitioning to the last token $\boldsymbol{x}_L$ and we can entirely mask out last token in the next transition (as soon as $\boldsymbol{x}_L$ is not contained within $\bar{X}$).

3. Consistent Estimation: When can we learn an attention layer by prompting it?

In this section, our interest is learning a ground-truth attention layer by sampling (prompt, next-token) pairs. Let $\mathcal{D}_{\mathcal{X}}$ denote the distribution of input prompts. We assume $\mathcal{D}_{\mathcal{X}}$ has a finite support and denote its support by Ω, which is a set of prompts. Finally, let $\boldsymbol{W}^{\text{GT}}$ denote the ground-truth attention weights which will be our generative model. Specifically, we will sample the next token $y \in [K]$ according to the model output $\boldsymbol{C}f_{\boldsymbol{W}^{\text{GT}}}(X)$ under Assumption 2.3. Let $\mathcal{D}_{\mathcal{X}\mathcal{Y}}$

be this distribution of (X, y) such that $X \sim \mathcal{D}_{\mathcal{X}}$ and $y \mid X$ is sampled from $\boldsymbol{C} f_{\boldsymbol{W}^{\text{GT}}}(X)$.

Throughout the section, we identify the conditions on the input prompt distribution $\mathcal{D}_{\mathcal{X}}$ and $\boldsymbol{W}^{\text{GT}}$ that guarantee consistent learning of the attention matrix $\boldsymbol{W}^{\text{GT}}$ in the population limit (given infinitely many samples from $\mathcal{D}_{\mathcal{XY}}$). We will investigate the maximum likelihood estimation procedure which is obtained by minimizing the negative log-likelihood. As a result, our estimation procedure corresponds to the following optimization problem.

$$\boldsymbol{W}_\star = \arg\min_{\boldsymbol{W} \in \mathcal{S}_{\boldsymbol{E}}} \mathcal{L}(\boldsymbol{W}) \quad \text{where} \tag{5}$$
$$\mathcal{L}(\boldsymbol{W}) = \mathbb{E}_{(X,y)\sim\mathcal{D}_{\mathcal{XY}}}[-\log(\boldsymbol{c}_y^\top \boldsymbol{X}^\top \mathbb{S}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{x}_L))].$$

Note that the prompt length is not necessarily the same. Here we use $\boldsymbol{x}_L$ to simplify the notation and it presents the last/query token of prompt $\boldsymbol{X}$.

Definition 3.1. The estimator (5) is **consistent** if $\boldsymbol{W}_\star = \boldsymbol{W}^{\text{GT}}$ when data $\mathcal{D}_{\mathcal{XY}}$ is sampled from $\boldsymbol{W}^{\text{GT}}$. Otherwise, the estimation is called inconsistent.

Observe that the optimization in (5) is performed over the subspace $\mathcal{S}_{\boldsymbol{E}}$. As discussed in Lemma 2.5, its orthogonal complement $\mathcal{S}_{\boldsymbol{E}}^\perp$ has no impact on the output of the attention model. Similarly, the gradient of $\mathcal{L}(\boldsymbol{W})$ is zero over $\mathcal{S}_{\boldsymbol{E}}^\perp$, thus $\mathcal{S}_{\boldsymbol{E}}^\perp$ is essentially the *null space* of the problem where token embeddings simply don't interact. For this reason, we make the following assumption.

Assumption 3.2. The ground truth obeys $\boldsymbol{W}^{\text{GT}} \in \mathcal{S}_{\boldsymbol{E}}$.

Learning the ground-truth model is challenging for two reasons: First, through each prompt, we only collect partial observations of the underlying model. It is not clear if these observations can be *stitched* to recover the full model. For instance, if we query an LLM on a subset of domains (e.g., only query about medicine and law), we might not be able to deduce its behavior on another domain (e.g., computer science, math). Second, optimization of self-attention is typically non-convex (Tarzanagh et al., 2023a) and does not result in global convergence. Fortunately, under Assumption 2.3, (5) becomes a convex problem (Li et al., 2024) which we will leverage in our analysis at the end of this section.

Section 2 established the equivalency between the CCMC model and the self-attention model under Assumption 3.2. This enables us to prove the consistency of estimation for the self-attention model through the consistency of estimation for the CCMC model. We define the population estimate of the transition matrix for the CCMC model as follows:

$$\boldsymbol{P}_\star = \arg\min_{\boldsymbol{P}} \mathcal{L}(\boldsymbol{P}) \quad \text{where} \tag{6}$$
$$\mathcal{L}(\boldsymbol{P}) = \mathbb{E}_{(X,y)\sim\mathcal{D}_{\mathcal{XY}}}[-\log(\mathbb{P}_{\boldsymbol{P}}(y|X))].$$

Through Theorem 2.6, there is a mapping between the optimal solution set of (5) and (6).

The consistency conditions on the input prompt distribution and the ground truth variable $\boldsymbol{W}^{\text{GT}}$, or equivalently $\boldsymbol{P}_\star = \boldsymbol{P}^{\text{GT}} \triangleq \boldsymbol{P}^{\boldsymbol{W}^{\text{GT}}}$, are related to *how well input prompts* Ω *cover the pairwise token relations*. To characterize this, for each last token (i.e., query token), we define an undirected co-occurrence graph as follows.

Definition 3.3. Let $\Omega_k \subset \Omega$ be the set of input prompts whose last tokens are equal to k for all $k \in [K]$. Define the co-occurrence graph $\mathcal{G}^{(k)}$ with K vertices as follows: There is an edge between i-th and j-th nodes in $\mathcal{G}^{(k)}$ iff there is a prompt $X \in \Omega_k$ such that its *key tokens* include both i and j. Here, *key tokens* are all tokens for self-attention and all but last tokens for cross-attention (i.e., $\bar{X}$ in (CA)).

A simple example highlighting the difference between self-attention and cross-attention is given in Figure 3. Based on this graph, the consistency of estimation is then translated to the connectivity of the graphs $\mathcal{G}^{(k)}$. This is summarized in the following theorem.

Theorem 3.4. *Let $\boldsymbol{P}^{GT}$ be a transition matrix with non-zero entries. Let $(\mathcal{G}^{(k)})_{k=1}^K$ be the co-occurrence graphs based on the input distribution $\mathcal{D}_{\mathcal{X}}$. Then, the estimation $\boldsymbol{P}_\star$ in (6) is consistent iff $\mathcal{G}^{(k)}$ is a connected graph for all $k \in [K]$.*

The proof of Theorem 3.4 is provided in Appendix C. This appendix also addresses the case where $\boldsymbol{P}^{\text{GT}}$ contains zero transition probabilities. The main proof idea is that, optimizing the log-likelihood of a specific prompt results in estimating the local Markov chain transitions over the tokens within that prompt. When two prompts are connected and optimized together, the optimization merges and expands these local chains.

The connectivity of the graph is essentially a *coverage condition* that ensures that prompts can fully sense the underlying Markov chain. For the self-attention model, this condition simplifies quite a bit.

Observation 1. For self-attention model, $\mathcal{G}^{(k)}$ is connected iff all tokens in $[K]$ appear at least in one prompt within Ω_k.

This observation follows from the fact that, for self-attention, the last token is always within the keys of the input prompt. Thus, all prompts within Ω_k overlap at the last token $x_L = k$. Thus, the node k naturally connects all tokens that appear within the prompts (at least once). Note that it is possible that some tokens do not co-occur at the same input sequence. In that case, Theorem 3.4 can be extended to the consistency of estimation for the observable Markov Chain as follows:

Corollary 3.5. *Given $k \in [K]$, let $\mathcal{C}_k \subset [K]$ be the set of all tokens that appear within some training prompt $X \in \Omega_k$ where X ends with token k. Let $\boldsymbol{P}^*$ be the transition model learned by the self-attention trained on $X \sim \mathcal{D}_{\mathcal{X}}$ with labels*

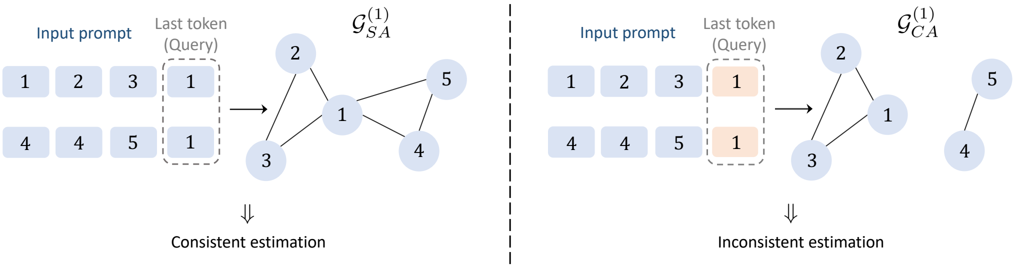


Figure 3. Illustration of co-occurrence graphs for the self-attention (**Left**) and cross-attention (**Right**) models. We fit the same input examples where the only difference is the use of self- vs cross-attention (which includes vs excludes the query token '1' from the list of key tokens). Following Theorem 3.4, in the cross-attention setting, where the token '1' is not contained in the prompt, the co-occurrence graph becomes disconnected, resulting in inconsistent estimation. In contrast, the estimation is consistent for the self-attention model since both inputs share the same query token within their key tokens.

sampled from the ground-truth model $\boldsymbol{P}^{GT}$. For any $k \in [K]$, we have that

$$\boldsymbol{\pi}^*_{k,\mathcal{C}_k} = \boldsymbol{\pi}^{GT}_{k,\mathcal{C}_k}.$$

Here $\boldsymbol{\pi}_{k,\mathcal{C}}$ denotes the probability distribution induced by normalizing the entries of $\boldsymbol{\pi}_k$ over the set $\mathcal{C}$.

The proof of Corollary 3.5 is provided in Appendix C.

The situation is more intricate for cross-attention – where the query token is distinct from keys – because there is no immediate connectivity between prompts. The difference between these two models are illustrated in Figure 3. For this reason, our theory on cross-attention is strictly more general than self-attention and its connection to Markov chain is of broader interest.

Our next result is an application of Theorems 2.6 and 3.4 and states the result for self-attention problem (5). It equivalently applies to the cross-attention variation of (5).

Corollary 3.6. *Let $\boldsymbol{W}^{GT}$ be the attention model underlying $\mathcal{D}_{\mathcal{X}\mathcal{Y}}$ for either the self-attention model or the cross-attention model and suppose Assumptions 2.3 and 3.2 hold. Then, $\boldsymbol{W}^{GT} = \boldsymbol{W}_\star$ in (5) iff all $\mathcal{G}^{(k)}$'s are connected.*

3.1. Gradient-based Optimization of Attention Weights

An important feature of our problem formulation (5) is its convexity, which was observed by (Li et al., 2024). The convexity arises from the fact that, we directly feed the attention probabilities to log-likelihood which results in a LogSumExp function. We next state the following stronger lemma which follows from Lemma 9 of (Li et al., 2024). A detailed discussion is provided in Appendix C.1.

Lemma 3.7. *Suppose Assumption 2.3 holds. If $\mathcal{G}^{(k)}$ is a connected graph for every $k \in [K]$, then $\mathcal{L}(\boldsymbol{W})$ of (5) is strictly convex over $\mathcal{S}_{\boldsymbol{E}}$ and $\boldsymbol{W}^{GT}$ is the unique finite solution.*

For a strictly convex function, we know that, when the minima is finite, it is unique. Therefore, Lemma 3.7 guarantees the gradient-based learnability of the ground-truth weights $\boldsymbol{W}^{\text{GT}}$ as follows.

Corollary 3.8. *Set $\boldsymbol{W}_0 = 0$ and run gradient iterations $\boldsymbol{W}_{t+1} \leftarrow \boldsymbol{W}_t - \eta\nabla\mathcal{L}(\boldsymbol{W}_t)$ for $t \geq 0$ with a suitable learning rate $\eta > 0$. If all $\mathcal{G}^{(k)}$'s are connected, then $\lim_{t\to\infty} \boldsymbol{W}_t = \boldsymbol{W}^{GT}$.*

4. Guarantees on Finite Sample Learning

Following the setting of Section 3, we sample a training dataset $\mathcal{T} = \{(X_i, y_i)\}_{i=1}^n$ from $\mathcal{D}_{\mathcal{X}\mathcal{Y}}$. In this section, we establish a sample complexity guarantee on the difference between $\|\hat{\boldsymbol{W}} - \boldsymbol{W}^{\text{GT}}\|_F$ where $\hat{\boldsymbol{W}}$ is trained on $\mathcal{T}$.

ERM problem. Given a training dataset $\mathcal{T}$, we consider the ERM problem with the following objective:

$$\hat{\boldsymbol{W}}_n = \arg\min_{\boldsymbol{W}\in\mathcal{S}_{\boldsymbol{E}}} \widehat{\mathcal{L}}_n(\boldsymbol{W}) \quad \text{where} \tag{7}$$

$$\widehat{\mathcal{L}}_n(\boldsymbol{W}) = \frac{1}{n}\sum_{i=1}^{n} -\log(\boldsymbol{c}_{y_i}^\top \boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{X}_i \boldsymbol{W} \boldsymbol{x}_{i,L_i})). \tag{8}$$

Our main aim in this section is to establish a sample complexity guarantee on $\|\hat{\boldsymbol{W}} - \boldsymbol{W}^{\text{GT}}\|_F$. We leverage the findings of Section 3 to achieve our aim with the following assumption:

Assumption 4.1. Recall the co-occurrence graphs in Definition 3.3. We assume that the co-occurrence graphs $(\mathcal{G}^{(k)})_{k=1}^K$ constructed from $\mathcal{D}_{\mathcal{X}}$ are connected.

We prove the following theorem, which will provide finite sample complexity guarantees for the loss function:

Theorem 4.2. *Suppose Assumptions 2.3 and 4.1 hold. Let $R_0 > 0$ be a finite constant based on the structure of $\boldsymbol{W}^{GT}$ and $\mathcal{D}_{\mathcal{X}}$. Then, if $n \geq R_0 K^2$, with probability at least $1 - 2\delta$*

$$\mathcal{L}(\hat{\boldsymbol{W}}_n) - \mathcal{L}(\boldsymbol{W}_\star) \lesssim \frac{K^2 \log \frac{n}{K\delta}}{n}.$$

The proof is provided in Appendix D.2. We apply the local covering arguments to achieve sample complexity guarantees by concentration inequalities. Using Lemma 3.7, we prove that $\hat{\boldsymbol{W}}_n$ is inside a local ball with sufficient samples and we achieve the fast rate $1/n$ with the smoothness of the loss function similar to (Bartlett et al., 2005) and (Srebro et al., 2010).

Now, we are ready to share our main contribution in this section with the following corollary:

Corollary 4.3. *Consider the setting in Theorem 4.2 and suppose Assumptions 2.3, 3.2, and 4.1 hold. Then, if* $n \geq R_0 K^2$*, with probability at least* $1 - 2\delta$

$$\|\hat{\boldsymbol{W}}_n - \boldsymbol{W}^{\mathrm{GT}}\|_F^2 \lesssim \frac{K^2 \log \frac{n}{K\delta}}{n}. \tag{9}$$

The proof of this corollary is provided in Appendix D.3 and follows from Lemma 3.7 and Theorem 4.2.

5. Learning from a Single Trajectory

In this section, we ask: Can we learn a ground-truth self-attention model by querying it once and collecting its output trajectory? The question of single-trajectory learning is fundamental for two reasons: First, modern language models train all tokens in parallel, that is we fit multiple next tokens per sequence. Secondly, learning from a single trajectory is inherently challenging due to dependent data and has been subject of intense research in reinforcement learning and control. Here, we initiate the statistical study of single trajectory learning for attention models.

Setting: We first describe the single trajectory sampling. Suppose we are given a ground truth attention matrix $\boldsymbol{W}^{\mathrm{GT}}$ and an initial prompt X_1 of length L. We feed X_1 to sample the next token y_1 and auto-regressively feed back each next token to sample a length n output trajectory. Overall, we obtain the training dataset $\mathcal{T} = \{(X_i, y_i)\}_{i=1}^n$ where $X_i = [X_1 \; y_1 \; \dots \; y_{i-1}]$. This sampling is done according to our self-attention model under Assumption 2.3. To proceed, we optimize the likelihood according to (7) and obtain $\hat{\boldsymbol{W}}_n$.

The consistency of estimation for the single trajectory sampling is defined as follows.

Definition 5.1. Recall the estimation $\hat{\boldsymbol{W}}_n$ of $\boldsymbol{W}^{\mathrm{GT}}$ in (7). The estimation $\hat{\boldsymbol{W}}_n$ is called consistent if and only if

$$\mathbb{P}\left(\lim_{n\to\infty} \hat{\boldsymbol{W}}_n = \boldsymbol{W}^{\mathrm{GT}}\right) = 1.$$

Let $\mathcal{T} \subset [K]$ be the set of tokens that appear in X_1. Recall that, our attention model samples the next token from the input prompt, thus, all generated tokens will be within $\mathcal{T}$. Thus, from a single trajectory, we can at most learn the local Markov chain induced by $\mathcal{T}$ and consistency is only possible if X_1 contains all $[K]$ tokens. Thus, we assume $\mathcal{T} = [K]$ and X_1 contains the full vocabulary going forward.

In what follows, we study two critical behaviors:

- **(Q1)** How does the distribution of generated tokens y_i evolve as a function of the time index i?
- **(Q2)** When is consistent estimation of the underlying attention model $\boldsymbol{W}^{\mathrm{GT}}$ possible?

5.1. Empirical Investigation

We first describe our experiments which elucidate why these questions are interesting and provide a strong motivation for the theory. Note that we have an equivalency between the attention models and the CCMC model, which is true for any sampling method. Throughout this section, we discuss the consistency of $\boldsymbol{P}^{\mathrm{GT}}$, which directly implies the consistency of $\boldsymbol{W}^{\mathrm{GT}}$.

The *Distribution Collapse* phenomenon. To gain more motivation, we randomly initialize a one-layer self-attention model and generate a single trajectory with a length of 500 starting from initial prompt $X_1 = [6]$. We track the evolution of token frequency as shown in the middle of Figure 4. As illustrated in the right side of Figure 4, when the generation time increases, the diversity of output tokens is greatly reduced, which eventually collapses to a singleton. This arises due to the *self-reinforcement* of majority tokens within the trajectory. This phenomenon also corresponds to the repetition problem found in text generation by language models, as demonstrated in the left side of Figure 4.

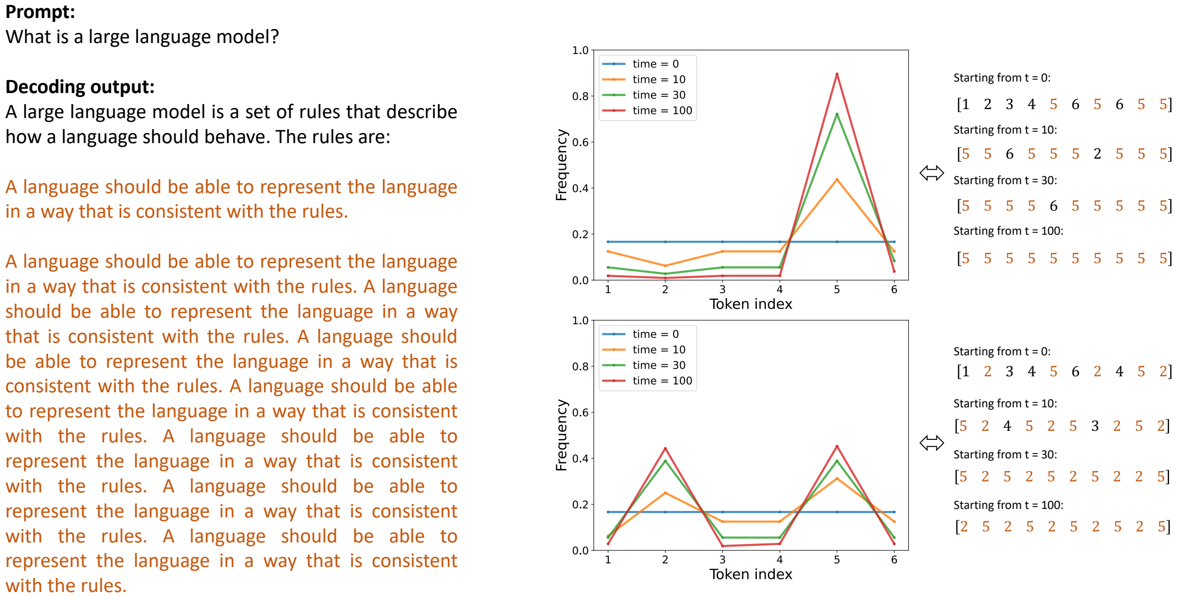
5.2. Theoretical Study of Single Trajectory Sampling

To learn the underlying self-attention model from a single trajectory, we must visit each token/state infinitely many times. Otherwise, we could not learn the probability transitions from that last token choice. Let $S_{k,n}$ be the number of occurrences of token k within X_n. Our first result shows that each token is guaranteed to be visited infinitely many times as the trajectory grows.

Lemma 5.2. *Let* $\boldsymbol{P}^{\mathrm{GT}}$ *be a transition matrix with non-zero entries. We have that* $\mathbb{P}(\lim_{n\to\infty} S_{k,n} = \infty) = 1$ *for all* $k \in [K]$.

The proof of this lemma can be found in Appendix E and follows from an application of the Borel-Cantelli lemma.

This lemma is insightful in light of Figure 4: Even if the distribution of the generated tokens collapses to a singleton, any token within the input prompt will keep appearing albeit with potentially vanishing frequencies. Next, we discuss the condition when the distribution collapse will happen in response to Figure 4. To characterize the condition for the distribution collapse, we consider a CCMC model with $K = 2$. Suppose the ground-truth transition matrix $\boldsymbol{P}^{\mathrm{GT}} = \begin{bmatrix} 1-p & 1-p \\ p & p \end{bmatrix}$. Without loss of generality, assume $p \leq$



Prompt:
What is a large language model?

Decoding output:
A large language model is a set of rules that describe how a language should behave. The rules are:

A language should be able to represent the language in a way that is consistent with the rules.

A language should be able to represent the language in a way that is consistent with the rules. A language should be able to represent the language in a way that is consistent with the rules. A language should be able to represent the language in a way that is consistent with the rules. A language should be able to represent the language in a way that is consistent with the rules. A language should be able to represent the language in a way that is consistent with the rules. A language should be able to represent the language in a way that is consistent with the rules.

Figure 4. Demonstration of Distribution Collapse/Repetition. **Left:** An example query where the GPT-2 response quickly degenerates into repetition. **Middle:** We generate two single self-attention trajectories with a vocabulary containing $K = 6$ tokens and plot the empirical token frequencies (for each token in the vocabulary). The upper figure is generated using a randomly initialized transition matrix, while the lower one is generated using the same transition matrix as the upper one except that the diagonal entries are set to 0, enforcing that the probability of query token $i \rightarrow$ next token i is 0. The frequency is calculated as the ratio of token occurrences to the sequence length at that time. **Right:** Trajectory snapshots with a 10-token window from time index i revealing that token 5 (upper) / tokens 2 and 5 (lower) dominate the trajectory. The lower right is dominated by two tokens because a single token cannot self-reinforce due to zero diagonals.

$1/2$. For brevity, we define the *weak token* as the token with a smaller transition probability p, which is Token 2 in our setting. The next result bounds the frequency of the weak token as the trajectory grows.

Lemma 5.3 (Distribution collapse)**.** *Consider the CCMC model with $K = 2$ defined in Section 5.1. Suppose that $\boldsymbol{X}_1$ includes all vocabulary at least once. Recall that $\boldsymbol{m}(X_t)$ denotes the empirical frequency of individual states where X_t is the state trajectory at time t. For any $t > t_0$ with a sufficiently large t_0, we have $\mathbb{E}[\boldsymbol{m}(X_t)_2] < t^{-q}$ where $q = 1 - p/(1-p)$. Furthermore, when $p < 1/2$,*

$$\lim_{t \to \infty} \mathbb{E}\left[\frac{\boldsymbol{m}(X_t)_2}{\boldsymbol{m}(X_t)_1}\right] = 0.$$

The proof is provided in Appendix E. Lemma 5.3 states that token 1 will dominate the trajectory when $p < 1/2$. As a result, the distribution collapse will happen as long as $p \neq 1/2$ due to the symmetry. So far, we show that all tokens are visited infinitely many but their frequencies can vanish. On the other hand, to learn the underlying transition matrix, we should also visit each transition (from state/token i to state/token j) infinitely many times (otherwise the estimated p would be 0). To study this question, we ask: *Will self-attention visit the transition from the weak token to itself forever?* Figure 5 shows the number of weak→weak transitions for varying p choices. We observe that this number grows super-logarithmic in trajectory length when p exceeds $1/3$ and it is sub-logarithmic when p is smaller than $1/3$. We argue that this sub-logarithmic (very slow) growth is actually an indicator of the fact that there are actually finitely many weak→weak throughout the trajectory, which would in turn make estimation of the second column of $\boldsymbol{P}^{\mathrm{GT}}$ inconsistent.

To justify this, we utilize our theory to study the growth of weak→weak transitions (albeit non-rigorously). Lemma 5.3 shows that the expected density of the weak token is t^{-q} throughout the trajectory. Let us treat this expectation as the true weak token probability at time t. Next, since the trajectory contains only $O(t^{-q})$ fraction weak tokens, due to the CCMC model, the chance of transition to a weak token (from any token) is $O(t^{-q})$. Combining these, we find that $\mathbb{P}(\text{weak} \rightarrow \text{weak}) = \mathbb{P}(\text{weak}|\text{weak})\mathbb{P}(\text{weak}) \propto t^{-2q}$. With this estimate at hand, we can use Borel-Cantelli to study finiteness of weak→weak transitions. Specifically $\int_{t=1}^{\infty} t^{-2q}$ is finite when $q \geq 1/2$ and infinite when $q < 1/2$. This translates to $p \leq 1/3$ and $p > 1/3$ respectively and remarkably coincides with the sub/super-logarithmic growth observed in Figure 5.

6. Related Work

Theoretical treatment of attention models. Yun et al. (2020); Edelman et al. (2022); Fu et al. (2023); Baldi & Vershynin (2023) focused on expressive power or inductive biases of attention-based models. Jelassi et al. (2022);

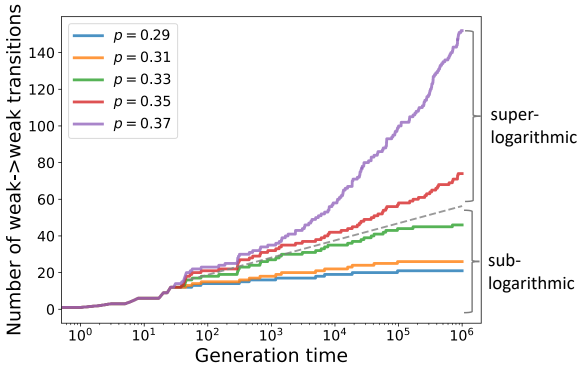


Figure 5. Growth of weak to weak transition

Li et al. (2023a); Oymak et al. (2023); Tarzanagh et al. (2023b;a) studied optimization and generalization dynamics of simplified attention models for supervised classification tasks. Tian et al. (2023); Li et al. (2024) explored the training dynamics of such models for the next-token prediction task. To the best of our knowledge, we are the first ones to establish the connection between (context-conditioned) Markov chains and self-attention models and leverage that to establish rigorous guarantees for learning the models. Although not directly related to this work, there is also a growing body of literature on the theoretical study of in-context learning (Xie et al., 2022; Garg et al., 2022; Li et al., 2023b; Akyürek et al., 2023; Von Oswald et al., 2023).

Learning Markov chains. The problem of estimating the transition matrix of a Markov chain from a single trajectory generated by the chain is a classical problem (Billingsley, 1961). There is a large literature on estimating the transition matrix under some structural constraints on the transition matrix such as low-rank assumption (Zhang & Wang, 2019; Shah et al., 2020; Stojanovic et al., 2023; Bi et al., 2023; Li et al., 2018; Zhu et al., 2022). Interestingly, Stojanovic et al. (2023) also considers matrix estimation from multiple transitions observed from IID sampled states. Note that the text generation from attention models is not quite Markovian due to the context-dependent masking of the base Markov chain in CCMC. Furthermore, we are interested in recovering the model weights $\boldsymbol{W}$ instead of the transition probabilities. A concurrent work (Makkuva et al., 2024) also explores the connection between a 1-layer attention model and Markov chains. The authors aim to learn a standard Markov chain using 1-layer self-attention whereas we show that self-attention is a non-Markovian model and we construct a general mapping between self-attention and a modified Markov chain dynamics.

Shortcomings in neural text generation. Multiple works (see, e.g., See et al., 2017; Holtzman et al., 2019; Welleck et al., 2020; Xu et al., 2022) have explored various issues with the language model generated texts, especially repetitive nature of such texts. Xu et al. (2022) argue that the self-enforcing behavior of language models leads to repetitions. This aligns with our formal analysis of self-attention models using CCMC. Several studies offer training-based solutions to mitigate repetition (Xu et al., 2022; Welleck et al., 2019; Lin et al., 2021), while others modify the decoding process (Fan et al., 2018; Holtzman et al., 2019; Welleck et al., 2020). Rather than proposing a new solution, our work rigorously characterizes the conditions under which repetition becomes inevitable in self-attention models.

Fu et al. (2021) analyzed repetition in a text generated by Markov models, attributing it to high-inflow words. Our work diverges significantly. Instead of assuming a Markovian process, we establish an equivalence between self-attention mechanisms and context-conditioned Markov chains. This leads to the non-Markovian generation, where the entire context influences the next token. Furthermore, our theoretical analysis extends beyond repetition, using this CCMC equivalence to explore the learnability of self-attention models from generated data.

In Appendix F, we share further related work on reinforcement learning and data-driven control.

7. Discussion

In this work, we have studied theoretical properties of the self-attention layer by formally linking its dynamics to (context-conditioned) Markov chains. Through this connection, we identify when a ground-truth self-attention layer is learnable by observing its output tokens. We develop consistency and finite sample learning guarantees for multiple prompts as well as for single trajectory learning, which reveal novel insights into the self-attention mechanism (such as prompt coverage conditions and distribution collapse in the learning from a single trajectory setup).

An important future direction is relaxing Assumption 2.3 to more general and realistic conditions. An initial way to relax this assumption is to assume that $\boldsymbol{C}\boldsymbol{E}^\top$ is equal to a column stochastic matrix. More broadly, instead of linear classifier $\boldsymbol{C}$, it is possible to incorporate a Multi-Layer Perceptron (MLP) into the model. However, these relaxations may cause the loss of convexity but provide a deeper understanding of the self-attention layer. In addition, it is also interesting to study the case of $d \ll K$, which is closely related to the low-rank adaption (LoRA) of the attention matrix.

Other possible future directions are (1) studying the multi-layer attention models and their connection to hierarchical Markov models and representation learning, and (2) analysis of the impact of the End-Of-Sequence (EOS) token, which is utilized to terminate the generation of outputs in modern language models.

Acknowledgements

This work was supported in part by the National Science Foundation grants CCF-2046816, CCF-2212426, CCF-2403075, the Office of Naval Research award N000142412289, UMich's MIDAS PODS program, Open Philanthropy, and an Adobe Data Science Research award. The authors thank Vijay Subramanian for the helpful discussion.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Akyürek, E., Schuurmans, D., Andreas, J., Ma, T., and Zhou, D. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=0g0X4H8yN4I.

Baldi, P. and Vershynin, R. The quarks of attention: Structure and capacity of neural attention building blocks. *Artificial Intelligence*, 319:103901, 2023. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2023.103901. URL https://www.sciencedirect.com/science/article/pii/S0004370223000474.

Bartlett, P. L., Bousquet, O., and Mendelson, S. Local rademacher complexities. *The Annals of Statistics*, 33 (4), August 2005. ISSN 0090-5364. doi: 10.1214/009053605000000282. URL http://dx.doi.org/10.1214/009053605000000282.

Bi, S., Yin, Z., and Weng, Y. A low-rank spectral method for learning markov models. *Optimization Letters*, 17(1): 143–162, 2023.

Billingsley, P. Statistical methods in markov chains. *The annals of mathematical statistics*, pp. 12–40, 1961.

Block, A., Simchowitz, M., and Tedrake, R. Smoothed online learning for prediction in piecewise affine systems. *arXiv preprint arXiv:2301.11187*, 2023.

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_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Chen, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning, 2019.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Dean, S., Mania, H., Matni, N., Recht, B., and Tu, S. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, 20(4):633–679, 2020.

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.

Fan, A., Lewis, M., and Dauphin, Y. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*, 2018.

Foster, D., Sarkar, T., and Rakhlin, A. Learning nonlinear dynamical systems from a single trajectory. In *Learning for Dynamics and Control*, pp. 851–861. PMLR, 2020.

Foster, D. J., Krishnamurthy, A., Simchi-Levi, D., and Xu, Y. Offline reinforcement learning: Fundamental barriers for value function approximation, 2022.

Fu, H., Guo, T., Bai, Y., and Mei, S. What can a single attention layer learn? a study through the random features lens. *arXiv preprint arXiv:2307.11353*, 2023.

Fu, Z., Lam, W., So, A. M.-C., and Shi, B. A theoretical analysis of the repetition problem in text generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 12848–12856, 2021.

Garg, S., Tsipras, D., Liang, P. S., and Valiant, G. What can transformers learn in-context? a case study of simple function classes. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 30583–30598. Curran Associates, Inc.,

2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c529dba08a146ea8d6cf715ae8930cbe-Paper-Conference.pdf.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 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.

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline rl?, 2022.

Kuznetsov, V. and Mohri, M. Generalization bounds for time series prediction with non-stationary processes. In *International conference on algorithmic learning theory*. Springer, 2014.

Kuznetsov, V. and Mohri, M. Time series prediction and online learning. In *Conference on Learning Theory*, pp. 1190–1213. PMLR, 2016.

Li, H., Wang, M., Liu, S., and Chen, P.-Y. A theoretical understanding of shallow vision transformers: Learning, generalization, and sample complexity. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=jClGv3Qjhb.

Li, X., Wang, M., and Zhang, A. Estimation of markov chain via rank-constrained likelihood. In *International Conference on Machine Learning*, pp. 3033–3042. PMLR, 2018.

Li, Y., Ildiz, M. E., Papailiopoulos, D., and Oymak, S. Transformers as algorithms: Generalization and stability in in-context learning. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19565–19594. PMLR, 23–29 Jul 2023b. URL https://proceedings.mlr.press/v202/li23l.html.

Li, Y., Huang, Y., Ildiz, M. E., Rawat, A. S., and Oymak, S. Mechanics of next token prediction with self-attention. *In International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024.

Lin, X., Han, S., and Joty, S. Straight to the gradient: Learning to use novel tokens for neural text generation. In *International Conference on Machine Learning*, pp. 6642–6653. PMLR, 2021.

Makkuva, A. V., Bondaschi, M., Girish, A., Nagle, A., Jaggi, M., Kim, H., and Gastpar, M. Attention with markov: A framework for principled analysis of transformers via markov chains, 2024.

Mania, H., Jordan, M. I., and Recht, B. Active learning for nonlinear system identification with guarantees. *arXiv preprint arXiv:2006.10277*, 2020.

Matni, N. and Tu, S. A tutorial on concentration bounds for system identification. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 3741–3749. IEEE, 2019.

Mohri, M. and Rostamizadeh, A. Rademacher complexity bounds for non-iid processes. *Advances in Neural Information Processing Systems*, 21, 2008.

Oymak, S. Stochastic gradient descent learns state equations with nonlinear activations. In *conference on Learning Theory*, pp. 2551–2579. PMLR, 2019.

Oymak, S. and Ozay, N. Non-asymptotic identification of lti systems from a single trajectory. In *2019 American control conference (ACC)*, pp. 5655–5661. IEEE, 2019.

Oymak, S. and Ozay, N. Revisiting ho–kalman-based system identification: Robustness and finite-sample analysis. *IEEE Transactions on Automatic Control*, 67(4):1914–1928, 2021.

Oymak, S., Rawat, A. S., Soltanolkotabi, M., and Thrampoulidis, C. On the role of attention in prompt-tuning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 26724–26768. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/oymak23a.html.

Press, O. and Wolf, L. Using the output embedding to improve language models, 2017.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. Improving language understanding by generative pre-training. *OpenAI blog*, 2018.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. Bridging offline reinforcement learning and imitation learning: A tale of pessimism, 2023.

Sarkar, T. and Rakhlin, A. Near optimal finite time identification of arbitrary linear dynamical systems. In *International Conference on Machine Learning*, pp. 5610–5618. PMLR, 2019.

Sattar, Y. and Oymak, S. Non-asymptotic and accurate learning of nonlinear dynamical systems. *The Journal of Machine Learning Research*, 23(1):6248–6296, 2022.

See, A., Liu, P. J., and Manning, C. D. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*, 2017.

Shah, D., Song, D., Xu, Z., and Yang, Y. Sample efficient reinforcement learning via low-rank matrix estimation. *Advances in Neural Information Processing Systems*, 33: 12092–12103, 2020.

Simchowitz, M., Mania, H., Tu, S., Jordan, M. I., and Recht, B. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pp. 439–473. PMLR, 2018.

Srebro, N., Sridharan, K., and Tewari, A. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper_files/paper/2010/file/76cf99d3614e23eabab16fb27e944bf9-Paper.pdf.

Stojanovic, S., Jedra, Y., and Proutiere, A. Spectral entrywise matrix estimation for low-rank reinforcement learning. *arXiv preprint arXiv:2310.06793*, 2023.

Sun, Y., Oymak, S., and Fazel, M. Finite sample identification of low-order lti systems via nuclear norm regularization. *IEEE Open Journal of Control Systems*, 1:237–254, 2022.

Tarzanagh, D. A., Li, Y., Thrampoulidis, C., and Oymak, S. Transformers as support vector machines. *arXiv preprint arXiv:2308.16898*, 2023a.

Tarzanagh, D. A., Li, Y., Zhang, X., and Oymak, S. Max-margin token selection in attention mechanism. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b.

Tian, Y., Wang, Y., Chen, B., and Du, S. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. *arXiv preprint arXiv:2305.16380*, 2023.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Tsiamis, A., Ziemann, I., Matni, N., and Pappas, G. J. Statistical learning theory for control: A finite sample perspective. *arXiv preprint arXiv:2209.05423*, 2022.

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_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., and Vladymyrov, M. Transformers learn in-context by gradient descent. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35151–35174. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/von-oswald23a.html.

Welleck, S., Kulikov, I., Roller, S., Dinan, E., Cho, K., and Weston, J. Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*, 2019.

Welleck, S., Kulikov, I., Kim, J., Pang, R. Y., and Cho, K. Consistency of a recurrent language model with respect to incomplete decoding. *arXiv preprint arXiv:2002.02492*, 2020.

Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RdJVFCHjUMI.

Xie, T. and Jiang, N. Q* approximation schemes for batch reinforcement learning: A theoretical comparison, 2020.

Xu, J., Liu, X., Yan, J., Cai, D., Li, H., and Li, J. Learning to break the loop: Analyzing and mitigating repetitions for neural text generation. *Advances in Neural Information Processing Systems*, 35:3082–3095, 2022.

Yu, B. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pp. 94–116, 1994.

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.

Zhan, W., Huang, B., Huang, A., Jiang, N., and Lee, J. D. Offline reinforcement learning with realizability and single-policy concentrability, 2022.

Zhang, A. and Wang, M. Spectral state compression of markov processes. *IEEE transactions on information theory*, 66(5):3202–3231, 2019.

Zhu, Z., Li, X., Wang, M., and Zhang, A. Learning markov models via low-rank optimization. *Operations Research*, 70(4):2384–2398, 2022.

Ziemann, I. and Tu, S. Learning with little mixing. *Advances in Neural Information Processing Systems*, 35: 4626–4637, 2022.

Ziemann, I., Tu, S., Pappas, G. J., and Matni, N. Sharp rates in dependent learning theory: Avoiding sample size deflation for the square loss. *arXiv preprint arXiv:2402.05928*, 2024.

Ziemann, I. M., Sandberg, H., and Matni, N. Single trajectory nonparametric learning of nonlinear dynamics. In *conference on Learning Theory*, pp. 3333–3364. PMLR, 2022.

Organization of Appendix

- Appendix A incorporates positional encoding into the CCMC.
- Appendix B provides the proof for the theoretical results in Section 2.
- Appendix C provides the proof for the theoretical results in Section 3.
- Appendix D provides the proof for the theoretical results in Section 4.
- Appendix E provides the proof for the theoretical results in Section 5.
- Appendix F shares further related work on reinforcement learning and data-driven control.

A. The Role of Positional Encoding

In Section 2, we have established an equivalence between the attention models and the CCMC model using a mask. This mask was defined by $\boldsymbol{m}(X)$ in (2), which is associated with the occurrences of different tokens in an input prompt. In this section, we incorporate positional encodings into the CCMC model to study its impact on the transition dynamics.

To proceed, suppose the input prompt is fixed to be length L. We will use absolute position encoding Vaswani et al. (2017) which adds a position vector $\boldsymbol{u}_i$ at position i for $1 \leq i \leq L$. Recalling $\boldsymbol{e}_{x_i}$ is the vocabulary embedding, this leads to the following input embedding $\boldsymbol{x}_i = \boldsymbol{u}_i + \boldsymbol{e}_{x_i}$. Let $\boldsymbol{U} := [\boldsymbol{u}_1 \; \ldots \; \boldsymbol{u}_L]^\top \in \mathbb{R}^{L \times d}$ be the positional embedding matrix with $\boldsymbol{u}_i \in \mathbb{R}^d$. The linear classifier $\boldsymbol{C}$ in the attention models predicts the next token ID. In addition to Lemma 2.3, we assume the following to ensure that there is no bias to the classifier output from position embeddings.

Assumption A.1. The projection of positional embedding onto the columns of $\boldsymbol{C}$ is zero, i.e., $\boldsymbol{C}\boldsymbol{U}^\top = \boldsymbol{0}$

To quantify the effect of positional embedding on the output of the attention model, we define the variables $\boldsymbol{a} \in \mathbb{R}^L, \boldsymbol{b} \in \mathbb{R}^K$, and $\boldsymbol{V} \in \mathbb{R}^{L \times K}$ as follows:

$$\boldsymbol{a} := \exp(\boldsymbol{U}\boldsymbol{W}\boldsymbol{u}_L) \qquad \boldsymbol{b} := \exp(\boldsymbol{E}\boldsymbol{W}\boldsymbol{u}_L)$$
$$\boldsymbol{V} := \exp(\boldsymbol{U}\boldsymbol{W}\boldsymbol{E}^\top)$$

where $\exp(\cdot)$ represents the element-wise exponential function. Then, we define the probability distribution characterizing the CCMC model as follows: $\mathbb{P}_{(\boldsymbol{P},\boldsymbol{U})}(x_{L+1} = j|X) =$

$$\frac{b_j \pi_{x_L,j} \sum_{i=1}^{L} a_i V_{i,x_L} \cdot \mathbf{1}(x_i = j)}{\sum_{k=1}^{K} b_k \pi_{x_L,k} \sum_{i=1}^{L} a_i V_{i,x_L} \cdot \mathbf{1}(x_i = k)} \tag{10}$$

where π_{ij} is based on $\boldsymbol{P}^{\boldsymbol{W}}$, defined in (4). The intuition behind (10) is that, in Section 2, the CCMC model is constructed by a mask with $K-$dimension whereas (10) can be considered as a mask with $K \times L$ dimension. Now, we are ready to share our main results of this section:

Lemma A.2. *Suppose Assumptions 2.3 and A.1 hold. Then, for any $\boldsymbol{W} \in \mathbb{R}^{d \times d}$, there exists the transition matrix $\boldsymbol{P}^{\boldsymbol{W}}$ such that for any (X, y) we have the following:*

$$\mathbb{P}_{\boldsymbol{P},\boldsymbol{U}}(y|X) = \boldsymbol{c}_y^\top \boldsymbol{X}^\top \mathbb{S}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{x}_L)$$

Note that the one-to-one map, consistency of estimation, and finite sample guarantee can be built upon Lemma A.2 for the CCMC model with positional embedding by following similar arguments to the previous sections.

Proof. Consider any $\boldsymbol{W} \in \mathbb{R}^{d \times d}$ and an arbitrary input (X, y) with positional embedding $\boldsymbol{U}$. Let $\boldsymbol{X} = \boldsymbol{M}\boldsymbol{E} + \boldsymbol{U}$, where $\boldsymbol{M}$ is the same mapping matrix defined in Section B.1. Define $\boldsymbol{a} := \exp(\boldsymbol{U}\boldsymbol{W}\boldsymbol{u}_L), \boldsymbol{b} := \exp(\boldsymbol{E}\boldsymbol{W}\boldsymbol{u}_L), \boldsymbol{V} := \exp(\boldsymbol{U}\boldsymbol{W}\boldsymbol{E}^\top)$

we have:

$$\begin{aligned}
\boldsymbol{c}_y^\top \boldsymbol{X}^\top \mathbb{S}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{x}_L) &= \boldsymbol{c}_y^\top (\boldsymbol{E}^\top \boldsymbol{M}^\top + \boldsymbol{U}^\top)\mathbb{S}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{x}_L) \\
&\overset{(a)}{=} \boldsymbol{c}_y^\top \boldsymbol{E}^\top \boldsymbol{M}^\top \mathbb{S}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{x}_L) \\
&\overset{(b)}{=} \frac{\sum_{i=1}^L \exp(\boldsymbol{x}_y^\top \boldsymbol{W}\boldsymbol{x}_L) \cdot \mathbf{1}(x_i = y)}{\sum_{k=1}^K \sum_{i=1}^L \exp(\boldsymbol{x}_k^\top \boldsymbol{W}\boldsymbol{x}_L) \cdot \mathbf{1}(x_i = k)} \\
&= \frac{\sum_{i=1}^L \exp\Big((\boldsymbol{e}_y + \boldsymbol{u}_y)^\top \boldsymbol{W}(\boldsymbol{e}_{x_L} + \boldsymbol{u}_{x_L})\Big) \cdot \mathbf{1}(x_i = y)}{\sum_{k=1}^K \sum_{i=1}^L \exp\Big((\boldsymbol{e}_k + \boldsymbol{u}_k)^\top \boldsymbol{W}(\boldsymbol{e}_{x_L} + \boldsymbol{u}_{x_L})\Big) \cdot \mathbf{1}(x_i = k)} \\
&= \frac{b_y \cdot \exp(\boldsymbol{e}_y^\top \boldsymbol{W}\boldsymbol{e}_{x_L}) \sum_{i=1}^L a_i \cdot V_{i,x_L} \cdot \mathbf{1}(x_i = y)}{\sum_{k=1}^K b_k \cdot \exp(\boldsymbol{e}_k^\top \boldsymbol{W}\boldsymbol{e}_{x_L}) \sum_{i=1}^L a_i \cdot V_{i,x_L} \cdot \mathbf{1}(x_i = k)}
\end{aligned} \tag{11}$$

where (a) comes from the assumption that $\boldsymbol{C}\boldsymbol{U}^\top = \mathbf{0}$, (b) comes from the assumption that $\boldsymbol{C}\boldsymbol{E}^\top = \boldsymbol{I}_K$ and the definition of $\boldsymbol{M}$. Comparing Equation (11) with Equation (10), we can set $\pi_k = \exp(\boldsymbol{e}_k^\top \boldsymbol{W}\boldsymbol{e}_{x_L}) / \sum_{j\in[K]} \exp(\boldsymbol{e}_j^\top \boldsymbol{W}\boldsymbol{e}_{x_L})$ to obtain $\boldsymbol{P}^{\boldsymbol{W}}$. □

B. Proof of Theorems in Section 2

In this section, we share the proofs of Lemmas 2.2, 2.5, and Theorem 2.6.

B.1. Proof of Lemma 2.2

Lemma B.1 (Restated Lemma 2.2)**.** *Recall the probability vector $\boldsymbol{\pi}^X \in \mathbb{R}^K$ from Definition 2.1. We have that*

$$f_{\boldsymbol{W}}(X) = \boldsymbol{X}^\top \boldsymbol{s}_X = \boldsymbol{E}^\top \boldsymbol{\pi}^X.$$

Proof. Suppose $\boldsymbol{X} = \boldsymbol{M}\boldsymbol{E}$ where $\boldsymbol{M} \in \mathbb{R}^{L\times K}$ is a universal mapping matrix, which specifies the token index for each entry. Specifically, $M_{jk} = \begin{cases} 1/L, & \boldsymbol{x}_j = \boldsymbol{e}_k \\ 0, & \boldsymbol{x}_j \neq \boldsymbol{e}_k \end{cases}$. Note that $\boldsymbol{M}^\top \mathbf{1}_L = \boldsymbol{m}(X) = \boldsymbol{m}$. Then, we have:

$$\boldsymbol{X}^\top \boldsymbol{s}_X = \boldsymbol{E}^\top \boldsymbol{M}^\top \boldsymbol{s}_X. \tag{12}$$

Then, it is sufficient to prove $(\boldsymbol{M}^\top \boldsymbol{s}_X)_k = \pi_k^X$ for any $k \in [K]$. Let $s_0 = \sum_{j\in[K]} \exp(\boldsymbol{e}_j^\top \boldsymbol{W}\boldsymbol{x}_L)$. To proceed, using the definition of $\boldsymbol{s}_X$, we get:

$$\begin{aligned}
(\boldsymbol{M}^\top \boldsymbol{s}_X)_k &= \frac{m_k \cdot \exp(\boldsymbol{e}_k^\top \boldsymbol{W}\boldsymbol{x}_L)}{\sum_{j\in[K]} m_j \cdot \exp(\boldsymbol{e}_j^\top \boldsymbol{W}\boldsymbol{x}_L)} \\
&= \frac{m_k \cdot \exp(\boldsymbol{e}_k^\top \boldsymbol{W}\boldsymbol{x}_L)/s_0}{\sum_{j\in[K]} m_j \cdot \exp(\boldsymbol{e}_j^\top \boldsymbol{W}\boldsymbol{x}_L)/s_0} \\
&= \frac{m_k \cdot \pi_{x_L,k}}{\sum_{j\in[K]} m_j \cdot \pi_{x_L,j}} \\
&= \pi_k^X
\end{aligned} \tag{13}$$

which completes the proof. □

B.2. Proof of Lemma 2.5

Lemma B.2 (Restated Lemma 2.5)**.** *For all $\boldsymbol{W} \in \mathbb{R}^{d\times d}$ and X: $f_{\boldsymbol{W}}(X) = f_{\boldsymbol{\Pi}_{\mathcal{S}_{\boldsymbol{E}}}(\boldsymbol{W})}(X)$.*

Proof. Let $\mathcal{S}_{\boldsymbol{E}}^\perp$ be the orthogonal complement of the subspace $\mathcal{S}_{\boldsymbol{E}}$ in $\mathbb{R}^{d\times d}$. Then, for any $\boldsymbol{W} \in \mathbb{R}^{d\times d}$, we have $\boldsymbol{W} = \Pi_{\mathcal{S}_{\boldsymbol{E}}}(\boldsymbol{W}) + \Pi_{\mathcal{S}_{\boldsymbol{E}}^\perp}(\boldsymbol{W})$.

By definition of $\mathcal{S}_{\boldsymbol{E}}$, for any $j \in [K]$, there exists $c_j \in \mathbb{R}$ such that $\boldsymbol{e}_i^\top \Pi_{\mathcal{S}_{\boldsymbol{E}}^\perp}(\boldsymbol{W})\boldsymbol{e}_j = c_j$ for $i \in [K]$.

As a result, using the definition of $f_{\boldsymbol{W}}(X)$ in (SA), we obtain that

$$\begin{aligned} f_{\boldsymbol{W}}(X) = \boldsymbol{X}^\top \mathbb{S}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{x}_L) &= \boldsymbol{X}^\top \mathbb{S}(\boldsymbol{X}(\Pi_{\mathcal{S}_{\boldsymbol{E}}}(\boldsymbol{W}) + \Pi_{\mathcal{S}_{\boldsymbol{E}}^\perp}(\boldsymbol{W}))\boldsymbol{x}_L) \\ &= \boldsymbol{X}^\top \mathbb{S}(\boldsymbol{X}\Pi_{\mathcal{S}_{\boldsymbol{E}}}(\boldsymbol{W})\boldsymbol{x}_L + c_L \mathbf{1}_L) \\ &= \boldsymbol{X}^\top \mathbb{S}(\boldsymbol{X}\Pi_{\mathcal{S}_{\boldsymbol{E}}}(\boldsymbol{W})\boldsymbol{x}_L) \\ &= f_{\Pi_{\mathcal{S}_{\boldsymbol{E}}}(\boldsymbol{W})}(X) \end{aligned}$$

which completes the proof. □

B.3. Proof of Theorem 2.6

Theorem B.3 (Restated Theorem 2.6)**.** *Suppose Assumption 2.3 holds. Let $\mathcal{P}$ be the set of transition matrices with non-zero entries. For each $\boldsymbol{P} \in \mathcal{P}$, there is a unique $\boldsymbol{W} \in \mathcal{S}_{\boldsymbol{E}}$ with $\boldsymbol{P}^{\boldsymbol{W}} = \boldsymbol{P}$. Thus, for any prompt $X \in [K]^L$ and next token $y = x_{L+1} \in [K]$*

$$\mathbb{P}_{\boldsymbol{P}}(y|X) = \boldsymbol{c}_y^\top \boldsymbol{X}^\top \mathbb{S}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{x}_L)$$

where $\boldsymbol{c}_y$ is the y-th row of the linear prediction head $\boldsymbol{C}$.

Proof. 1) We prove that there exists $\boldsymbol{W} \in \mathcal{S}_{\boldsymbol{E}}$ satisfying the lemma. Using Lemma 2.2, for any $\boldsymbol{W} \in \mathcal{S}_{\boldsymbol{E}}$, let $\boldsymbol{\pi}_i^{\boldsymbol{W}} = \mathbb{S}(\boldsymbol{E}\boldsymbol{W}\boldsymbol{e}_i)$, we have:

$$\boldsymbol{c}_y^\top \boldsymbol{X}^\top \mathbb{S}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{x}_L) = \boldsymbol{c}_y^\top \boldsymbol{E}^\top \boldsymbol{\pi}^X = \pi_y^X = \frac{m_y \cdot \pi_{x_L,y}^{\boldsymbol{W}}}{\boldsymbol{m}^\top \boldsymbol{\pi}_{x_L}^{\boldsymbol{W}}}. \tag{14}$$

Comparing the equation above with

$$\mathbb{P}_{\boldsymbol{P}}(y|X) = \frac{m_y \cdot \pi_{x_L,y}}{\boldsymbol{m}^\top \boldsymbol{\pi}_{x_L}},$$

it is sufficient to prove that for any given $\boldsymbol{\pi}$, there exists a solution $\boldsymbol{W}$ for the following problem:

$$\boldsymbol{\pi} = \exp(\boldsymbol{E}\boldsymbol{W}\bar{\boldsymbol{x}}). \tag{15}$$

It is equivalent to solving the following linear system:

$$\boldsymbol{E}\boldsymbol{w} = \dot{\boldsymbol{\pi}} \tag{16}$$

where $\boldsymbol{w} = \boldsymbol{W}\bar{\boldsymbol{x}}, \dot{\boldsymbol{\pi}} = \log \boldsymbol{\pi}$. Since the rows of $\boldsymbol{E}$ are linearly independent from Assumption 2.3, $\boldsymbol{E}$ is right invertible. Combining with Lemma 2.5 implies that there exists at least one solution to the problem above.

2) We prove the uniqueness as follows: Let's assume the inverse. There exists $\boldsymbol{W}_1, \boldsymbol{W}_2 \in \mathcal{S}_{\boldsymbol{E}}$ such that $\boldsymbol{W}_1 \neq \boldsymbol{W}_2$ and we have the following for any (X, y):

$$\boldsymbol{c}_y^\top \boldsymbol{X}^\top \mathbb{S}(\boldsymbol{X}\boldsymbol{W}_1\boldsymbol{x}_L) = \boldsymbol{c}_y^\top \boldsymbol{X}^\top \mathbb{S}(\boldsymbol{X}\boldsymbol{W}_2\boldsymbol{x}_L) \tag{17}$$

As $\boldsymbol{W}_1, \boldsymbol{W}_2 \in \mathcal{S}_{\boldsymbol{E}}$ and $\boldsymbol{W}_1 \neq \boldsymbol{W}_2$, there exists $i, j, k \in [K]$ such that $\Pi_{(\boldsymbol{e}_i - \boldsymbol{e}_j)\boldsymbol{e}_k^\top}(\boldsymbol{W}_1 - \boldsymbol{W}_2) > 0$. Then, let's consider $X = [i, j]$ and $y = k$ and $\boldsymbol{x}_L = k$. (If the attention model is self-attention, then we can include k into X as well). Let $\boldsymbol{s}_{X,\boldsymbol{W}_1} = \mathbb{S}(\boldsymbol{X}\boldsymbol{W}_1\boldsymbol{x}_L)$ and $\boldsymbol{s}_{X,\boldsymbol{W}_2} = \mathbb{S}(\boldsymbol{X}\boldsymbol{W}_2\boldsymbol{x}_L)$. Let $u_{mn} = \boldsymbol{e}_m^\top \boldsymbol{W}_n \boldsymbol{e}_k$ for $n \in \{1, 2\}, m \in \{i, j\}$. As we have $\Pi_{(\boldsymbol{e}_i - \boldsymbol{e}_j)\boldsymbol{e}_k^\top}(\boldsymbol{W}_1 - \boldsymbol{W}_2) > 0$, then we obtain that $u_{j1} - u_{i1} \neq u_{j2} - u_{i2}$, which implies that $\boldsymbol{s}_{X,\boldsymbol{W}_1} = \boldsymbol{s}_{X,\boldsymbol{W}_2}$. As $\boldsymbol{C}\boldsymbol{E}^\top = \boldsymbol{I}$ by Assumption 2.3, (17) cannot hold, which is a contradiction. This completes the proof. □

C. Proof of Theorems in Section 3

In this section, we will analyze the case where the ground-truth transition matrix $\boldsymbol{P}^{\text{GT}}$ may have zero transitions. Note that the transition matrix that has zero probability transitions is not important for the attention models as the equivalency between the attention models and the CCMC model is constructed for the non-zero transitions. We provide a detailed analysis for the sake of the Markov chain community. First, we share a supplementary lemma for this section:

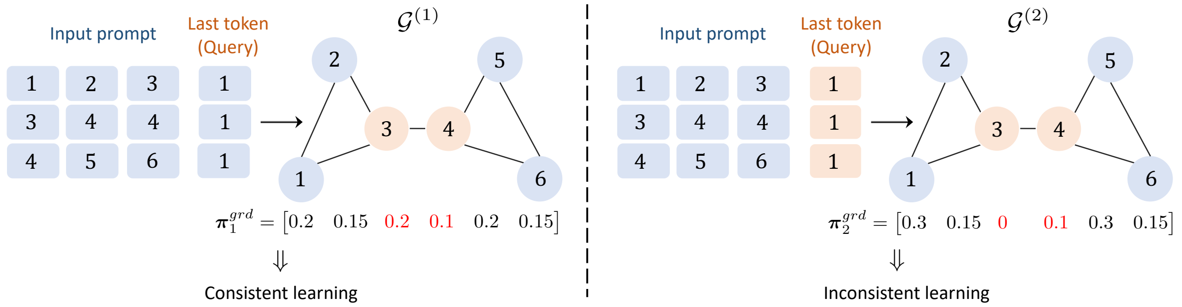


Figure 6. Illustration of the Co-occurrence Graphs for the Cross-attention Models with Possible Zero Transition Probability in $\boldsymbol{P}^{\text{GT}}$. In this figure, we draw the co-occurrence graphs of the input prompt distribution whose support consists of three elements under the cross-attention model. In the left figure, the first column of the transition matrix $\boldsymbol{P}^{\text{GT}}$ does not include any zero transition whereas, in the right figure, the first column of the transition matrix $\boldsymbol{P}^{\text{GT}}$ includes zero transition. For the left figure, the co-occurrence graph $\mathcal{G}^{(1)}$ is connected with respect to $\boldsymbol{P}^{\text{GT}}$. However, in the right figure, the co-occurrence graph $\mathcal{G}^{(1)}$ is not connected with respect to $\boldsymbol{P}^{\text{GT}}$ as there is no path between 2nd vertex and the 5th vertex using non-zero vertex.

Lemma C.1. *Define the function $f : \mathbb{R}^n_{>0} \to \mathbb{R}$ for fixed $(y_i)_{i=1}^n$ positive variables and define $\boldsymbol{x}^*$ as follows:*

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n_{>0}} f(\boldsymbol{x}) := \arg\min_{\boldsymbol{x}\in\mathbb{R}^n_{>0}} -\sum_{i=1}^n y_i \log(x_i) \tag{18}$$

$$\textit{subject to} \quad \sum_{i=1}^n x_i = 1. \tag{19}$$

Then, we have $c \in \mathbb{R}$ such that $c = \frac{x_i^*}{y_i}$ for all $i \in [n]$.

Proof. Let $\mathcal{L}(\boldsymbol{x}, \lambda)$ be the Lagrangian function:

$$\mathcal{L}(\boldsymbol{x}, \lambda) = -\sum_{i=1}^n y_i \log(x_i) + \lambda\left(1 - \sum_{i=1}^n x_i\right).$$

As a result of KKT condition, we have

$$-\frac{y_i}{x_i^*} - \lambda = 0 \quad \forall i \in [n] \tag{20}$$

which completes the proof. □

Definition C.2. (i) Let Ω_k be the set of input prompts that are inside the support of $\mathcal{D}_{\mathcal{X}}$ and whose queries are the k-th token. Then, we define an undirected co-occurrence graph $(\mathcal{G}^{(k)})_{k=1}^K$ with K vertices such that the vertices $i, j \in [K]$ are connected in $\mathcal{G}^{(k)}$ if there exists an input prompt in Ω_k that includes both the i-th and j-th tokens.

(ii) Let $\boldsymbol{P}^{\text{GT}}$ be the ground-truth transition matrix. For an arbitrary query $k \in [K]$, the co-occurrence graph $\mathcal{G}^{(k)}$ is said to be 'connected with respect to $\boldsymbol{P}^{\text{GT}}$' if it satisfies the following: For every pair of vertices $(i, j) \in [K] \times [K]$, $P^{\text{GT}}_{ki} \neq 0$, $P^{\text{GT}}_{kj} \neq 0$, there exists a path of $(v_m)_{m=1}^M$ such that $v_1 = i$, $v_M = j$, and $P^{\text{GT}}_{kv_m} \neq 0$ for every $m \in [M]$.

To explain the notion of the 'connected graph with respect to $\boldsymbol{P}^{\text{GT}}$', we demonstrate a dataset for different $\boldsymbol{P}^{\text{GT}}$ in Figure 6. Note that the following theorem reduces to Theorem 3.4 if $\boldsymbol{P}^{\text{GT}}$ has non-zero transition probabilities.

Theorem C.3 (Stronger version of Theorem 3.4)**.** *Let $\boldsymbol{P}^{GT}$ be the transition matrix of a Markov chain that determines the next token under the CCMC model. Let $(\mathcal{G}^{(k)})_{k=1}^K$ be the co-occurrence graphs based on the input prompt distribution $\mathcal{D}_{\mathcal{X}}$. Then, the estimation of $\boldsymbol{P}^{GT}$ in (6) with the prompt distribution $\mathcal{D}_{\mathcal{X}}$ is consistent if and only if $\mathcal{G}^{(k)}$ is connected with respect to $\boldsymbol{P}^{GT}$ (see Definition C.2 (ii)) for every $k \in [K]$.*

Proof. For an arbitrary $k \in [K]$, we are going to prove that the k-th column of $\boldsymbol{P}^{\text{GT}}$ and $\boldsymbol{P}_\star$ are equivalent if and only if $\mathcal{G}^{(k)}$ is connected with respect to $\boldsymbol{P}^{\text{GT}}$. The proof of this statement is sufficient to prove Theorem C.3.

Let $\boldsymbol{\pi}_*$ and $\boldsymbol{\pi}^{\mathrm{GT}}$ be the k-th column of $\boldsymbol{P}_\star$ and $\boldsymbol{P}^{\mathrm{GT}}$, respectively. Let Ω_k be the set of input prompts such that the query is the k-th token. Let $\mathcal{S}_{\boldsymbol{\pi}^{\mathrm{GT}}}$ be the set of token IDs i such that $P^{\mathrm{GT}}_{ki} \neq 0$. Let X be an arbitrary prompt inside the set Ω_k. Let $[X]$ represent the set of token IDs inside the input prompt X. We change the notation of $\mathbb{P}_{\boldsymbol{P}}(y|X)$ to $\mathbb{P}_{\boldsymbol{\pi}}(y|X)$ as we will deal with the input prompts whose queries are the same. We first minimize the population risk for this specific input sequence X.

$$\begin{aligned}\boldsymbol{\pi}_*^X &= \arg\min_{\boldsymbol{\pi}} \mathbb{E}_y\left[-\log \mathbb{P}_{\boldsymbol{\pi}}(y|X)\right] \\ &= \arg\min_{\boldsymbol{\pi}} -\sum_{i\in\mathcal{S}_{\boldsymbol{\pi}^{\mathrm{GT}}}} \mathbb{P}_{\boldsymbol{\pi}^{\mathrm{GT}}}(y=i|X)\log\left(\mathbb{P}_{\boldsymbol{\pi}}(y=i|X)\right). \end{aligned} \tag{21}$$

Applying Lemma C.1 on (21) where $K = n$, $y_i = \mathbb{P}_{\boldsymbol{\pi}^{\mathrm{GT}}}(y = i|X)$ and $x_i = \mathbb{P}_{\boldsymbol{\pi}}(y = i|X)$, we know that there exists $c_1 \in \mathbb{R}$ such that

$$\frac{\mathbb{P}_{\boldsymbol{\pi}^{\mathrm{GT}}}(y=i|X)}{\mathbb{P}_{\boldsymbol{\pi}_*^X}(y=i|X)} = c_1 \qquad \forall i \in \mathcal{S}_{\boldsymbol{\pi}^{\mathrm{GT}}}. \tag{22}$$

From (1), we know that the output probabilities of the CCMC model are a linear transformation of the transition probabilities based on the occurrences of tokens inside an input prompt. As a result, the linear transformation is one-to-one for the tokens IDs i such that $i \in \mathcal{S}_{\boldsymbol{\pi}^{\mathrm{GT}}} \cap [X]$. Then, there exists $c_2 > 0$ such that

$$\frac{(\pi_*^{\boldsymbol{X}})_i}{(\pi^{\mathrm{GT}})_i} = c_2 \qquad \forall i \in \mathcal{S}_{\boldsymbol{\pi}^{\mathrm{GT}}} \cap [X]. \tag{23}$$

Note that $\boldsymbol{\pi}_* = \boldsymbol{\pi}^{\mathrm{GT}}$ satisfies (23) for all $i \in [K]$. This means that $\boldsymbol{\pi}^{\mathrm{GT}}$ is a solution to the population risk minimization problem in (6). What is remaining is that there is no other $\boldsymbol{\pi} \neq \boldsymbol{\pi}^{\mathrm{GT}}$ such that $\boldsymbol{\pi}$ is a solution to the population risk minimization problem. As $\boldsymbol{\pi} = \boldsymbol{\pi}^{\mathrm{GT}}$ satisfies (23) for all $i \in \mathcal{S}_{\boldsymbol{\pi}^{\mathrm{GT}}} \cap [X]$, if there exists any other $\boldsymbol{\pi}$ that minimizes (6), then this $\boldsymbol{\pi}$ should satisfy (23) for all possible Ω_k.

Proof of $\Leftarrow$: Now, we know that $\mathcal{G}^{(k)}$ is connected with respect to $\boldsymbol{P}^{\mathrm{GT}}$ and we want to prove the consistency of estimation. Let's assume the inverse: There exists a probability vector $\boldsymbol{\pi} \in \mathbb{R}^K$ such that for an arbitrary index $\bar{i} \in [K]$ we have $\pi_{\bar{i}} > \pi^{\mathrm{GT}}_{\bar{i}}$ and π minimizes (6). Then, there must exist $\bar{j} \in [K]$ such that $\pi_{\bar{j}} < \pi^{\mathrm{GT}}_{\bar{j}}$ as both $\boldsymbol{\pi}$ and $\boldsymbol{\pi}^{\mathrm{GT}}$ are probability vectors. This implies that

$$\frac{\pi_{\bar{i}}}{\pi^{\mathrm{GT}}_{\bar{i}}} > 1 > \frac{\pi_{\bar{j}}}{\pi^{\mathrm{GT}}_{\bar{j}}} \tag{24}$$

Since $\mathcal{G}^{(k)}$ is connected with respect to $\boldsymbol{P}^{\mathrm{GT}}$, there exists a path of $(v_m)_{m=1}^M \in \mathcal{G}^{(k)}$ such that $v_1 = \bar{i}$ and $v_M = \bar{j}$ and $v_m \in \mathcal{S}_{\boldsymbol{\pi}^{\mathrm{GT}}}$ for every $m \in [M]$. Since the path $(v_m)_{m=1}^M$ is inside the co-occurrence graph $\mathcal{G}^{(k)}$, there exists an input prompt sequence $(X_m)_{m=1}^{M-1}$ such that $X_m \in \Omega_k$ and X_m includes the tokens with ID v_m and v_{m+1}. Using (23) for the input prompt sequence $(X_m)_{m=1}^{M-1}$, we obtain the following using the fact that these input sequences include the k^{th} token:

$$\frac{\pi_{v_m}}{\pi^{\mathrm{GT}}_{v_m}} = \frac{\pi_{v_{m+1}}}{\pi^{\mathrm{GT}}_{v_{m+1}}} \qquad \forall m \in [M-1] \tag{25}$$

Combining (25) and the fact that $v_1 = \bar{i}$ and $v_{M-1} = \bar{j}$, we obtain

$$\frac{\pi_{\bar{i}}}{\pi^{\mathrm{GT}}_{\bar{i}}} = \frac{\pi_{\bar{j}}}{\pi^{\mathrm{GT}}_{\bar{j}}}$$

which contradicts with (24). This means that if $\mathcal{G}^{(k)}$ is connected with respect to $\boldsymbol{P}^{\mathrm{GT}}$, then the only solution that minimizes (6) is $\boldsymbol{\pi}^{\mathrm{GT}}$, which is equivalent to the consistency of estimation in Definition 3.1.

Proof of $\Rightarrow$: Now, we know that the estimation of $\boldsymbol{\pi}^{\mathrm{GT}}$ in (6) is consistent and we want to prove that $\mathcal{G}^{(k)}$ is connected with respect to $\boldsymbol{P}^{\mathrm{GT}}$. Let's assume the inverse: There exists $i, j \in \mathcal{S}_{\boldsymbol{\pi}^{\mathrm{GT}}}$ such that there is no path between them in $\mathcal{G}^{(k)}$ using the vertices in $\mathcal{S}_{\boldsymbol{\pi}^{\mathrm{GT}}}$. Then, there exists a partition $\mathcal{S}_1$ and $\mathcal{S}_2$ such that $\mathcal{S}_1 \cap \mathcal{S}_2 = \emptyset$, $\mathcal{S}_1 \cup \mathcal{S}_2 = \mathcal{S}_{\boldsymbol{\pi}^{\mathrm{GT}}}$, $i \in \mathcal{S}_1$, $j \in \mathcal{S}_2$, and there is no path from any element of $\mathcal{S}_1$ to any element $\mathcal{S}_2$ in $\mathcal{G}^{(k)}$ using the vertices in $\mathcal{S}_{\boldsymbol{\pi}^{\mathrm{GT}}}$. We construct a probability vector

$\boldsymbol{\pi} \in \mathbb{R}^K$ such that $\boldsymbol{\pi}$ minimizes (6) and $\boldsymbol{\pi} \neq \boldsymbol{\pi}^{\mathrm{GT}}$:

$$\begin{aligned}
&\bar{\boldsymbol{\pi}} \in \mathbb{R}^K \\
&\bar{\pi}_k = 2\pi_k^{\mathrm{GT}} \qquad \forall k \in \mathcal{S}_1 \\
&\bar{\pi}_k = \pi_k^{\mathrm{GT}} \qquad \forall k \in \mathcal{S}_2 \\
&\pi_k = \begin{cases} \frac{\bar{\pi}_k}{\sum_{k'=1}^K \bar{\pi}_{k'}}, & \text{if } k \in \mathcal{S}_{\boldsymbol{\pi}^{\mathrm{GT}}} \\ 0, & \text{if } k \notin \mathcal{S}_{\boldsymbol{\pi}^{\mathrm{GT}}} \end{cases}
\end{aligned}$$

Note that this constructed $\boldsymbol{\pi}$ satisfies (23) for all $X \in \Omega_k$, which implies that $\boldsymbol{\pi}$ is a minimizer of (6) and $\boldsymbol{\pi} \neq \boldsymbol{\pi}^{\mathrm{GT}}$. This is a contradiction to the consistency of estimation, which completes the proof. □

Corollary C.4 (Restated Corollary 3.5)**.** *Given $k \in [K]$, let $\mathcal{C}_k \subset [K]$ be the set of all tokens that appear within some training prompt $X \in \Omega_k$ where X ends with token k. Let $\boldsymbol{P}^*$ be the transition model learned by the self-attention trained on $X \sim \mathcal{D}_{\mathcal{X}}$ with labels sampled from the ground-truth model $\boldsymbol{P}^{GT}$. For any $k \in [K]$, we have that*

$$\boldsymbol{\pi}^*_{k,\mathcal{C}_k} = \boldsymbol{\pi}^{GT}_{k,\mathcal{C}_k}.$$

Here $\boldsymbol{\pi}_{k,\mathcal{C}}$ denotes the probability distribution induced by normalizing the entries of $\boldsymbol{\pi}_k$ over the set $\mathcal{C}$.

Proof. The first observation for the proof is that the co-occurrence graph of self-attention consists of a star graph and some isolated vertices. Here, the center vertex of the star graph is the last token, and the isolated vertices correspond to the unobserved tokens associated with that last token. This means that when we exclude the tokens that are not inside the set $\mathcal{C}_k$, the remaining graph is connected. Finally, when we apply the proof of Theorem C.3 for a specific column, we obtain the advertised result. □

C.1. Strict Convexity and Smoothness of Loss Function

In this subsection, we discuss the convexity, strict convexity, and smoothness of the loss functions. First, we analyze the empirical loss function $\widehat{\mathcal{L}}_n(\boldsymbol{W})$, then we connect our analysis with the population loss function $\mathcal{L}(\boldsymbol{W})$. Throughout the section, we omit the subscript of n in $\widehat{\mathcal{L}}_n(\boldsymbol{W})$ as we are analyzing the loss function for any arbitrary n. First, we define the following subspace:

Definition C.5. Define the subspace $\mathcal{S}_{\mathcal{T}}$ as the span of all matrices $(\boldsymbol{e}_i - \boldsymbol{e}_j)\boldsymbol{e}_k^\top$ for all $i, j, k \in [K]$ such that there exists an input prompt in the dataset that includes the i^{th} token, j^{th} token is the next token, and k^{th} token is the query or the last token.

The following Lemma proves the strict convexity inside a subspace, which is a slightly generalized version of Lemma 9 in (Li et al., 2024). Even though the proofs are almost the same as (Li et al., 2024), we restate the proofs for the sake of completeness and notational coherence.

Lemma C.6 (Stronger version of Lemma 3.7, (Li et al., 2024))**.** *Suppose Assumptions 2.3 and 4.1 hold. Then $\mathcal{L}(\boldsymbol{W})$ and $\widehat{\mathcal{L}}(\boldsymbol{W})$ is convex on $\mathbb{R}^{d\times d}$. Furthermore, $\mathcal{L}(\boldsymbol{W})$ and $\widehat{\mathcal{L}}(\boldsymbol{W})$ are strictly convex on $\mathcal{S}_{\boldsymbol{E}}$ and $\mathcal{S}_{\mathcal{T}}$, respectively.*

Proof. First, we are going to prove the convexity and the strict convexity for $\widehat{\mathcal{L}}(\boldsymbol{W})$. Then, we apply our findings to $\mathcal{L}(\boldsymbol{W})$. Recall from Definition 2.4 that $\mathcal{S}_{\boldsymbol{E}}$ is the span of all matrices $(\boldsymbol{e}_i - \boldsymbol{e}_j)\boldsymbol{e}_k^\top$ for $i, j, k \in [K]$.

- **First Case: $\boldsymbol{W} \in \mathcal{S}_{\boldsymbol{E}}$.** Let $g : \mathcal{S}_{\boldsymbol{E}} \to \mathbb{R}^{K\times K}$ such that $g(\boldsymbol{W}) = \boldsymbol{E}\boldsymbol{W}\boldsymbol{E}^\top$. By definition, this function is linear. In addition to that, this function g is invertible on $g(\mathcal{S}_{\boldsymbol{E}})$ by Assumption 2.3 and the domain of the function is $\mathcal{S}_{\boldsymbol{E}}$. Note that Assumption 2.3 ensures $\operatorname{rank}(\boldsymbol{E}) = K$.

Let $\boldsymbol{E}' = \boldsymbol{C}' = \boldsymbol{I}_k$, $(\boldsymbol{X}_i', y_i')_{i=1}^n$ be a dataset constructed from $(\boldsymbol{X}_i, y_i)_{i=1}^n$ such that $y_i' = y_i$ and $\boldsymbol{X}_i' = \boldsymbol{X}_i\boldsymbol{E}^\dagger$. Then, for any $\boldsymbol{W}' \in \mathbb{R}^{K\times K}$, we have the following:

$$\widehat{\mathcal{L}} \circ g^{-1}(\boldsymbol{W}') = \frac{1}{n}\sum_{i=1}^{n} -\log\left((\boldsymbol{c}'_{y_i})^\top (\boldsymbol{X}_i')^\top \mathbb{S}(\boldsymbol{X}_i'\boldsymbol{W}'\boldsymbol{x}'_{i,L_i})\right)$$

Using Lemma C.7 and C.8, we know that $\widehat{\mathcal{L}} \circ g^{-1}(\boldsymbol{W}')$ is convex on $\mathbb{R}^{K\times K}$ and strictly convex on $g(\mathcal{S}_{\mathcal{T}})$. Using these two facts and Lemma C.7, we have $\widehat{\mathcal{L}}(\boldsymbol{W})$ is convex on $\mathcal{S}_{\boldsymbol{E}}$ and strictly convex on $\mathcal{S}_{\mathcal{T}} \cap \mathcal{S}_{\boldsymbol{E}} = \mathcal{S}_{\mathcal{T}}$.

• **Second Case: $\boldsymbol{W} \notin \mathcal{S}_{\boldsymbol{E}}$.** Using Lemma 2.5, we have the following for any $0 \leq \lambda \leq 1$:

$$\widehat{\mathcal{L}}(\lambda \boldsymbol{W}_1 + (1-\lambda)\boldsymbol{W}_2) = \widehat{\mathcal{L}}(\lambda \Pi_{\mathcal{S}_K}(\boldsymbol{W}_1) + \lambda \Pi_{\mathcal{S}_K}(\boldsymbol{W}_2)) \tag{26}$$

Then, using (26), we have the following:

$$\begin{aligned}\lambda\widehat{\mathcal{L}}(\boldsymbol{W}_1) + (1-\lambda)\widehat{\mathcal{L}}(\boldsymbol{W}_2) &= \lambda\widehat{\mathcal{L}}(\Pi_{\mathcal{S}_{\boldsymbol{E}}}(\boldsymbol{W}_1)) + (1-\lambda)\widehat{\mathcal{L}}(\Pi_{\mathcal{S}_{\boldsymbol{E}}}(\boldsymbol{W}_2)) \\ &\overset{(a)}{\geq} \widehat{\mathcal{L}}(\lambda\Pi_{\mathcal{S}_{\boldsymbol{E}}}(\boldsymbol{W}_1) + \lambda\Pi_{\mathcal{S}_{\boldsymbol{E}}}(\boldsymbol{W}_2)) = \widehat{\mathcal{L}}(\lambda\boldsymbol{W}_1 + (1-\lambda)\boldsymbol{W}_2)\end{aligned}$$

where (a) follows from the convexity of $\widehat{\mathcal{L}}(\boldsymbol{W})$ inside $\mathcal{S}_{\boldsymbol{E}}$. This implies that $\widehat{\mathcal{L}}(\boldsymbol{W})$ is convex when $\boldsymbol{W} \notin \mathcal{S}_{\boldsymbol{E}}$. Note that $\mathcal{S}_{\mathcal{T}} \subset \mathcal{S}_{\boldsymbol{E}}$, therefore we do not look at the strict convexity in this case.

For the loss function $\mathcal{L}(\boldsymbol{W})$, the same procedure can be applied. By Assumption 4.1, the subspace of $\mathcal{S}_{\mathcal{T}}$ for the population dataset becomes $\mathcal{S}_{\boldsymbol{E}}$ as $\mathcal{G}^{(k)}$ are connected for every $k \in [K]$. □

Lemma C.7 (Lemma 10, (Li et al., 2024)). *Let $T : \mathcal{X} \to \mathcal{Y}$ be an invertible linear map. If a function $f : \mathcal{Y} \to \mathbb{R}$ is convex/strictly convex on $\mathcal{Y}$, then $f \circ T(x)$ is a convex/strictly convex function on $\mathcal{X}$.*

Proof. Let $x_1 \neq x_2 \in \mathcal{X}$ be arbitrary variables. Let $y_1 = T(x_1)$ and $y_2 = T(x_2)$. Since T is an invertible map, $y_1 \neq y_2$. Since T is a linear map, $T(\lambda x_1 + (1-\lambda)x_2) = \lambda y_1 + (1-\lambda)y_2$ for $0 < \lambda < 1$. Then, we obtain the following

$$\begin{aligned}\lambda(f \circ T(x_1)) + (1-\lambda)(f \circ T(x_2)) &= \lambda f(y_1) + (1-\lambda) f(y_2) \\ &\overset{(a)}{>} f(\lambda y_1 + (1-\lambda)y_2) \\ &= f \circ T(\lambda x_1 + (1-\lambda)x_2)\end{aligned}$$

where (a) follows from the strict convexity of the function f. This implies that $f \circ T(x)$ is a strictly convex function on $\mathcal{X}$. Note that if $y_1 = y_2$, then we cannot achieve (a). Additionally, if f is convex instead of strictly convex, then $>$ in (a) is changed to $\geq$, and $f \circ T(x)$ is convex. □

Lemma C.8 (Lemma 11, (Li et al., 2024)). *Let $\boldsymbol{E} = \boldsymbol{I}_d$. Let $f : \mathbb{R}^{d \times d} \to \mathbb{R}^{d^2}$ be a linear transformation defined as $f(\boldsymbol{W}) = \boldsymbol{v}$ where $v_{i \times d + j} = \boldsymbol{e}_i^T \boldsymbol{W} \boldsymbol{e}_j$. Then, $\widehat{\mathcal{L}} \circ f^{-1}(\boldsymbol{v})$ is convex. Furthermore, $\widehat{\mathcal{L}} \circ f^{-1}(\boldsymbol{v})$ is strictly convex on $f(\mathcal{S}_{\mathcal{T}})$, where $\mathcal{S}_{\mathcal{T}}$ is defined in Definition C.5.*

Proof. • **We first prove that $\mathcal{L} \circ f^{-1}(\boldsymbol{v})$ is convex.** Let $\ell : \mathbb{R}^{d^2} \times \mathbb{R}^{T \times d} \times \mathbb{R} \to \mathbb{R}$ be defined as follows:

$$\ell(\boldsymbol{v}, \boldsymbol{X}, y) := -\log\left(\boldsymbol{c}_y^\top \boldsymbol{X}^\top \mathbb{S}(\boldsymbol{X}\big(f^{-1}(\boldsymbol{v})\big)\boldsymbol{x}_L)\right).$$

Then, we have the following:

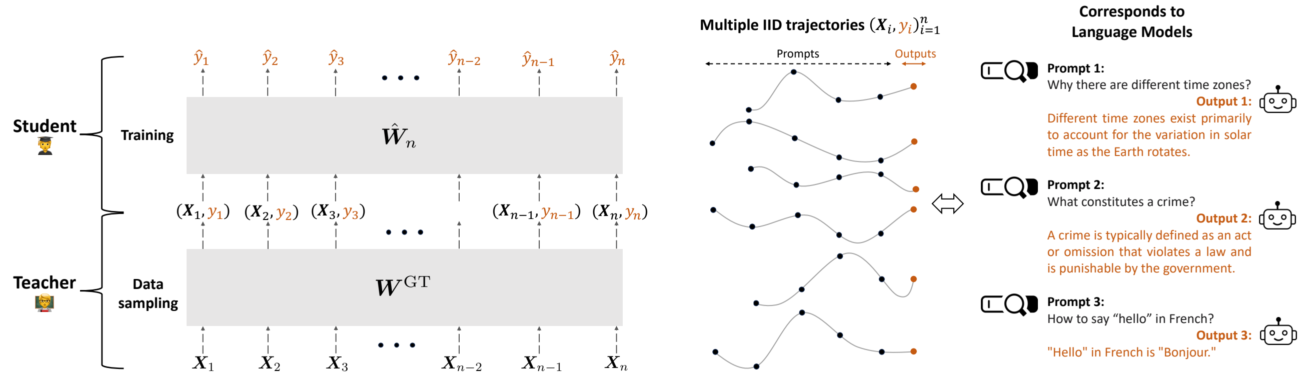
$$\widehat{\mathcal{L}} \circ f^{-1}(\boldsymbol{v}) = \frac{1}{n}\sum_{i=1}^{n} -\log\left(\boldsymbol{c}_{y_i}^\top \boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{X}_i\big(f^{-1}(\boldsymbol{v})\big)\boldsymbol{x}_{i,L_i})\right) = \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{v}, \boldsymbol{X}_i, y_i). \tag{27}$$

Note that the summation of convex functions is convex. Therefore, it is sufficient to prove the convexity of $\mathcal{L} \circ f^{-1}(\boldsymbol{v})$ by proving the convexity of $\bar{\ell}(\boldsymbol{v}, \boldsymbol{X}, y)$ for an arbitrary pair of input sequence and label $(\boldsymbol{X}, y)$. For the simplicity of notation, we use $\ell(\boldsymbol{v})$ instead of $\ell(\boldsymbol{v}, \boldsymbol{X}, y)$. Let k be the last token of $\boldsymbol{X}$. By Assumption 2.3 and log-loss, we know that

$$\begin{aligned}\ell(\boldsymbol{v}) := \ell(\boldsymbol{v}, \boldsymbol{X}, y) &= -\log\left(\frac{\boldsymbol{m}(\boldsymbol{X})_y \cdot e^{v_{y \times d + k}}}{\sum_{j \in [K]} \boldsymbol{m}(\boldsymbol{X})_j \cdot e^{v_{j \times d + k}}}\right) \\ &= \log\left(\sum_{j \in [K]} \boldsymbol{m}(\boldsymbol{X})_y \cdot e^{v_{j \times d + k}}\right) - \log(\boldsymbol{m}(\boldsymbol{X})_y \cdot e^{v_{y \times d + k}}).\end{aligned}$$

Let $\boldsymbol{z} \in \mathbb{R}^{d^2}$ be a vector such that the $(j \times d + k)^{\text{th}}$ element of $\boldsymbol{z}$ is $z_{j \times d + k} = \boldsymbol{m}(\boldsymbol{X})_j \cdot e^{v_{j \times d + k}}$ for $k \in [K]$, otherwise $z_i = 0$. Then, the Hessian matrix of $\bar{\ell}(\boldsymbol{v})$ is

$$\nabla^2 \ell(\boldsymbol{v}) = \frac{1}{(\boldsymbol{1}^\top \boldsymbol{z})^2}\left((\boldsymbol{1}^\top \boldsymbol{z})\text{diag}(\boldsymbol{z}) - \boldsymbol{z}\boldsymbol{z}^\top\right)$$


Student
Training
$\hat{y}_1$ $\hat{y}_2$ $\hat{y}_3$ $\hat{y}_{n-2}$ $\hat{y}_{n-1}$ $\hat{y}_n$
$\hat{W}_n$
(X_1, y_1) (X_2, y_2) (X_3, y_3) (X_{n-1}, y_{n-1}) (X_n, y_n)
Teacher
Data sampling
W^{GT}
X_1 X_2 X_3 X_{n-2} X_{n-1} X_n
Multiple IID trajectories $(X_i, y_i)_{i=1}^n$
Prompts
Outputs
Corresponds to Language Models

Prompt 1:
Why there are different time zones?
Output 1:
Different time zones exist primarily to account for the variation in solar time as the Earth rotates.

Prompt 2:
What constitutes a crime?
Output 2:
A crime is typically defined as an act or omission that violates a law and is punishable by the government.

Prompt 3:
How to say "hello" in French?
Output 3:
"Hello" in French is "Bonjour."

Figure 7. **Left:** Illustration of finite sample learning where the next tokens are sampled from the ground-truth model, which corresponds to single outputs from multiple IID trajectories. **Right:** In practice, the scenario is analogous to querying language models with prompts on different topics and using the responses to train a tiny model. In Theorem 3.4, we characterize the condition when the tiny model can estimate the ground-truth model consistently.

For any $\boldsymbol{u} \in \mathbb{R}^{d^2}$, we obtain that

$$\boldsymbol{u}^\top \nabla^2 \ell(\boldsymbol{v}) \boldsymbol{u} = \frac{1}{(\mathbf{1}^\top \boldsymbol{z})^2}\left(\left(\sum_{j=1}^{d^2} z_j\right)\left(\sum_{j=1}^{d^2} u_j^2 z_j\right) - \left(\sum_{j=1}^{d^2} u_j z_j\right)^2\right) \geq 0. \tag{28}$$

Since $\boldsymbol{z}_i \geq 0$, $i \in [d^2]$, (28) follows from the Cauchy-Schwarz inequality $(\boldsymbol{\alpha}^\top \boldsymbol{\alpha})(\boldsymbol{\beta}^\top \boldsymbol{\beta}) \geq (\boldsymbol{\alpha}^\top \boldsymbol{\beta})^2$ applied to the vectors with $\alpha_i = u_i \sqrt{z_i}$ and $\beta_i = \sqrt{z_i}$. The equality condition holds $k\boldsymbol{\alpha} = \boldsymbol{\beta}$ for $k \neq 0$. This means that $\ell(\boldsymbol{v})$ is convex.

• **Next, we will show that $\widehat{\mathcal{L}} \circ f^{-1}(\boldsymbol{v})$ is strictly convex on $f(\mathcal{S}_\mathcal{T})$.** Assume that $\widehat{\mathcal{L}} \circ f^{-1}(\boldsymbol{v})$ is not strictly convex on $f(\mathcal{S}_\mathcal{T})$. Using the convexity of $\widehat{\mathcal{L}} \circ f^{-1}(\boldsymbol{v})$, this implies that there exist $\boldsymbol{u}, \boldsymbol{v} \in f(\mathcal{S}_\mathcal{T})$, $\|\boldsymbol{u}\|_2 > 0$ such that

$$\boldsymbol{u}^\top \left(\nabla^2 \widehat{\mathcal{L}} \circ f^{-1}(\boldsymbol{v})\right) \boldsymbol{u} = 0$$

Combining this with the convexity of $\ell(\boldsymbol{v})$ and (27), we have the following:

$$\boldsymbol{u}^\top \left(\nabla^2 \ell(\boldsymbol{v}, \boldsymbol{X}_i, \boldsymbol{y}_i)\right) \boldsymbol{u} = 0 \qquad \forall i \in [n] \tag{29}$$

Now, we are going to prove that $\|\boldsymbol{u}\|_2 = 0$ if (29) holds. As $\boldsymbol{u} \in f(\mathcal{S}_\mathcal{T})$, there exists $\boldsymbol{W} \in \mathcal{S}_\mathcal{T}$ such that $f(\boldsymbol{W}) = \boldsymbol{u}$. As the function f preserves the norm, $\|\boldsymbol{W}\|_F > 0$. By definition of $\mathcal{S}_\mathcal{T}$, there exist $\bar{i}, \bar{j}, \bar{k} \in [K]$ and $(\boldsymbol{X}_{\bar{n}}, y_{\bar{n}}) \in \mathcal{T}$ such that $\langle (\boldsymbol{e}_{\bar{i}} - \boldsymbol{e}_{\bar{j}})\boldsymbol{e}_{\bar{k}}^T, \boldsymbol{W} \rangle > 0$, $\boldsymbol{X}_{\bar{n}}$ includes the $\bar{j}^{\text{th}}$ token, the last token of $\boldsymbol{X}_{\bar{n}}$ is the $\bar{k}^{\text{th}}$ token, and $y_{\bar{n}} = \bar{i}$. On the other hand, by the generation of the dataset $\mathcal{T}$, the next token should be inside the input prompt. Then, $z_{\bar{i} \times d + \bar{k}}$ and $z_{\bar{j} \times d + \bar{k}}$ in (28) are non-zero for this input sequence $\boldsymbol{X}_{\bar{n}}$. Using the equality condition of Cauchy-Schwartz Inequality in (28), we obtain that $u_{\bar{i} \times d + \bar{k}} - u_{\bar{j} \times d + \bar{k}} = 0$. This implies that

$$\begin{aligned}
0 &= u_{\bar{i} \times d + \bar{k}} - u_{\bar{j} \times d + \bar{k}} \\
&= \boldsymbol{e}_{\bar{i}}^T \boldsymbol{W} \boldsymbol{e}_{\bar{k}} - \boldsymbol{e}_{\bar{j}}^T \boldsymbol{W} \boldsymbol{e}_{\bar{k}} \\
&= (\boldsymbol{e}_{\bar{i}} - \boldsymbol{e}_{\bar{j}})^T \boldsymbol{W} \boldsymbol{e}_{\bar{k}} = \langle (\boldsymbol{e}_{\bar{i}} - \boldsymbol{e}_{\bar{j}})\boldsymbol{e}_{\bar{k}}^T, \boldsymbol{W} \rangle
\end{aligned}$$

which contradicts with the fact that $\|\boldsymbol{u}\|_2 > 0$. This completes the proof. □

D. Proof Theorems in Section 4

In this section, we first share supplementary lemmas for the proof of Theorem 4.2.

D.1. Supplementary Lemmas for Proof of Theorem 4.2

Definition D.1. Let $\boldsymbol{B}(\boldsymbol{W}_\star, r) \subset \mathbb{R}^{d\times d}$ be a ball centered at a point $\boldsymbol{W}_\star$ with radius r defined as follows:

$$\boldsymbol{B}(\boldsymbol{W}_\star, r) = \{\boldsymbol{W} \in \mathcal{S}_{\boldsymbol{E}} \quad | \quad \|\boldsymbol{W} - \boldsymbol{W}_\star\|_F \le r\}.$$

Lemma D.2. *Suppose that Assumption 2.3 holds. Then, for any (X, y) where the token ID y exists in X, and $\boldsymbol{W} \in \boldsymbol{B}(\boldsymbol{W}_\star, r)$ the absolute loss difference satisfies the following:*

$$|\ell(\boldsymbol{c}_y^\top \boldsymbol{X}^\top \mathbb{S}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{x}_L)) - \ell(\boldsymbol{c}_y^\top \boldsymbol{X}^\top \mathbb{S}(\boldsymbol{X}\boldsymbol{W}_\star\boldsymbol{x}_L))| \le 2r \max_{i\in[K]} \|\boldsymbol{e}_i\|_2^2.$$

This implies that the loss function is $(2\max_{i\in[K]} \|\boldsymbol{e}_i\|_2^2)$-Lipschitz.

Proof. Let $\boldsymbol{x}^{occ}$ be the token occurrence vector, i.e., $\boldsymbol{x}_i^{occ}$ is the number of occurrences of the i-th token inside the input sequence $\boldsymbol{X}$. Let x_L be the last token ID of $\boldsymbol{X}$. Let $\boldsymbol{u} \in \mathbb{R}^K$ be the vector such that $u_k = \boldsymbol{e}_k^\top \boldsymbol{W} \boldsymbol{e}_{x_L}$ for $k \in [K]$. Then, the loss function will be the following:

$$\ell(\boldsymbol{c}_y^\top \boldsymbol{X}^\top \mathbb{S}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{x}_L)) = -\log\left(\frac{\boldsymbol{x}_y^{occ}\mathrm{e}^{u_y}}{\sum_{i=1}^K \boldsymbol{x}_i^{occ}\mathrm{e}^{u_i}}\right) = -\log(\boldsymbol{x}_y^{occ}) - u_y + \log\left(\sum_{i=1}^K \boldsymbol{x}_i^{occ}\mathrm{e}^{u_i}\right)$$

Let $\boldsymbol{u}^*$ be the vector such that $\boldsymbol{u}_k^* = \boldsymbol{e}_k^\top \boldsymbol{W}_\star \boldsymbol{e}_{x_L}$ for $k \in [K]$. Then, the loss difference will be the following:

$$\begin{aligned}\ell(\boldsymbol{c}_y^\top \boldsymbol{X}^\top \mathbb{S}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{x}_L)) - \ell(\boldsymbol{c}_y^\top \boldsymbol{X}^\top \mathbb{S}(\boldsymbol{X}\boldsymbol{W}_\star\boldsymbol{x}_L)) &= \log\left(\sum_{i=1}^K \boldsymbol{x}_i^{occ}\mathrm{e}^{u_i}\right) - u_y - \log\left(\sum_{i=1}^K \boldsymbol{x}_i^{occ}\mathrm{e}^{u_i^*}\right) + u_y^* \\ &\le \left|\left(\sum_{i=1}^K \boldsymbol{x}_i^{occ}\mathrm{e}^{u_i}\right) - \log\left(\sum_{i=1}^K \boldsymbol{x}_i^{occ}\mathrm{e}^{u_i^*}\right)\right| + |u_y - u_y^*|\end{aligned}$$

Let $f : \mathbb{R}^K \to \mathbb{R}$ be defined as $f(\boldsymbol{u}) = \log\left(\sum_{i=1}^K \boldsymbol{x}_i^{occ}\mathrm{e}^{u_i}\right)$. Then, the derivative of the function f is the following:

$$\nabla_j f(\boldsymbol{u}) = \frac{\boldsymbol{x}_j^{occ}\mathrm{e}^{u_j}}{\sum_{i=1}^K \boldsymbol{x}_i^{occ}\mathrm{e}^{u_i}} \qquad \forall j \in [K]$$

Note that f is a continuous function. Using the Intermediate Value Theorem, for any $\boldsymbol{u}, \boldsymbol{u}^*$, there exists $\boldsymbol{v} \in \mathbb{R}^K$ such that $\boldsymbol{v} \in [\boldsymbol{u}, \boldsymbol{u}^*]$ and it satisfies the following:

$$|f(\boldsymbol{u}) - f(\boldsymbol{u}^*)| = \nabla f(\boldsymbol{v})^\top (\boldsymbol{u} - \boldsymbol{u}^*) \le \sum_{j=1}^K \nabla_j f(\boldsymbol{v}) \|\boldsymbol{u} - \boldsymbol{u}^*\|_\infty = \|\boldsymbol{u} - \boldsymbol{u}^*\|_\infty$$

As a result, we obtain that

$$\begin{aligned}|\ell(\boldsymbol{c}_y^\top \boldsymbol{X}^\top \mathbb{S}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{x}_L)) - \ell(\boldsymbol{c}_y^\top \boldsymbol{X}^\top \mathbb{S}(\boldsymbol{X}\boldsymbol{W}_\star\boldsymbol{x}_L))| &\le \left|\left(\sum_{i=1}^K \boldsymbol{x}_i^{occ}\mathrm{e}^{u_i}\right) - \log\left(\sum_{i=1}^K \boldsymbol{x}_i^{occ}\mathrm{e}^{u_i^*}\right)\right| + |u_y - u_y^*| \\ &\le 2\|\boldsymbol{u} - \boldsymbol{u}^*\|_\infty \\ &= 2\max_{i\in[K]} \boldsymbol{e}_i^\top (\boldsymbol{W} - \boldsymbol{W}_\star)\boldsymbol{e}_{x_L} \\ &\le 2r \max_i \|\boldsymbol{e}_i\|_2^2\end{aligned}$$

□

Lemma D.3. *Let $\boldsymbol{W} \in \boldsymbol{B}(\boldsymbol{W}_\star, r)$ be an arbitrary attention matrix. Then, for any $\boldsymbol{W}_\star \in \mathbb{R}^{d\times d}$ and $\delta > 0$, we have the following with probability at least $1 - 2\delta$*

$$|\mathcal{L}(\boldsymbol{W}) - \mathcal{L}(\boldsymbol{W}_\star) - \widehat{\mathcal{L}}(\boldsymbol{W}) + \widehat{\mathcal{L}}(\boldsymbol{W}_\star)| < r \max_{i\in[K]} \|\boldsymbol{e}_i\|_2^2 \sqrt{\frac{8\log(1/\delta)}{n}}$$

Proof. We are going to utilize McDiarmid's Inequality. First, we check whether the assumption of McDiarmid's Inequality is satisfied. Let $\mathcal{T} = ((\boldsymbol{X}_i, \boldsymbol{x}_{i,L_i}, y_i))_{i=1}^n$ be the dataset. Let $\mathcal{X}_i$ be the sample space of $(\boldsymbol{X}_i, \boldsymbol{x}_{i,L_i}, y_i)$. Let the function $f : \mathcal{X}_1 \times \mathcal{X}_2 \times \dots \mathcal{X}_n \to \mathbb{R}$ be defined as follows:

$$f(\mathcal{T}) := \widehat{\mathcal{L}}(\boldsymbol{W}) - \widehat{\mathcal{L}}(\boldsymbol{W}_\star) = \frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{c}_{y_i}^\top \boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{X}_i \boldsymbol{W} \boldsymbol{x}_{i,L_i})) - \frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{c}_{y_i}^\top \boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{X}_i \boldsymbol{W}_\star \boldsymbol{x}_{i,L_i}))$$

Let $\mathcal{T}' = ((\boldsymbol{X}_i', \boldsymbol{x}_{i,L_i}', y_i'))_{i=1}^n$ be the dataset such that the samples are different for only the j^{th} sample from $\mathcal{T}$. In other words, if $j \neq i$, then $(\boldsymbol{X}_i', \boldsymbol{x}_{i,L_i}', y_i') = (\boldsymbol{X}_i, \boldsymbol{x}_{i,L_i}, y_i)$. Then, we are going to look at the following difference:

$$\begin{aligned}
|f(\mathcal{T}) - f(\mathcal{T}')| = \frac{1}{n}\Big| & \ell(\boldsymbol{c}_{y_j}^\top \boldsymbol{X}_j^\top \mathbb{S}(\boldsymbol{X}_j \boldsymbol{W} \boldsymbol{x}_{Lj})) - \ell(\boldsymbol{c}_{y_j}^\top \boldsymbol{X}_j^\top \mathbb{S}(\boldsymbol{X}_j \boldsymbol{W}_\star \boldsymbol{x}_{Lj})) \\
& - \Big(\ell(\boldsymbol{c}_{y_j'}^\top \boldsymbol{X}_j'^\top \mathbb{S}(\boldsymbol{X}_j' \boldsymbol{W} \boldsymbol{x}_{Lj'})) - \ell(\boldsymbol{c}_{y_j'}^\top \boldsymbol{X}_j'^\top \mathbb{S}(\boldsymbol{X}_j' \boldsymbol{W}_\star \boldsymbol{x}_{Lj}'))\Big)\Big| \\
\le \frac{1}{n}\Big| & \ell(\boldsymbol{c}_{y_j}^\top \boldsymbol{X}_j^\top \mathbb{S}(\boldsymbol{X}_j \boldsymbol{W} \boldsymbol{x}_{Lj})) - \ell(\boldsymbol{c}_{y_j}^\top \boldsymbol{X}_j^\top \mathbb{S}(\boldsymbol{X}_j \boldsymbol{W}_\star \boldsymbol{x}_{Lj}))\Big| \\
& + \frac{1}{n}\Big|\ell(\boldsymbol{c}_{y_j'}^\top \boldsymbol{X}_j'^\top \mathbb{S}(\boldsymbol{X}_j' \boldsymbol{W} \boldsymbol{x}_{Lj}')) - \ell(\boldsymbol{c}_{y_j'}^\top \boldsymbol{X}_j'^\top \mathbb{S}(\boldsymbol{X}_j' \boldsymbol{W}_\star \boldsymbol{x}_{Lj}'))\Big| \\
\overset{(a)}{\le} & \frac{4r \max_{i\in[K]} \|\boldsymbol{e}_i\|_2^2}{n}
\end{aligned}$$

where (a) follows from Lemma D.2. Now, using McDiarmid's Inequality, we obtain the following for any $\varepsilon > 0$:

$$\begin{aligned}
\mathbb{P}\left(|\,\mathbb{E}[\widehat{\mathcal{L}}(\boldsymbol{W})] - \mathbb{E}[\widehat{\mathcal{L}}(\boldsymbol{W}_\star)] - \widehat{\mathcal{L}}(\boldsymbol{W}) + \widehat{\mathcal{L}}(\boldsymbol{W}_\star)| > \varepsilon\right) &\le 2\exp\left(\frac{2\varepsilon^2}{\sum_{j=1}^n \frac{16r^2 \max_{i\in[K]} \|\boldsymbol{e}_j\|_2^4}{n^2}}\right) \\
&= 2\exp\left(\frac{n\varepsilon^2}{8r^2 \max_{i\in[K]} \|\boldsymbol{e}_i\|_2^4}\right)
\end{aligned}$$

This completes the proof. □

Lemma D.4. *For any $\boldsymbol{W}_\star \in \mathbb{R}^{d\times d}$, and $\delta > 0$, we have the following with probability at least $1 - 2\delta$*

$$\begin{aligned}
\sup_{\boldsymbol{W}\in\boldsymbol{B}(\boldsymbol{W}_\star, r)} & |\mathcal{L}(\boldsymbol{W}) - \mathcal{L}(\boldsymbol{W}_\star) - \widehat{\mathcal{L}}(\boldsymbol{W}) + \widehat{\mathcal{L}}(\boldsymbol{W}_\star)| \\
& \le \inf_{r>\varepsilon>0}\left\{2\varepsilon \max_{i\in[K]} \|\boldsymbol{e}_i\|_2^2 + r \max_{i\in[K]} \|\boldsymbol{e}_i\|_2^2 \sqrt{\frac{8K^2 \log((r+2/\varepsilon)/\delta)}{n}}\right\}
\end{aligned}$$

Proof. For any $\varepsilon > 0$, let $\mathcal{A}_\varepsilon : \mathcal{N}(\mathcal{S}_{\boldsymbol{E}}, \varepsilon)$ be the minimal ε cover of $\mathcal{S}_{\boldsymbol{E}} \cap \boldsymbol{B}(\boldsymbol{W}_\star, r)$ in terms of the Frobenius norm. Let the function $g : \mathcal{S}_{\boldsymbol{E}} \times \mathcal{S}_{\boldsymbol{E}} \to \mathbb{R}$ be defined as $g(\boldsymbol{W}, \boldsymbol{W}') = \mathcal{L}(\boldsymbol{W}) - \mathcal{L}(\boldsymbol{W}') - (\widehat{\mathcal{L}}(\boldsymbol{W}) - \widehat{\mathcal{L}}(\boldsymbol{W}'))$. Then, we have the following:

$$\sup_{\boldsymbol{W}\in\boldsymbol{B}(\boldsymbol{W}_\star, r)} |g(\boldsymbol{W}, \boldsymbol{W}_\star)| \le \sup_{\boldsymbol{W}\in\boldsymbol{B}(\boldsymbol{W}_\star, r)} \min_{\boldsymbol{W}'\in\mathcal{A}_\varepsilon} |g(\boldsymbol{W}, \boldsymbol{W}')| + \max_{\boldsymbol{W}\in\mathcal{A}_\varepsilon} |g(\boldsymbol{W}, \boldsymbol{W}_\star)|$$

By definition, there exists $\boldsymbol{W}' \in \mathcal{A}_\varepsilon$ such that $\boldsymbol{W}' \in \boldsymbol{B}(\boldsymbol{W}, \varepsilon)$. Then, using Lemma D.2, we have the following:

$$\sup_{\boldsymbol{W}\in\boldsymbol{B}(\boldsymbol{W}_\star, r)} \min_{\boldsymbol{W}'\in\mathcal{A}_\varepsilon} |g(\boldsymbol{W}, \boldsymbol{W}')| \le 2\varepsilon \max_{i\in[K]} \|\boldsymbol{e}_i\|_2^2$$

On the other hand, note that the cardinality of $\mathcal{A}_\varepsilon$ is finite and it is upper bounded by $(r + 2/\varepsilon)^{K^2}$ from Corollary 4.2.13 (Vershynin, 2018). Then, we apply union bound to Lemma D.3 and obtain the following with probability at least $1 - 2\delta$:

$$\begin{aligned}\max_{\boldsymbol{W}\in\mathcal{A}_\varepsilon} |g(\boldsymbol{W}, \boldsymbol{W}_\star)| &\le r \max_{i\in[K]} \|\boldsymbol{e}_i\|_2^2 \sqrt{\frac{8\log(|\mathcal{A}_\varepsilon|/\delta)}{n}} \\ &\le r \max_{i\in[K]} \|\boldsymbol{e}_i\|_2^2 \sqrt{\frac{8K^2\log((r+2/\varepsilon)/\delta)}{n}}\end{aligned}$$

Combining all of the results, for any $\varepsilon > 0$, we derive the following with probability at least $1 - 2\delta$

$$\sup_{\boldsymbol{W}\in\boldsymbol{B}(\boldsymbol{W}_\star, r)} |g(\boldsymbol{W}, \boldsymbol{W}_\star)| \le 2\varepsilon \max_{i\in[K]} \|\boldsymbol{e}_i\|_2^2 + r \max_{i\in[K]} \|\boldsymbol{e}_i\|_2^2 \sqrt{\frac{8K^2\log((r+2/\varepsilon)/\delta)}{n}}$$

which completes the proof. □

D.2. Proof of Theorem 4.2

Theorem D.5 (Restated Theorem 4.2)**.** *Suppose Assumptions 2.3 and 4.1 hold. Let $R_0 > 0$ be a finite constant based on the structure of $\boldsymbol{W}^{GT}$ and $\mathcal{D}_{\mathcal{X}}$. Then, if $n \ge R_0K^2$, with probability at least $1 - 2\delta$*

$$\mathcal{L}(\hat{\boldsymbol{W}}_n) - \mathcal{L}(\boldsymbol{W}_\star) \lesssim \frac{K^2 \log \frac{n}{K\delta}}{n}.$$

Proof. Let $r_0 \in \mathbb{R}$ and consider the ball $\boldsymbol{B}(\boldsymbol{W}_\star, r_0)$. From Lemma C.6, we know that $\mathcal{L}(\boldsymbol{W})$ is strictly convex on $\mathcal{S}_{\boldsymbol{E}}$. Additionally, note that the function $\mathcal{L}(\boldsymbol{W})$ is differentiable twice and its second derivative is continuous. As the $\boldsymbol{B}(\boldsymbol{W}_\star, r_0)$ is a compact set, there exists a positive constant $\alpha > 0$ such that the eigenvalues of the Hessian of $\mathcal{L}(\boldsymbol{W})$ are lower bound by α. It means that the loss function $\mathcal{L}(\boldsymbol{W})$ is $\alpha-$strongly convex in the ball $\boldsymbol{B}(\boldsymbol{W}_\star, r_0)$.

Now, let $\boldsymbol{e}_{max} := \max_{i\in[K]} \|\boldsymbol{e}_i\|_2$ and let's set the following values:

$$\varepsilon = \frac{2K\boldsymbol{e}_{max}^2}{\alpha n} \tag{30}$$

$$D_\delta = \sqrt{8K^2\log(3/(\varepsilon\delta))} \tag{31}$$

$$r = \frac{2^5 D_\delta \boldsymbol{e}_{max}^2}{\alpha\sqrt{n}} \tag{32}$$

$$n \ge \frac{2^{11} D_\delta^2 \boldsymbol{e}_{max}^4}{\alpha^2(\min(r_0, 1))^2} \tag{33}$$

From (30) and (32), we have $\varepsilon < r/\sqrt{n}$, which shows that we can utilize this ε in Lemma D.4. From (32) and (33), we have $r \le \min(r_0, 1)$, which implies that the loss function $\mathcal{L}(\boldsymbol{W})$ is $\alpha-$strongly convex inside the ball $\boldsymbol{B}(\boldsymbol{W}_\star, r)$. We are going to show that $\hat{\boldsymbol{W}}_n \in \boldsymbol{B}(\boldsymbol{W}_\star, r)$ if n satisfies (33). Let $\boldsymbol{W}_{out}$ be an arbitrary point on the boundary of the ball $\boldsymbol{B}(\boldsymbol{W}_\star, r)$. Let $\boldsymbol{W}_{inn}$ be the boundary point of the ball $\boldsymbol{B}(\boldsymbol{W}, r/2)$ such that it is on the line segment between $\boldsymbol{W}_\star$ and $\boldsymbol{W}_{out}$. By the strong convexity of $\mathcal{L}(\boldsymbol{W})$ on $\boldsymbol{B}(\boldsymbol{W}_\star, r)$, we have the following:

$$\mathcal{L}(\boldsymbol{W}_{out}) - \mathcal{L}(\boldsymbol{W}_{inn}) \ge \frac{\alpha r^2}{8} \tag{34}$$

We apply Lemma D.4 to both $\boldsymbol{W}_{inn}$ and $\boldsymbol{W}_{out}$. As $r \le \min(r_0, 1)$, we have $\log((r + 2/\varepsilon)/\delta) < \log(3/(\varepsilon\delta))$. Then, we obtain the following for any $\varepsilon > 0$ with probability at least $1 - 2\delta$

$$\begin{aligned}|\mathcal{L}(\boldsymbol{W}_{inn}) - \mathcal{L}(\boldsymbol{W}_\star) - \widehat{\mathcal{L}}(\boldsymbol{W}_{inn}) + \widehat{\mathcal{L}}(\boldsymbol{W}_\star)| &\le 2\varepsilon\boldsymbol{e}_{max}^2 + r\boldsymbol{e}_{max}^2 \sqrt{\frac{8K^2\log((r+2/\varepsilon)/\delta)}{n}} \\ &\overset{(a)}{\le} \frac{4D_\delta r \boldsymbol{e}_{max}^2}{\sqrt{n}} \end{aligned} \tag{35}$$

where (a) follows from the fact that $\varepsilon \leq D_\delta r/\sqrt{n}$. Similarly, we have the following with probability at least $1-2\delta$

$$|\mathcal{L}(\boldsymbol{W}_{out}) - \mathcal{L}(\boldsymbol{W}_\star) - \widehat{\mathcal{L}}(\boldsymbol{W}_{out}) + \widehat{\mathcal{L}}(\boldsymbol{W}_\star)| \leq 2\varepsilon \boldsymbol{e}_{max}^2 + r\boldsymbol{e}_{max}^2 \sqrt{\frac{8K^2 \log((r+2/\varepsilon)/\delta)}{n}} \tag{36}$$

$$\leq \frac{4D_\delta r \boldsymbol{e}_{max}^2}{\sqrt{n}} \tag{37}$$

Combining (34) and (37), we obtain the following for any $\varepsilon > 0$

$$\widehat{\mathcal{L}}(\boldsymbol{W}_{out}) \geq \mathcal{L}(\boldsymbol{W}_{inn}) - \mathcal{L}(\boldsymbol{W}_\star) + \widehat{\mathcal{L}}(\boldsymbol{W}_\star) + \frac{\alpha r^2}{8} - \left(\frac{4D_\delta r \boldsymbol{e}_{max}^2}{\sqrt{n}}\right) \tag{38}$$

From (35), we obtain the following:

$$\widehat{\mathcal{L}}(\boldsymbol{W}_{inn}) \leq \mathcal{L}(\boldsymbol{W}_{inn}) - \mathcal{L}(\boldsymbol{W}_\star) + \widehat{\mathcal{L}}(\boldsymbol{W}_\star) + \left(\frac{4D_\delta r \boldsymbol{e}_{max}^2}{\sqrt{n}}\right) \tag{39}$$

From (31), (32), and (33), we have that

$$\frac{\alpha r^2}{4} > \frac{8D_\delta r \boldsymbol{e}_{max}^2}{\sqrt{n}} \tag{40}$$

Combining (38), (39), and (40), we obtain the following

$$\widehat{\mathcal{L}}(\boldsymbol{W}_{out}) > \widehat{\mathcal{L}}(\boldsymbol{W}_{inn}) \tag{41}$$

Note that (41) is valid for any boundary point of $\boldsymbol{B}(\boldsymbol{W}_\star, r)$. Due to the convexity of $\widehat{\mathcal{L}}(\boldsymbol{W})$ from Lemma C.6, $\hat{\boldsymbol{W}}_n$ should be inside the ball $\boldsymbol{B}(\boldsymbol{W}_\star, r)$ with probability at least $1-2\delta$ as a result of (41). Now, let's use the smoothness of the loss function $\mathcal{L}(\boldsymbol{W})$. From (Li et al., 2024), we know that $\mathcal{L}(\boldsymbol{W})$ is $2\boldsymbol{e}_{max}^2\sqrt{L_{max}}-$smooth. As $\hat{\boldsymbol{W}}_n$ is inside the ball $\boldsymbol{B}(\boldsymbol{W}_\star, r)$, we have the following with at least probability $1-2\delta$:

$$\mathcal{L}(\hat{\boldsymbol{W}}_n) - \mathcal{L}(\boldsymbol{W}) \leq \frac{\boldsymbol{e}_{max}\sqrt{L_{max}}r^2}{2} \overset{(a)}{\lesssim} \frac{K^2 \log\frac{n}{K\delta}}{n} \tag{42}$$

where (a) follows from (32). This completes the proof. □

D.3. Proof of Corollary 4.3

Corollary D.6 (Restated Corollary 4.3)**.** *Consider the setting in Theorem 4.2 and suppose Assumptions 2.3, 3.2, and 4.1 hold. Then, if $n \geq R_0 K^2$, with probability at least $1-2\delta$*

$$\|\hat{\boldsymbol{W}}_n - \boldsymbol{W}^{\text{GT}}\|_F^2 \lesssim \frac{K^2 \log\frac{n}{K\delta}}{n}. \tag{43}$$

Proof. We show in the proof of Theorem D.5 that with probability at least $1-2\delta$ that $\hat{\boldsymbol{W}}_n \in \boldsymbol{B}(\boldsymbol{W}_\star, r)$.

With the same argument in the proof of Theorem D.5, we have the strong convexity inside $\boldsymbol{B}(\boldsymbol{W}^\star, r_0)$: From Lemma C.6, we know that $\mathcal{L}(\boldsymbol{W})$ is strictly convex on $\mathcal{S}_{\boldsymbol{E}}$. Additionally, note that the function $\mathcal{L}(\boldsymbol{W})$ is differentiable twice and its second derivative is continuous. As the $\boldsymbol{B}(\boldsymbol{W}_\star, r_0)$ is a compact set, there exists a positive constant $\alpha > 0$ such that the eigenvalues of the Hessian of $\mathcal{L}(\boldsymbol{W})$ are lower bound by α. It means that the loss function $\mathcal{L}(\boldsymbol{W})$ is $\alpha-$strongly convex in the ball $\boldsymbol{B}(\boldsymbol{W}_\star, r_0)$. Recall that $r \leq \min(r_0, 1)$ from (30), (31), (32), and (33). Therefore, the loss function $\mathcal{L}(\boldsymbol{W})$ is $\alpha-$strongly convex in the $\boldsymbol{B}(\boldsymbol{W}_\star, r)$ as well.

By $\alpha-$strong convexity of $\mathcal{L}(\boldsymbol{W})$ on $\boldsymbol{B}(\boldsymbol{W}_\star, r)$, we obtain that

$$\mathcal{L}(\boldsymbol{W}_\star) - \mathcal{L}(\hat{\boldsymbol{W}}_n) \geq \frac{\alpha\|\boldsymbol{W}_\star - \hat{\boldsymbol{W}}_n\|_F}{2}$$

In addition to that, the estimation of $\boldsymbol{W}^{\text{GT}}$ is consistent by Theorem 3.4 using Assumption 4.1. This implies that $\boldsymbol{W}^{\text{GT}} = \boldsymbol{W}_\star$. This completes the proof. □

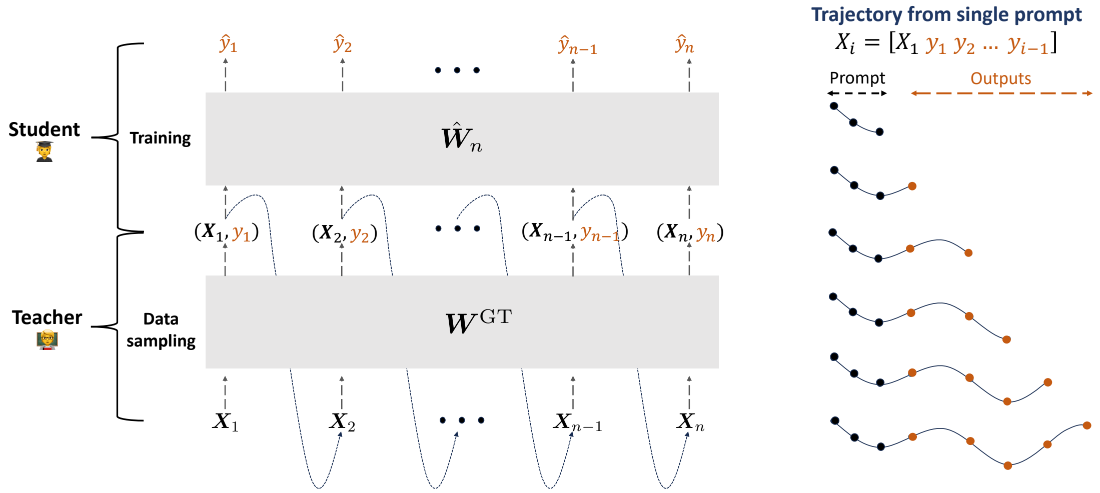

Trajectory from single prompt
$X_i = [X_1\ y_1\ y_2 \dots y_{i-1}]$
Prompt
Outputs
Student
Training
Teacher
Data sampling
$\hat{y}_1$
$\hat{y}_2$
$\hat{y}_{n-1}$
$\hat{y}_n$
$\hat{\boldsymbol{W}}_n$
$(\boldsymbol{X}_1, y_1)$
$(\boldsymbol{X}_2, y_2)$
$(\boldsymbol{X}_{n-1}, y_{n-1})$
$(\boldsymbol{X}_n, y_n)$
$\boldsymbol{W}^{\mathrm{GT}}$
$\boldsymbol{X}_1$
$\boldsymbol{X}_2$
$\boldsymbol{X}_{n-1}$
$\boldsymbol{X}_n$

Figure 8. Illustration of single-trajectory learning where next tokens are sampled from a single trajectory. The setting is analogous to asking language models a broad question and constantly collecting the responses.

E. Proof of Theorems in Section 5

E.1. Proof of lemma 5.2

Definition E.1. Given sample space $\Omega = \{0,1\}^\infty$ and let $A_{k,t} = \mathbf{1}(y_t = k)$. Let $\mathcal{F}_{k,t}, t \geq 0$ be a filtration with $\mathcal{F}_{k,0} = \{\emptyset, \Omega\}$ and $\mathcal{F}_{k,t} := \sigma(A_{k,j} | j \in [t])$, where σ refers to the σ-algebra.

Lemma E.2 (Restated Lemma 5.2)**.** *Let $\boldsymbol{P}^{GT}$ be a transition matrix with non-zero entries. For all $k \in [K]$, $\mathbb{P}(\lim_{n\to\infty} S_{k,n} = \infty) = 1$.*

Proof. Let $\bar{\pi}_k = \min_{i\in[K]} \pi_{ik}$. Consider any token $k \in [K]$, Then we have:

$$\Pr(A_{k,t}|\mathcal{F}_{k,t-1}) \overset{(a)}{=} \mathbb{P}(y_t = k) = \sum_{i=1}^{K} \mathbb{P}(y_t = k|\bar{x}_t = i)p(\bar{x}_t = i) \tag{44}$$

To proceed, consider a single trajectory starting from $X_1 = [K]$, then at time t, the length of X_t is $t + K - 1$. The minimum of $\mathbb{P}(y_t = k|\bar{x}_t = i)$ can then be written as

$$\begin{aligned}
\min_{i\in[K]} \mathbb{P}(y_t = k|\bar{x}_t = i) &= \min_{i\in[K]} \frac{\pi_{ik} S_{k,t}}{S_{k,t}\pi_{ik} + \sum_{j\in[K], j\neq k} S_{j,t}\pi_{ij}} \\
&\geq \min_{i\in[K]} \frac{\pi_{ik} S_{k,t}}{S_{k,t}\pi_{ik} + (t + K - S_{k,t} - 1)(1 - \pi_{ik})} \\
&= \min_{i\in[K]} \frac{\pi_{ik} S_{k,t}}{S_{k,t}(2\pi_{ik} - 1) + (t + K - 1)(1 - \pi_{ik})} \\
&= \min_{i\in[K]} \frac{\pi_{ik}}{(2\pi_{ik} - 1) + \frac{(t+K-1)(1-\pi_{ik})}{S_{k,t}}} \\
&\geq \min_{i\in[K]} \frac{\pi_{ik}}{(2\pi_{ik} - 1) + (t + K - 1)(1 - \pi_{ik})} \\
&= \min_{i\in[K]} \frac{\pi_{ik}}{t + K - 2 + (3 - K - t)\pi_{ik}} \\
&\overset{(a)}{\geq} \frac{\bar{\pi}_k}{t}
\end{aligned} \tag{45}$$

where (a) holds for sufficiently large t where $t \gtrsim K/\bar{\pi}_k$. Then we have:

$$\Pr(A_{k,t}|\mathcal{F}_{k,t-1}) \geq \frac{\bar{\pi}_k}{t} \sum_{i=1}^{K} p(\bar{x}_t = i) = \frac{\bar{\pi}_k}{t} \tag{46}$$

Hence, the sum of the probabilities over infinite iterations diverges to infinity, i.e., $\sum_{t=1}^{\infty} \Pr(A_{k,t}|\mathcal{F}_{k,t-1}) = \infty$. Using the second Borel-Cantelli lemma, we have:

$$\Pr\Big(\limsup_{t\to\infty} A_{k,t}\Big) = 1 \tag{47}$$

which reveals that token k can be observed infinitely many times in the trajectory. This implies $\mathbb{P}(\lim_{n\to\infty} S_{k,n} = \infty) = 1$ where $S_{k,n} = \sum_{t=1}^{n} A_{k,t}$. □

E.2. Proof of lemma 5.3

Lemma E.3 (Restated Distribution collapse Lemma 5.3)**.** *Consider the CCMC model with $K = 2$ defined in Section 5.1. Suppose that $\boldsymbol{X}_1$ includes all vocabulary at least once. Recall that $\boldsymbol{m}(X_t)$ denotes the empirical frequency of individual states where X_t is the state trajectory at time t. For any $t > t_0$ with a sufficiently large t_0, we have:*

$$\mathbb{E}[\boldsymbol{m}(X_t)_2] < t^{-q}$$

where $q = 1 - p/(1-p)$. Furthermore, when $p < 1/2$,

$$\lim_{t\to\infty} \mathbb{E}\left[\frac{\boldsymbol{m}(X_t)_2}{\boldsymbol{m}(X_t)_1}\right] = 0.$$

Proof. Let S_t and L_t be the number of token 2 in the input prompt and the length of the input prompt at iteration t, respectively. Note that $S_1 \geq 1$. Then, at iteration t, we have:

$$\begin{aligned}\mathbb{E}[S_t] &= S_{t-1} + \frac{S_{t-1}p}{S_{t-1}p + (L_t - S_{t-1})(1-p)} \\ &= S_{t-1}(1 + \frac{p}{S_{t-1}(2p-1) + L_t(1-p)})\end{aligned} \tag{48}$$

Since the probability of selecting a specific token is weighted by the number of occurrences, when $p < 1/2$, the model tends to sample token 1 over 2. Moreover, due to the positive reinforcement nature, selecting token 1 as the label will further increase the probability of selecting token 1 in the next round. Thus, for any $p < 1/2$, there exists a sufficiently large t_0 such that when $t > t_0, L_t \gg S_{t-1}$. Hence:

$$\begin{aligned}\mathbb{E}[S_t] &= S_{t-1}(1 + \frac{p}{S_{t-1}(2p-1) + L_t(1-p)}) \\ &\approx S_{t-1}(1 + \frac{p}{L_t(1-p)}) \\ &= S_{t-1}(1 + \bar{p}/L_t)\end{aligned} \tag{49}$$

where $\bar{p} = p/(1-p)$. Recall that $\boldsymbol{m}(X_t)_2 = \frac{S_t}{L_t}$. For $t > t_0$ where t_0 is sufficiently large, we have:

$$\begin{aligned}\mathbb{E}[\boldsymbol{m}(X_t)_2] &= \mathbb{E}\left[\frac{S_t}{L_t}\right] \\ &= \mathbb{E}\left[\frac{S_{t-1}(1+\bar{p}/L_t)}{L_{t-1}(1+1/L_{t-1})}\right] \\ &= \mathbb{E}[\boldsymbol{m}(X_{t-1})_2] \cdot \frac{1+\bar{p}/L_t}{1+1/L_{t-1}} \\ &\approx \mathbb{E}[\boldsymbol{m}(X_{t-1})_2](1 - q/L_t)\end{aligned} \tag{50}$$

where $q \coloneqq 1 - \bar{p} = (1-2p)/(1-p)$. Applying this equation recursively, we get:

$$\mathbb{E}\left[\frac{\boldsymbol{m}(X_t)_2}{\boldsymbol{m}(X_1)_2}\right] = \prod_{\tau=1}^{t}(1 - q/\tau) \overset{(a)}{\leq} \prod_{\tau=1}^{t}\exp(-q/\tau) = \exp(-q\sum_{\tau=1}^{t} 1/\tau) \overset{(b)}{<} \exp(-q\ln t) = t^{-q} \tag{51}$$

where (a) comes from $1 + x \leq e^x$ for any x and (b) comes from the fact that $\sum_{\tau=1}^{t} 1/\tau > \ln t$ when t is large. By Euler-Maclaurin formula, $H_t = \sum_{\tau=1}^{t} 1/\tau = \ln t + \gamma + 1/(2t) - \varepsilon_t \approx \ln t + \gamma \geq \ln t$ where $\gamma \approx 0.5772$. As a result, for a sufficiently big t, $\mathbb{E}[\boldsymbol{m}(X_t)_2] < t^{-q}$ as $\boldsymbol{m}(X_1)_2$ is a finite number. Moving forward, since $\mathbb{E}[\boldsymbol{m}(X_t)_2] + \mathbb{E}[\boldsymbol{m}(X_t)_1] = 1$, we have:

$$\mathbb{E}\left[\frac{\boldsymbol{m}(X_t)_2}{\boldsymbol{m}(X_t)_1}\right] = \mathbb{E}\left[\frac{\boldsymbol{m}(X_t)_2}{1-\boldsymbol{m}(X_t)_2}\right] = \frac{\mathbb{E}[\boldsymbol{m}(X_t)_2]}{1-\mathbb{E}[\boldsymbol{m}(X_t)_2]} < \frac{t^{-q}}{2-t^{-q}} = \frac{1}{2t^q - 1} \tag{52}$$

As $p < 1/2$, $q \in (0, 1)$, when $t \to \infty$, the ratio $\lim_{t\to\infty} \frac{1}{2t^q-1}$ goes to zero. Combining the fact that $\frac{\boldsymbol{m}(X_t)_2}{\boldsymbol{m}(X_t)_1} \geq 0$, it implies:

$$\lim_{t\to\infty} \mathbb{E}\left[\frac{\boldsymbol{m}(X_t)_2}{\boldsymbol{m}(X_t)_1}\right] = 0 \tag{53}$$

□

F. Further Related Work on Reinforcement Learning / Data-Driven Control

The coverage condition that we have found for the consistency of estimation is also related to the data coverage condition in offline reinforcement learning (Chen & Jiang, 2019; Xie & Jiang, 2020; Zhan et al., 2022; Jin et al., 2022; Foster et al., 2022; Rashidinejad et al., 2023). In these works, given a dataset collected according to offline policies, we wish to learn optimal policy, which raises a distribution shift challenge. Their statistical analysis relies on the data coverage conditions during dataset collection to withstand the distribution shift. Our work is related to these at a high level since we provide necessary and sufficient conditions on the input prompt distribution for the consistent estimation of a ground truth attention matrix. Furthermore, we establish finite sample complexity guarantees under the coverage conditions that provide the consistency of estimation.

Our work also relates to the literature on the statistical aspects of time-series prediction (Kuznetsov & Mohri, 2014; 2016; Simchowitz et al., 2018; Mohri & Rostamizadeh, 2008) and learning (non)linear dynamics (Dean et al., 2020; Ziemann & Tu, 2022; Dean et al., 2020; Tsiamis et al., 2022; Sarkar & Rakhlin, 2019; Sun et al., 2022; Mania et al., 2020; Oymak & Ozay, 2021; Block et al., 2023). Learning dynamical systems from a single trajectory has attracted significant attention in the recent literature (Ziemann et al., 2022; Oymak, 2019; Sattar & Oymak, 2022; Oymak & Ozay, 2019; Matni & Tu, 2019; Foster et al., 2020; Ziemann et al., 2024). As long as the stochastic process is mixing (e.g. ergodic Markov chain, stable dynamical system), the samples from the trajectory are approximately independent, and the underlying hypothesis can be learned under suitable assumptions (Yu, 1994). There is also a recent emphasis on mitigating the need for mixing (Simchowitz et al., 2018; Ziemann & Tu, 2022). Unlike dynamical systems, MDPs, or Markov chains, the token generation process of self-attention is non-Markovian as it depends on the whole past trajectory. Thus, our work initiates the statistical and consistency study of learning self-attention process by highlighting its unique nature and challenges.