How Transformers Learn Causal Structure with Gradient Descent

Eshaan Nichani [1] **Alex Damian** [1] **Jason D. Lee** [1]

Abstract

The incredible success of transformers on sequence modeling tasks can be largely attributed to the *self-attention* mechanism, which allows information to be transferred between different parts of a sequence. Self-attention allows transformers to encode causal structure which makes them particularly suitable for sequence modeling. However, the process by which transformers learn such causal structure via gradient-based training algorithms remains poorly understood. To better understand this process, we introduce an in-context learning task that requires learning latent causal structure. We prove that gradient descent on a simplified two-layer transformer learns to solve this task by encoding the latent causal graph in the first attention layer. The key insight of our proof is that the gradient of the attention matrix encodes the mutual information between tokens. As a consequence of the data processing inequality, the largest entries of this gradient correspond to edges in the latent causal graph. As a special case, when the sequences are generated from in-context Markov chains, we prove that transformers learn an *induction head* (Olsson et al., 2022). We confirm our theoretical findings by showing that transformers trained on our in-context learning task are able to recover a wide variety of causal structures.

1. Introduction

The transformer architecture (Vaswani et al., 2017) has revolutionized the field of deep learning, and has led to state-of-the art performance on tasks spanning language modeling (Brown et al., 2020), computer vision (Dosovitskiy et al., 2020), reinforcement learning (Chen et al., 2021), and the sciences (Jumper et al., 2021). The basic primitive of a transformer is a *self-attention* head, a sequence-to-sequence mapping in which each token in the output is a weighted linear combination of, i.e "attends to," the other tokens in the sequence. Prior work (Elhage et al., 2021) has sought to understand which specific computational operations, or "circuits," are implemented by self-attention layers in trained transformers. However, the process by which such circuits arise when transformers are trained from scratch via gradient-based algorithms is still unknown.

One hallmark capability of transformers is *in-context learning* (Brown et al., 2020), which is the ability to learn from information present in the input context without needing to update the model parameters. For example, given a prompt of input-output pairs, in-context learning is the ability to predict the output corresponding to a new input. Prior work has shown that this in-context learning ability relies on the existence of specific circuits called *induction heads* (Olsson et al., 2022). Given a prompt of the form $[\cdots, A, B, \cdots, A]$, an induction head copies the token which follows the previous occurrence of A, in this case being B. This can be implemented using two attention layers: the first performs the operation of "copying" the previous token, while the second compares this previous token to the last token of the context. By copying the previous token, the first attention layer thus implicitly encodes the causal structure of a Markov chain.

As another example, consider the setting of learning a function class in-context, introduced by Garg et al. (2022). Each prompt sequence is formed by sampling a new function f from some function class $\mathcal{F}$, and generating the prompt $[x_1, f(x_1), x_2, f(x_2), \dots, x_n, f(x_n), x_{test}]$ for i.i.d inputs $x_1, \dots, x_n, x_{test}$. The model must learn to estimate $f(x_{test})$ from the $(x_i, f(x_i))$ pairs given in-context. This setting has proved a useful testbed to understand which in-context learning algorithms can be implemented by transformers; see Section 1.2 for additional discussion. Such prompts also possess special causal structure. Conditioned on f, the $2k$-th token in the prompt only depends on the $(2k-1)$-th token, and is independent of the rest of the sequence. The model must thus learn to associate each $f(x_k)$, at position $2k$, with its corresponding x_k, at position $2k-1$. We view this instance as a problem with both a global causal structure, which comes from pairing x_k with $f(x_k)$, and an in-context transition which comes from the specific $f \in \mathcal{F}$ sampled for the sequence.

[1] Princeton University. Correspondence to: Eshaan Nichani <eshnich@princeton.edu>.

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

Transformers are clearly able to model causal structure, yet despite the necessity of doing so for performing in-context learning tasks, we still do not understand how such structures are learned by gradient descent when training from scratch. We thus ask the following question:

How do transformers learn causal structure with gradient descent?

1.1. Our Contributions

In this work, we analyze the gradient descent dynamics of an autoregressive two-layer attention-only transformer, and prove that it recovers latent causal structure. Our specific contributions are as follows:

- We introduce a novel family of in-context learning problems, which we call *random sequences with causal structure* (Task 2.4). The task fixes a latent causal structure, unknown to the transformer, and samples each sequence from a different distribution which respects the causal structure.
- When the latent causal graph is a tree, we prove that gradient descent on a simplified two-layer transformer solves this task by encoding the causal graph in the first attention layer in order to perform in-context estimation of the transition distribution (Theorem 4.4). As a special case of Theorem 4.4, we show that when the sequences are generated from in-context Markov chains, the transformer learns an induction head.
- The proof of Theorem 4.4 relies on showing that the gradient of the first attention layer automatically computes the χ^2-mutual information between pairs of tokens. As a result of the data processing inequality, the largest entries of this gradient correspond to edges in the latent causal graph, and hence the first attention layer converges to the adjacency matrix of this graph.
- When the causal graph is not a tree, we explicitly construct a multi-head transformer which solves this task by distributing the latent causal graph across many heads. We show empirically that transformers trained by gradient descent on this task learn our construction.

1.2. Related Work

In-Context Learning. Brown et al. (2020) demonstrated that GPT-3 can perform in-context learning, which has led to much subsequent work on understanding how such in-context learning ability arises. Xie et al. (2021) presents a Bayesian perspective on in-context learning by looking at the log-likelihood of out-of-distribution prompt sequences. Olsson et al. (2022) posits that in-context learning relies on the emergence of induction heads.

Garg et al. (2022) formalizes the setting of learning a function class in-context, and shows that transformers can be trained to in-context learn various simple function classes such as (sparse) linear models or shallow neural networks. This bears similarity to transformer neural processes (Nguyen & Grover, 2022), which recasts uncertainty-aware meta learning as an in-context learning task. Many recent works have sought to understand which in-context learning algorithms can be efficiently expressed by a transformer. Akyürek et al. (2023); Bai et al. (2023); Von Oswald et al. (2023) show that transformers can implement gradient descent to solve in-context linear regression, while Fu et al. (2023) constructs transformers which implement higher-order learning algorithms such as Newton's method. Giannou et al. (2023) constructs a transformer that can express general-purpose computational operations. However, these works are solely concerned with the representational capabilities of transformers, and do not answer the question of whether gradient descent indeed learns such constructions.

Training dynamics of transformers. Prior works have primarily studied the optimization dynamics of a single attention layer. Lu et al. (2021); Li et al. (2023) show that a single attention layer trained via gradient descent learns to encode topic structure. Snell et al. (2021) studies the dynamics of a simplified attention layer on sequence-to-sequence translation tasks. Tarzanagh et al. (2023b;a) show an equivalence between the optimization dynamics of a single attention layer and a certain SVM problem. Ahn et al. (2023) shows that the global optimum of a single linear attention layer trained on in-context linear regression implements a single step of preconditioned gradient descent. Mahankali et al. (2023); Zhang et al. (2023) study the optimization dynamics of a single linear attention layer for performing in-context linear regression. Huang et al. (2023) shows that gradient descent on a single softmax attention layer learns to solve linear regression in-context when the input data are orthogonal. Jelassi et al. (2022) analyzes the gradient descent dynamics of the position-position block of a single layer vision transformer, and shows that it converges to a solution encoding spacial structure. Tian et al. (2023a) studies the optimization dynamics of a single attention layer for a specific toy dataset, while Tian et al. (2023b) shows that jointly training a self-attention and MLP layer corresponds to the optimization dynamics of a certain modified MLP module. Boix-Adsera et al. (2023) shows that transformers with diagonal attention matrices display incremental learning behavior.

Bietti et al. (2023) studies a synthetic ICL task, and shows that an induction head is formed during training. They demonstrate heuristically that a few gradient steps on a modifed transformer architecture approximately learns the induction head. Our synthetic task handles more general causal structure, and requires attending to all prior instances of the final token rather than the most recent one. Furthermore, Bietti et al. (2023) requires the alphabet size S to be significantly larger than the context length T; we, however, assume that $T \gg S$, and our main theorem (Theorem 4.4) is an end-to-end guarantee on learning the causal structure and obtaining vanishing population loss.

Concurrent Work. A number of concurrent works also study the ability of transformers to solve synthetic in-context learning tasks. Reddy (2023) shows empirically that an induction head suddenly emerges during the training of a two-attention layer transformer on a specific in-context learning task. Akyürek et al. (2024) demonstrates that transformers can learn regular languages in-context, where each prompt consists of strings generated from a prompt-dependent formal language. This is due to the ability of transformers to compute in-context n-gram counts, via a generalization of the induction head mechanism using multiple attention layers. In Section 6, we show that a two-layer transformer with n heads can also compute such n-gram counts.

Edelman et al. (2024) study the formation of induction heads on the task of learning Markov chains in context, which is equivalent to our Task 2.4 when the latent causal graph is the chain graph. Their theoretical analysis focuses on two steps of gradient descent on a simplified linear transformer model, for Markov chains over two states. An interesting empirical observation of theirs is that transformers trained to learn n-grams in-context undergo a sequential learning procedure, by first predicting using the unigram counts, then the bigram counts, and so on.

2. Setup

2.1. Transformer Architecture

Let $[S]$ be a finite alphabet. Transformers are models mapping sequences $s_{1:T} := (s_1, \dots, s_T) \in [S]^T$ of length T to a length T sequence of vectors $z_1, \dots, z_T \in \mathbb{R}^{d_{out}}$. A transformer first embeds the sequence $s_{1:T}$ as a matrix $X = \left[x_1, x_2, \dots, x_T\right]^\top \in \mathbb{R}^{T\times d}$, where d is the embedding dimension. This is parameterized by the token embeddings $E \in \mathbb{R}^{d\times S}$ and positional embeddings $P \in \mathbb{R}^{d\times T}$:

$$\mathrm{embed}(s_{1:T}; (E, P))_i := Ee_{s_i} + Pe_i \text{ for } i = 1, \dots, T.$$

Transformers consist of two types of layers: attention layers and MLP layers. Throughout, we focus on decoder-based, attention-only transformers. These are models in which every layer is a *causal self-attention layer*, defined below:

Definition 2.1 (Causal self-attention head)**.** For a vector $v \in \mathbb{R}^k$, define the *softmax* function $\mathcal{S} : \mathbb{R}^k \to \mathbb{R}^k$ by $\mathcal{S}(v)_i := \frac{\exp(v_i)}{\sum_{j=1}^k \exp(v_j)}$. For a matrix $A \in \mathbb{R}^{d\times d}$, define the operator $\mathrm{attn}(\cdot; A) : \mathbb{R}^{T\times d} \to \mathbb{R}^{T\times d}$ by

$$\mathrm{attn}(h; A) := \mathcal{S}\big(\mathrm{MASK}(hAh^\top)\big)h \in \mathbb{R}^{T\times d}, \tag{1}$$

where $\mathrm{MASK}(M)_{i,j}$ is $M_{i,j}$ when $i \ge j$ and $-\infty$ otherwise, and the softmax function $\mathcal{S}$ is applied row-wise.

In Definition 2.1, the masking operator ensures tokens only attend to previous tokens in the sequence, and the softmax normalizes the output so that each row sums to 1. The amount that token i attends to token j, for $j \le i$, is thus $\mathcal{S}\big(\mathrm{MASK}(XAX^\top)\big)_{i,j} = \mathcal{S}\big(X_{\le i}A^\top x_i\big)_j$, where $X_{\le i} \in \mathbb{R}^{i\times d}$ is the submatrix formed by the first i rows of X.

A single attention head is parameterized by the tuple of $d\times d$ matrices (Q, K, V), referred to as the query, key, and value matrices, and maps X to the sequence $\mathrm{attn}(X; QK^\top)V^\top$.

A decoder-based transformer aggregates multiple causal self-attention heads over many layers:

Definition 2.2 (Decoder-based transformer)**.** Let L be the depth, $\{m_\ell\}_{\ell\in[L]}$ be the number of heads per layer, and d be the embedding dimension. For $\ell \in [L]$, $i \in [m_\ell]$, let $(Q_i^{(\ell)}, K_i^{(\ell)}, V_i^{(\ell)})$ be the query, key, and value matrices for the ith head in the ℓth layer. Let $W_O \in \mathbb{R}^{d_{out}\times d}$ be the output layer and let $E \in \mathbb{R}^{d\times S}$ and $P \in \mathbb{R}^{d\times T}$ be the token and positional embeddings respectively. Define the parameter vector $\theta := \{(Q_i^{(\ell)}, K_i^{(\ell)}, V_i^{(\ell)})\}_{\ell\in[L], i\in[m_\ell]} \cup \{E, P, W_O\}$. A decoder-based transformer $\mathrm{TF}_\theta : [S]^T \to \mathbb{R}^{T\times d_{out}}$ operates on $s_{1:T}$ by

$$\begin{aligned} h^{(0)} &= \mathrm{embed}(s_{1:T}; (E, P)) \\ h^{(\ell)} &= h^{(\ell-1)} + \sum_{i=1}^{m_\ell} \mathrm{attn}\left(h^{(\ell-1)}; Q_i^{(\ell)} K_i^{(\ell)\top}\right) V_i^{(\ell)\top} \\ \mathrm{TF}_\theta(s_{1:T}) &= h^{(L)} W_O^\top. \end{aligned} \tag{2}$$

We remark that $h^{(\ell)} \in \mathbb{R}^{T\times d}$ for $\ell = 0, \dots, L$.

Disentangled Transformer. Prior works on mechanistic interpretability have introduced the *residual stream* viewpoint to understand the behavior of trained transformers (Elhage et al., 2021). The residual stream exists as a memory and communication channel that various attention heads read and write to. Information in the residual stream is stored in low-dimensional subspaces of intermediate layers $h^{(\ell)}$. For a single attention layer $\mathrm{attn}(\cdot; QK^\top)V^\top$, the query and key matrices "read" information from the relevant subspace, and the value matrix "writes" the output to a new subspace of the residual stream. The weight matrices thus act as associative memories (Bietti et al., 2023), storing various embeddings within the residual stream.

While this residual stream viewpoint provides intuition for the flow of information through the forward pass of a transformer, from an interpretability perspective it is difficult to know which subspaces contain which information. The outputs of each attention layer are added together and thus their informations may overlap with each other, leading to a memory bottleneck (Elhage et al., 2021; Bietti et al., 2023). Friedman et al. (2023) thus consider a transformer model in which the residual stream is disentangled, and the outputs of each attention layer are *appended* to the residual stream. The dimension of the residual stream thus grows with the depth. We formalize this as a *disentangled transformer*, defined below:

Definition 2.3 (Disentangled Transformer)**.** Let L be the depth, and $\{m_\ell\}_{\ell \in [L]}$ be the number of heads per layer. Define the set of dimensions $d_0, \dots, d_L$ by $d_0 = S + T$ and $d_\ell = (1 + m_\ell) d_{\ell-1}$. Let $\{\widetilde{A}_i^\ell\}$ be the attention matrices with $\widetilde{A}_i^{(\ell)} \in \mathbb{R}^{d_{\ell-1} \times d_{\ell-1}}$, let $\widetilde{W}_O \in \mathbb{R}^{d_{out} \times d_L}$ be the output matrix, and let $\widetilde{\theta} = \{\widetilde{A}_i^{(\ell)}\}_{\ell \in [L], i \in [m_\ell]} \cup \{\widetilde{W}_O\}$. A disentangled transformer $\widetilde{\mathrm{TF}}_{\widetilde{\theta}}$ acts on a sequence $s_{1:T}$ by:

$$h^{(0)} = \tilde{X} = [\tilde{x}_1, \dots, \tilde{x}_T]^\top \text{ where } \tilde{x}_t = [e_{s_t}, e_t] \in \mathbb{R}^{d_0}$$

$$h^{(\ell)} = \Big[h^{(\ell-1)}, \mathrm{attn}(h^{(\ell-1)}; \widetilde{A}_1^{(\ell)}), \dots, \mathrm{attn}(h^{(\ell-1)}; \widetilde{A}_{m_\ell}^{(\ell)})\Big]$$

$$\widetilde{\mathrm{TF}}_{\widetilde{\theta}}(s_{1:T}) = h^{(L)} \widetilde{W}_O^\top.$$

We remark that $h^{(\ell)} \in \mathbb{R}^{T \times d_\ell}$ for $\ell = 0, \dots, L$.

In addition to disentangling the residual stream, Definition 2.3 replaces the query and key matrices with a single attention matrix $\widetilde{A} := QK^\top$ and absorbs the value matrices into $\widetilde{W}_O$. By allowing d_ℓ to grow with the depth, this disentangled transformer is actually *equivalent* to a decoder-based attention-only transformer (see Theorem A.1 for the formal statement). Given this equivalence, throughout the rest of the paper we study the disentangled transformer.

When the target is a vector in $\mathbb{R}^{d_{out}}$ rather than a sequence in $\mathbb{R}^{T \times d_{out}}$, it is customary to use the embedding of the last token, i.e. $\mathrm{TF}_\theta(s_{1:T}) = W_O h_T^{(L)}$ and similarly for $\widetilde{\mathrm{TF}}_{\widetilde{\theta}}$.

2.2. Problem Setup: Random Sequences with Causal Structure

Let $\mathcal{G} = ([T], \mathcal{E})$ be a directed acyclic graph on $[T] = \{1, \dots, T\}$ with edge set $\mathcal{E}$, which represents the latent causal structure. We assume that $(j \to i) \in \mathcal{E}$ only if $j < i$, i.e. each token can only point to future tokens. For a position $i \in [T]$, we let $p(i)$ denote the set of parents of i, i.e. $p(i) := \{j : (j \to i) \in \mathcal{E}\}$. We let $\mathcal{R}$ denote the set of root nodes, i.e $\mathcal{R} = \{i : p(i) = \emptyset\}$. For most of the paper, we assume that each position has at most one parent, i.e. $|p(i)| \le 1$ for all $i \in [T]$. See Section 6 for the generalization to multiple parents. When $|p(i)| = 1$, we overload notation and use $p(i) \in [T]$ to denote the unique parent of i.

We will also assume there exists a prior P_π over irreducible and aperiodic Markov chains π on $[S] = \{1, \dots, S\}$. For each π, we will use μ_π to denote the unique stationary measure of π. Then each sequence $[s_1, \dots, s_T]$ and its corresponding target y are generated by the following procedure:

Task 2.4 (Random Sequence with Causal Structure)**.**

1. First, draw $\pi \sim P_\pi$.
2. For $i = 1, \dots, T-1$, sample $s_i \sim \mu_\pi$ if $p(i) = \emptyset$. Otherwise sample $s_i \sim \pi(\cdot | s_{p(i)})$.
3. Draw $s_T \sim \mathrm{Unif}([S])$ and $s_{T+1} \sim \pi(\cdot | s_T)$
4. Return the input $x = s_{1:T}$ and the target $y = s_{T+1}$.

Because $s_T \sim \mathrm{Unif}([S])$, T is a root node of $\mathcal{G}$, i.e. $T \in \mathcal{R}$.

2.3. Examples

Markov Chains and Induction Heads. First, consider the case where $p(i) = i - 1$. The sequence $s_1, \dots, s_{T-1}$ is a Markov chain conditioned on π, with transition matrix π. Task 2.4 reduces to the problem of *estimating the Markov chain* π *in-context*. This can be solved via an induction head (Olsson et al., 2022) which, when presented with a prompt $\mathcal{P} = [\cdots, A, B, \cdots, A, C, \cdots, A]$, averages over the tokens following the previous occurrences of A, (in this case B and C). Explicitly, the output of an induction head on the sequence $s_{1:T}$ will be the empirical estimate for $\pi(\cdot \mid s_T)$:

$$\mathrm{TF}_\theta(s_{1:T})_{s'} = \frac{|\{i \,:\, (s_{i-1}, s_i) = (s_T, s')\}|}{|\{i \,:\, s_{i-1} = s_T\}|}.$$

In the limit as $T \to \infty$, this converges to $\pi(\cdot \mid s_T)$.

In-Context Learning. Consider the in-context learning setup from Garg et al. (2022). This corresponds to the causal graph where $p(2k-1) = \emptyset$ and $p(2k) = 2k-1$. Sequences are generated by sampling $f : [S] \to [S]$ from $\mathcal{F}$ and using the transition matrix $\pi(s'|s) = \mathbf{1}(s' = f(s))$. To learn this function class in-context, the transformer must learn to associate the (x, y) pairs in positions $2k-1$ and $2k$.

3. What does the Transformer Learn?

3.1. Experiments

We train a series of two-layer disentangled transformers with one head per layer on Task 2.4, for varying latent graphs $\mathcal{G}$. The prior P_π is chosen so that each row of π is sampled i.i.d from the Dirichlet distribution with parameter α, i.e $\pi(\cdot \mid s) \sim \mathrm{Dir}(\alpha \cdot 1_S)$, for varying values of α. We train using gradient descent on the cross entropy loss with initial learning rate 1 and cosine decay over 2^{17} steps.

We observe that the weights of the trained disentangled transformers exhibit consistent structure. First, all of the entries of $\widetilde{A}^{(1)}$ remain small except the position-position block (red box under $\widetilde{A}^{(1)}$ in Figure 2(a)), which converges to the

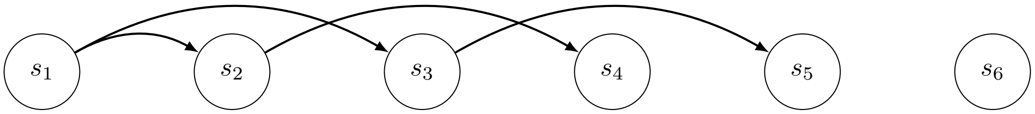


Figure 1. **Random Sequence with Causal Structure:** The causal structure is defined by the graph $\mathcal{G}$, denoted by the arrows. In this figure, $p(1)=\emptyset$, $p(2)=\{1\}$, $p(3)=\{1\}$, $p(4)=\{2\}$ and $p(5)=\{3\}$. Sequences are generated by sampling $\pi\sim P_\pi$, $s_1\sim\mu_\pi$, $s_2\sim\pi(\cdot|s_1)$, $s_3\sim\pi(\cdot|s_1)$, $s_4\sim\pi(\cdot|s_2)$, $s_5\sim\pi(\cdot|s_3)$, and finally $s_6\sim\mathrm{Unif}([S])$. The target y for this sequence is drawn from $\pi(\cdot|s_6)$.

adjacency matrix of the graph $\mathcal{G}$. Next, all of the entries of $\widetilde{A}^{(2)}$ are small, except the token/token block comparing the $h^{(0)}$ component of the residual stream of token i to the $\mathrm{attn}(h^{(0)},\widetilde{A}^{(1)})$ component of the residual stream of token j for $j\le i$ (red box under $\widetilde{A}^{(2)}$ in Figure 2(a)). Finally, all of the entries of the output matrix W_O are small except the token/token block which returns the value of the first component of the output of the second attention $\mathrm{attn}(h^{(1)},A^{(2)})$ (red box under W_O in Figure 2(a)). In Figure 4, we observe that this weight pattern persists for different latent graphs $\mathcal{G}$.

In the following section, we explicitly define this construction and describe the corresponding dynamics of the forward pass in Figure 2(b).

3.2. Construction

In Figure 2(a) we observe that the attention weights $\widetilde{A}^{(1)},\widetilde{A}^{(2)}$ and output weight $\widetilde{W}_O$ are of the form

$$\begin{aligned}
\widetilde{A}^{(1)} &= \begin{bmatrix} 0_{S\times S} & 0_{S\times T} \\ 0_{T\times S} & A^{(1)} \end{bmatrix} \\
\widetilde{A}^{(2)} &= \left[\begin{array}{cc|cc} 0_{S\times S} & 0_{S\times T} & A^{(2)} & 0_{S\times T} \\ 0_{T\times S} & 0_{T\times T} & 0_{T\times S} & 0_{T\times T} \\ \hline 0_{S\times S} & 0_{S\times T} & 0_{S\times S} & 0_{S\times T} \\ 0_{T\times S} & 0_{T\times T} & 0_{T\times S} & 0_{T\times T} \end{array}\right] \\
\widetilde{W}_O &= \left[\begin{array}{c|c|cc|c} 0_{S\times d} & 0_{S\times d} & I_S & 0_{S\times T} & 0_{S\times d} \end{array}\right]
\end{aligned} \tag{3}$$

for matrices $A^{(1)}\in\mathbb{R}^{T\times T}$ and $A^{(2)}\in\mathbb{R}^{S\times S}$. We now explicitly construct the $A^{(1)}$ and $A^{(2)}$ from Figure 2(a) that solve Task 2.4. Indeed, we show that this construction solves the task by estimating the *empirical transition* matrix $\hat{\pi}_{s_{1:T}}$:

$$\hat{\pi}_{s_{1:T}}(s'\mid s) := \frac{|\{(j\to i)\in\mathcal{E}\;:\;(s_j,s_i)=(s,s')\}|}{|\{(j\to i)\in\mathcal{E}\;:\;s_j=s\}|}. \tag{4}$$

Construction 3.1. There exists a two-layer disentangled transformer $f_{\widetilde{\theta}}=\widetilde{\mathrm{TF}}_{\widetilde{\theta}}$ such that

$$f_{\widetilde{\theta}}(s_{1:T})_{s'}\approx\hat{\pi}_{s_{1:T}}(s'\mid s_T). \tag{5}$$

Proof. Set $A^{(1)}$ to be the (scaled) adjacency matrix of $\mathcal{G}$, i.e $A^{(1)}_{i,j}=\beta_1\mathbf{1}(j=p(i))$, and $A^{(2)}=\beta_2 I_S$, where $\beta_1,\beta_2\to\infty$. We will now show that the output of the disentangled transformer approximates $\hat{\pi}_{s_{1:T}}(\cdot\mid s_T)$.

First Attention. Note that by the construction of $\widetilde{A}^{(1)}$, $\tilde{X}\widetilde{A}^{(1)}\tilde{X}^\top=A^{(1)}$, which is the scaled adjacency matrix of $\mathcal{G}$. If i is not a root node (i.e. $i\in\overline{\mathcal{R}}$, $p(i)\neq\emptyset$), then

$$\mathcal{S}(\tilde{X}\widetilde{A}^{(1)}\tilde{X}^\top)_{i,j}=\mathbf{1}(j=p(i)) \tag{6}$$

so i attends to its parent $p(i)$. Therefore, the output of the first attention is the token at position $p(i)$, i.e. $\mathrm{attn}(\tilde{X};\widetilde{A}^{(1)})_i=\tilde{x}_{p(i)}$. The transformer then appends $\tilde{x}_{p(i)}$ to the residual stream of token i.

When i is a root node (i.e. $i\in\mathcal{R}$, $p(i)=\emptyset$), then for all j, $(\tilde{X}\widetilde{A}^{(1)}\tilde{X}^\top)_{ij}=0$. Therefore after the softmax, i will attend equally to all previous tokens:

$$\mathcal{S}(\tilde{X}\widetilde{A}^{(1)}\tilde{X}^\top)_{i,j}=\tfrac{1}{i}\quad\text{for all}\quad j\le i. \tag{7}$$

Thus the first attention layer averages all of the tokens in the sequence: $\mathrm{attn}(\tilde{X};\widetilde{A}^{(1)})_i=\frac{1}{i}\sum_{j\le i}\tilde{x}_j$. It then copies this average into the residual stream.

Second Attention. We next show that the Tth token attends to all prior tokens whose parents are equal to s_T. It then averages them and copies them into the residual stream.

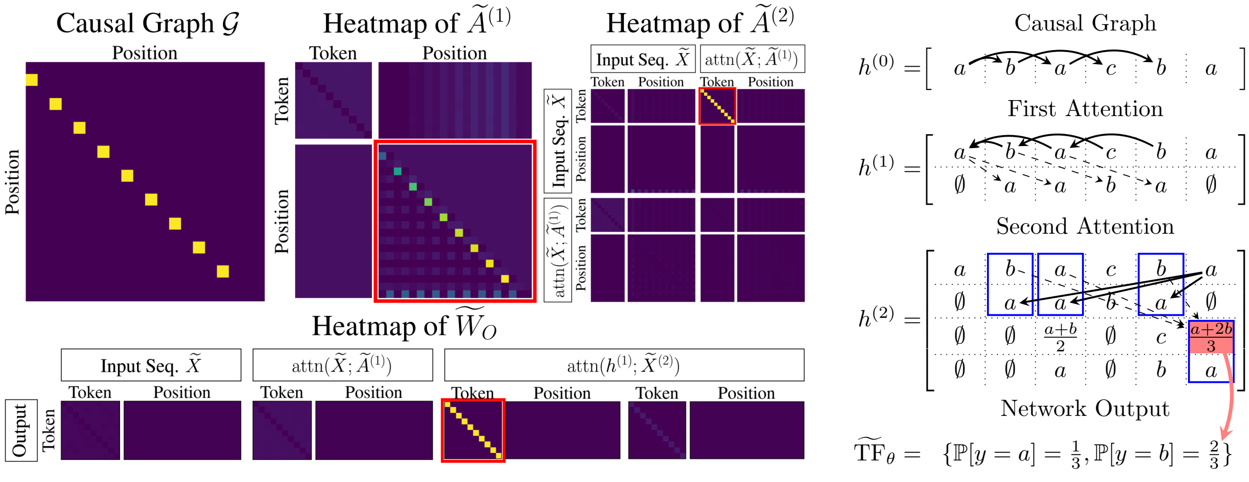
After the first attention layer, the residual stream is $h^{(1)}_j=[\tilde{x}_j,\mathrm{attn}(\tilde{X};\widetilde{A}^{(1)})_j]^\top$. The second attention layer compares the Tth token of the original sequence $\tilde{x}_T$ to the output of the first attention at all other positions. Explicitly, the attention pattern is equal to:

$$\begin{aligned}
h^{(1)\top}_T\widetilde{A}^{(2)}h^{(1)}_j &= \beta_2\cdot\tilde{x}_T^\top\begin{bmatrix} A^{(2)} & 0_{S\times T} \\ 0_{T\times S} & 0_{T\times T}\end{bmatrix}\mathrm{attn}(\tilde{X};\widetilde{A}^{(1)})_j \\
&= \beta_2\cdot\begin{cases}\mathbf{1}(s_{p(i)}=s_T) & i\in\overline{\mathcal{R}} \\ \frac{1}{i}\sum_{j\le i}\mathbf{1}(s_j=s_T) & i\in\mathcal{R}.\end{cases}
\end{aligned}$$

As $\beta_2\to\infty$, the softmax converges to a hard max, and so the Tth token attends equally to all tokens i such that $s_{p(i)}=s_T$. The attention then averages all of these tokens, so the Tth token in the residual stream is equal to $h^{(2)}_T=\left[\tilde{x}_T,\frac{1}{T}\sum_{j\le T}\tilde{x}_j,Z,\tilde{x}_T\right]$ where

$$Z:=\frac{\sum_{s_{p(i)}=s_T}\tilde{x}_i}{|\{i\,:\,s_{p(i)}=s_T\}|} \tag{8}$$

is the average of the tokens whose parent is equal to s_T.


Causal Graph $\mathcal{G}$
Position
Position
Heatmap of $\widetilde{A}^{(1)}$
Token
Position
Token
Position
Heatmap of $\widetilde{A}^{(2)}$
Input Seq. $\widetilde{X}$
attn($\widetilde{X}$; $\widetilde{A}^{(1)}$)
Token
Position
Token
Position
Input Seq. $\widetilde{X}$
Token
Position
attn($\widetilde{X}$; $\widetilde{A}^{(1)}$)
Token
Position
Heatmap of $\widetilde{W_O}$
Input Seq. $\widetilde{X}$
attn($\widetilde{X}$; $\widetilde{A}^{(1)}$)
attn($h^{(1)}$; $\widetilde{X}^{(2)}$)
Token
Position
Token
Position
Token
Position
Token
Position
Output
Token
Causal Graph
$h^{(0)} = [\ a\ b\ a\ c\ b\ a\ ]$
First Attention
$h^{(1)} =$
a b a c b a
∅ a a b a ∅
Second Attention
$h^{(2)} =$
a b a c b a
∅ a a b a ∅
∅ ∅ $\frac{a+b}{2}$ ∅ c $\frac{a+2b}{3}$
∅ ∅ a ∅ b a
Network Output
$\widetilde{\mathrm{TF}}_\theta = \ \{\mathbb{P}[y=a]=\frac{1}{3}, \mathbb{P}[y=b]=\frac{2}{3}\}$

(a) The Weights of a Trained Transformer

(b) Understanding the Forward Pass

Figure 2. We visualize the weights and depict the forward pass of trained transformers on tasks with two different causal graphs. **(a) The Weights of a Trained Transformer:** We plot the weights of a two layer disentangled transformer trained on Task 2.4 with $S = 10$ and $T = 20$ when the causal graph is the in-context learning graph where $p(2i) = 2i - 1$ for all $i > 0$. All entries of $A^{(1)}, A^{(2)}, W_O$ remain small except the three blocks highlighted in red. The highlighted block in $A^{(1)}$ converges to the adjacency matrix of the causal graph $\mathcal{G}$, and the highlighted blocks in $A^{(2)}, W_O$ converge to the identity matrix I_S. **(b) Understanding the Forward Pass:** The solid arrows represent the causal graph $\mathcal{G}$ defined in Figure 1 and $h^{(0)}$ denotes the unmodified input sequence. The first attention *reverses* this causal pattern, as every token attends to its parent (solid arrows). It then *appends* this parent token to the residual stream (dashed arrows). In the second attention layer, each token i attends to all previous tokens j whose parent token $p(j)$ has the same value, i.e. $s_i = s_{p(j)}$ (solid arrows), and appends the *average of these tokens* into the residual stream (dashed arrows). Finally, the transformer returns the third entry in the last column (red box), which is the average of all of the tokens whose parent token has the same value as the last token.

Output Layer. W_O reads from the third block in this stream, which we denoted by Z in (8) above. It then returns the token embedding of Z which is equal to:

$$f_{\widetilde{\theta}}(s_{1:T}) = \frac{\sum_{s_{p(i)}=s_T} e_{s_i}}{|\{i : s_{p(i)}=s_T\}|} = \hat{\pi}_{s_{1:T}}(\cdot|s_T). \qquad (9)$$

□

See Figure 2(b) for a breakdown of this forward pass through the transformer for a specific sequence.

3.3. The Reduced Model

Motivated by the sparsity pattern in Figure 2(a) and Equation (3), we consider training a simplified two-layer transformer architecture where the sparsity in Equation (3) is fixed, and only $A^{(1)}$ and $A^{(2)}$ are trained. The transformer $\widetilde{\mathrm{TF}}_{\widetilde{\theta}}$ can be rewritten as the following reduced model:

Lemma 3.2. *Let $\theta = (A^{(1)}, A^{(2)})$, and let $\widetilde{\theta} = (\widetilde{A}^{(1)}, \widetilde{A}^{(2)}, \widetilde{W}_O)$ be defined in Equation (3). Let $f_\theta = \widetilde{\mathrm{TF}}_{\widetilde{\theta}}$ be a two-layer disentangled transformer parameterized by θ. Then if $\overline{X} = [\overline{x}_1, \ldots, \overline{x}_T]^T$ where $\overline{x_i} = e_{s_i}$,*

$$f_\theta(s_{1:T}) = \overline{X}^\top \mathcal{S}\Big(\mathcal{S}(\mathrm{MASK}(A^{(1)}))\overline{X}A^{(2)^\top}\overline{x}_T\Big). \qquad (10)$$

Due to the masking, we restrict $A^{(1)}$ to be lower diagonal.

Our goal is to analyze the gradient descent dynamics of f_θ under the cross entropy loss. However, if the token s' does not appear in $s_{1:T}$, then $f_\theta(s_{1:T})_{s'}$ is 0 and the cross entropy loss is infinite. As such, we perturb the predictions by some small $\epsilon > 0$. The perturbed population loss is thus:

$$L(\theta) = -\mathop{\mathbb{E}}_{\pi, s_{1:T}}\left[\sum_{s' \in [S]} \pi(s'|s_T)\log\left(f_\theta(s_{1:T})_{s'} + \epsilon\right)\right] \qquad (11)$$

In the following sections, we will study the gradient descent dynamics of the reduced model (10) on the loss (11).

4. Main Results

4.1. Training Algorithm

Our training algorithm is stage-wise gradient descent on the population loss (11) using the reduced model (10). The model is initialized at $A^{(1)} = 0_{T\times T}, A^{(2)} = \beta_0 I_{S\times S}$ for small initialization scale β_0. The first stage is gradient descent on $A^{(1)}$ with learning rate η_1 for τ_1 timesteps. The second stage is gradient descent on $A^{(2)}$ with learning rate η_2 for τ_2 timesteps. Pseudocode is given in Algorithm 1.

```
Algorithm 1 Training Algorithm
Input: init size $\beta_0$; learning rates $\eta_1, \eta_2$; times $\tau_1, \tau_2$
  Initialize $A^{(1)}(0) = 0_{T \times T}, A^{(2)}(0) = \beta_0 \cdot I_{S \times S}$
  for $t = 1, \dots, \tau_1$ do
    $A^{(1)}(t) \leftarrow A^{(1)}(t-1) - \eta_1 \nabla_{A^{(1)}} L(\theta^{(t-1)})$ {Stage 1}
    $\theta^{(t)} = (A^{(1)}(t), A^{(2)}(0))$
  end for
  for $t = \tau_1, \dots, \tau_1 + \tau_2$ do
    $A^{(2)}(t) \leftarrow A^{(2)}(t-1) - \eta_2 \nabla_{A^{(2)}} L(\theta^{(t-1)})$ {Stage 2}
    $\theta^{(t)} = (A^{(1)}(\tau_1), A^{(2)}(t))$
  end for
  $\hat{\theta} \leftarrow \theta^{(\tau_1 + \tau_2)}$
Output: $\hat{\theta}$.
```

We require the following assumptions on the prior P_π:

Assumption 4.1 (Assumptions on prior P_π.)**.** There exists $\gamma > 0$ such that almost surely over the draw of π:

- (Transition lower bounded): $\min_{s,s'} \pi(s' \mid s) > \gamma/S$.
- (Non-degeneracy of chain): The chain does not immediately mix to the stationary measure μ_π in one step:
$$\textstyle\sum_s \|\pi(\cdot \mid s) - \mu_\pi(\cdot)\|_2^2 \geq \gamma^2/S$$
- (Symmetry): For any permutation σ on $[S]$, $\sigma^{-1}\pi\sigma \overset{d}{=} \pi$.
- (Constant mean): $\mathbb{E}_\pi[\pi] = \frac{1}{S} 1_S 1_S^\top$.

The final two assumptions imply that the marginal distributions of $\pi(s' \mid s)$ are equal for any $s' \neq s$, and likewise for $\pi(s \mid s)$, and that these distributions have mean $1/S$. We remark that Assumption 4.1 is satisfied with probability 0.99 for some $\gamma = \Theta(1)$ when each row of π is sampled i.i.d from a Dirichlet distribution with parameter $\alpha = \Theta(1)$.

Additionally, we assume that a non-vanishing fraction of nodes have a parent.

Assumption 4.2. Let $r := |\mathcal{R}|/T$. Then $r \leq 1 - \gamma$.

Throughout the proof, we let $C_{\gamma,S}$ denote an absolute constant that depends *polynomially* on γ^{-1} and S. If $A \leq C_{\gamma,S} B$, then we write $A = O_{\gamma,S}(B)$ or $A \lesssim_{\gamma,S} B$. For convenience, we also drop the dependence on γ, S, and write $O(\cdot)$ or $\lesssim$.

4.2. Main Theorem

The minimum possible value for the (unperturbed) loss is the mean entropy of π, averaged over the prior P_π:

$$L^* := -\mathbb{E}_\pi\Big[\tfrac{1}{S} \textstyle\sum_{s,s'} \pi(s' \mid s) \log \pi(s' \mid s)\Big]. \tag{12}$$

We also define the effective sequence length as follows:

Definition 4.3 (Effective Sequence Length)**.** Decompose $\mathcal{G} = \bigcup_{i=1}^k \mathcal{T}_i$ where $\mathcal{T}_i$ are disjoint trees. Let L_i denote the number of leaves of tree $\mathcal{T}_i$. Then, $T_{\text{eff}} := \frac{T}{\max_{i=1}^k L_i}$.

The effective sequence length roughly captures the number of independent samples present in the sequence $s_{1:T}$, and is related to the mixing time of the process on $\mathcal{G}$. For both the Markov chain and in-context learning examples, we see that $L_i = 1$ and thus $T_{\text{eff}} = T$.

Our main theorem is the following:

Theorem 4.4 (Guarantee for Algorithm 1)**.** *Assume that the effective sequence length satisfies $T_{\mathit{eff}} \geq poly(\gamma^{-1}, S)$. There exist $\epsilon, \eta_1, \eta_2, \tau_1, \tau_2$ such that the output of Algorithm 1, $\hat{\theta} = (\hat{A}^{(1)}, \hat{A}^{(2)})$, satisfies*

$$L(\hat{\theta}) - L^* \lesssim \frac{\log T}{T_{\mathit{eff}}^{c\gamma}} \quad \text{and} \quad \mathcal{S}(\hat{A}^{(1)})_{i,p(i)} \geq 1 - O\left(\frac{1}{T}\right).$$

for $i \in \overline{\mathcal{R}}$, for some constant $c > 0$ (independent of γ, S).

Algorithm 1 thus approximately minimizes the loss by encoding the adjacency matrix of $\mathcal{G}$ in the first attention layer $\hat{A}^{(1)}$. Furthermore, we show that the trained model $\hat{\theta}$ achieves good prediction on transitions π which are out of distribution:

Theorem 4.5 (OOD Generalization)**.** *Let $\tilde{\pi}$ have transition lower bounded as $\min_{s,s'} \tilde{\pi}(s' \mid s) \geq \gamma/S$ and let $\hat{\theta}$ be the trained model from Theorem 4.4. Let $s_{1:T}$ be generated by steps 2-4 of Task 2.4. Then with probability at least 0.99 over the draw of $s_{1:T}$,*

$$\sup_{s'} \left| f_{\hat{\theta}}(s_{1:T})_{s'} - \tilde{\pi}(s' \mid s_T) \right| \lesssim \frac{\log T}{T_{\mathit{eff}}^{c\gamma}}. \tag{13}$$

We remark that the only assumption needed on $\tilde{\pi}$ is the lower bound on the transition; it does not need to be close to typical draw from the prior P_π.

5. Proof Sketch

5.1. Stage 1: Learning the Causal Graph

The first step of the proof is to show that during the first stage of training, the first attention layer $A^{(1)}$ learns the latent causal graph $\mathcal{G}$.

5.1.1. The Oracle Algorithm

We begin by describing an efficient algorithm for learning the graph $\mathcal{G}$. The goal is to recover the parent node $p(i)$ for each i. The key idea is that as a result of the data generating process, $s_{p(i)}$ is the node which maximizes mutual information with s_i.

We briefly recall the definition of an f-divergence.

Definition 5.1. Let f be a convex function with $f(1) = 0$. The f-divergence between two probability distributions P, Q on state space $\mathcal{X}$ is defined as

$$D_f(P||Q) := \sum_{x \in \mathcal{X}} Q(x) f\left(\frac{P(x)}{Q(x)}\right). \tag{14}$$

The f mutual information between two random variables Y, Z, denoted by $I_f(Y; Z)$, is

$$I_f(Y; Z) := D_f(P_{Y,Z}||P_Y \otimes P_Z). \tag{15}$$

Given a latent variable C, the conditional mutual information $I_f(Y; Z \mid C)$ is defined as

$$I_f(Y; Z \mid C) := \mathbb{E}_C\left[D_f(P_{(Y,Z)|C}||P_{Y|C} \otimes P_{Z|C})\right].$$

Information measures admit a *data processing inequality*:

Lemma 5.2 (Data Processing Inequality)**.** *Let I_f be an information measure, and let $W \to Y \to Z$ be a Markov chain. Then $I_f(W; Z) \le I_f(Y; Z)$.*

The data processing inequality suggests an efficient algorithm for recovering $\mathcal{G}$. For non-root nodes $i \in \overline{\mathcal{R}}$, $s_j \to s_{p(i)} \to s_i$ form a Markov chain conditioned on π. Therefore by the data processing inequality, $p(i) \in \arg\max_{j<i} I_f(s_i; s_j \mid \pi)$. Otherwise, if $i \in \mathcal{R}$, s_j and s_i are independent given π, and thus $I_f(s_i; s_j \mid \pi) = 0$. To recover the graph $\mathcal{G}$, one can compute the conditional mutual informations $I_f(s_i; s_j \mid \pi)$. If $I_f(s_i; s_j \mid \pi) = 0$ for all $j < i$, then i is a root node. Otherwise, $p(i) = \arg\max_{j<i} I_f(s_i; s_j \mid \pi)$. Pseudocode for this oracle algorithm is given in Algorithm 2.

Algorithm 2 Oracle Algorithm

```
$\mathcal{E} \leftarrow \emptyset$
for $i = 1, \dots, T-1$: do
  if $\max_{j<i} I_f(s_i; s_j \mid \pi) > 0$ then
    $p(i) \leftarrow \arg\max_{j<i} I_f(s_i, s_j \mid \pi)$.
    $\mathcal{E} \leftarrow \mathcal{E} \cup \{(p(i) \to i)\}$.
  end if
end for
```

We remark that Algorithm 2 is a special case of the celebrated Chow-Liu algorithm (Chow & Liu, 1968), when a topological ordering of the tree is known a priori: the tree consisting of edges $(p(i) \to i)$ where $p(i) = \arg\max_{j<i} I_f(s_i; s_j \mid \pi)$ is indeed the max-weight spanning forest when the edge weights are the conditional mutual informations.

5.1.2. The Gradient Descent Dynamics

We next compute the gradient with respect to $A^{(1)}$. Let $A^{(1)}_i \in \mathbb{R}^i$ denote the ith row of $A^{(1)}$ (restricted to the first i entries, since $A^{(1)}$ is lower triangular). Define $J : \mathbb{R}^k \to \mathbb{R}^{k \times k}$ by $J(v) = \mathrm{diag}(v) - vv^\top$; J is the Jacobian of the softmax function $\mathcal{S}$, in that $\nabla_u \mathcal{S}(u) = J(\mathcal{S}(u))$.

The following lemma computes the gradient with respect to $A^{(1)}$; a heuristic derivation is deferred to Appendix D.3.

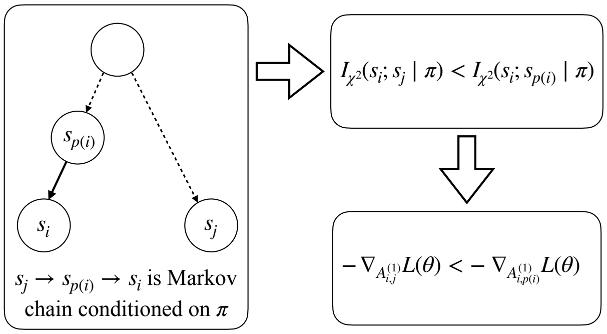

$s_{p(i)}$
s_i
s_j
$s_j \to s_{p(i)} \to s_i$ is Markov chain conditioned on π
$I_{\chi^2}(s_i; s_j \mid \pi) < I_{\chi^2}(s_i; s_{p(i)} \mid \pi)$
$-\nabla_{A^{(1)}_{i,j}} L(\theta) < -\nabla_{A^{(1)}_{i,p(i)}} L(\theta)$

Figure 3. By the data processing inequality, $A^{(1)}_{i,p(i)}$ grows faster than $A^{(1)}_{i,j}$.

Lemma 5.3.

$$\nabla_{A^{(1)}_i} L(\theta) = -\frac{\beta_0}{ST} J\Big(\mathcal{S}(A^{(1)}_i)\Big)\Big(g_i + O(T_{\mathit{eff}}^{-1/2})\Big), \tag{16}$$

where the jth entry of g_i, $g_{i,j}$, is

$$g_{i,j} := \mathbb{E}_\pi\left[\sum_{s,s'} \frac{\pi(s' \mid s)}{\mu_\pi(s')} \mathbb{P}_X[s_j = s, x_i = s']\right] - 1.$$

For non-root nodes $i \in \overline{\mathcal{R}}$, $(s_i, s_{p(i)})$ has joint distribution $\mathbb{P}[s_i = s', s_{p(i)} = s] = \mu_\pi(s)\pi(s'|s)$, and thus $g_{i,p(i)}$ is

$$g_{i,p(i)} = \mathbb{E}_\pi\left[\sum_{s,s'} \frac{\pi(s' \mid s)^2 \mu_\pi(s)^2}{\mu_\pi(s')\mu_\pi(s)} - 1\right]. \tag{17}$$

It turns out that this expression is exactly equal to the χ^2-mutual information, I_{χ^2}, between s_i and $s_{p(i)}$ conditioned on π. The χ^2-divergence is the f-divergence obtained by setting $f(z) = (z-1)^2$. Therefore

$$g_{i,p(i)} = I_{\chi^2}(s_i; s_{p(i)} \mid \pi). \tag{18}$$

By Cauchy-Schwarz, we can also upper bound $g_{i,j}$ by the sum of two χ^2-mutual informations:

$$g_{i,j} \le \frac{1}{2} I_{\chi^2}(s_i; s_{p(i)} \mid \pi) + \frac{1}{2} I_{\chi^2}(s_i; s_j \mid \pi). \tag{19}$$

Applying the data processing inequality[1], we obtain that for $j \ne p(i)$

$$g_{i,j} < I_{\chi^2}(s_i; s_{p(i)} \mid \pi) = g_{i,p(i)}. \tag{20}$$

Therefore $g_{i,j}$ is maximized at $j = p(i)$, and the gradient is aligned with the adjacency matrix of the causal graph $\mathcal{G}$. In fact, the gradient descent dynamics mimic Algorithm 2!

[1] By the assumptions on the prior P_π (Assumption 4.1), the data processing inequality is indeed strict.

Maintaining the inductive hypothesis that $\arg\max_j A^{(1)}_{i,j} = p(i)$, we see by the gradient formula in Lemma 5.3 that $\arg\max\left[-\nabla_{A^{(1)}_i} L(\theta)_j\right] = p(i)$. Thus $A^{(1)}_{i,p(i)}$ continues to grow faster than the other entries throughout stage 1. This growth continues until $\mathcal{S}(A^{(1)})_{i,p(i)} \approx 1$.

For root nodes $i \in \mathcal{R}$, i is independent of j. Since both s_i and s_j have the marginal μ_π, one has

$$g_{i,j} = \mathbb{E}_\pi\left[\sum_{s,s'} \frac{\pi(s' \mid s)}{\mu_\pi(s')}\mu_\pi(s')\mu_\pi(s)\right] - 1 = 0 \qquad (21)$$

and thus $\nabla_{A^{(1)}_i} L(\theta) \approx 0$. Therefore at the end of stage 1, $\mathcal{S}(A^{(1)})_{i,j} \approx \frac{1}{i}$ for all $j < i$.

Altogether, at the end of stage 1, $A^{(1)}$ satisfies

$$\mathcal{S}(A^{(1)})_{i,j} \approx \begin{cases} \mathbf{1}(j = p(i)) & i \in \overline{\mathcal{R}} \\ \frac{1}{i} & i \in \mathcal{R} \end{cases}. \qquad (22)$$

The precise quantitative bound is given in Corollary D.6, and requires controlling the various error terms throughout multiple steps of gradient descent.

5.2. Stage 2: Decreasing the Loss

We next show that during the second stage, $A^{(2)}$ grows large in the direction $I_S - \frac{1}{S}1_S 1_S^\top$. By a symmetry argument, one can show that $\nabla_{A^{(2)}} L(\theta)$ is proportional to $I_S - \frac{1}{S}1_S 1_S^\top$. Writing $A^{(2)} = \beta I_S + \frac{\beta - \beta_0}{S} 1_S 1_S^\top$, it suffices to show that $\nabla_\beta L(\theta) < 0$.

In Lemma D.8, we show that $-\nabla_\beta L(\theta)$ can be approximated by a quantity which is an f-mutual information for some convex f defined in terms of β. We show that this quantity is strictly positive (Lemma D.9) until $\beta = \Theta(\log T_{\text{eff}})$. Thus at the end of stage 2, $\beta = \Theta(\log T_{\text{eff}})$.

To conclude the proof of Theorem 4.4, we must show that $f_{\hat\theta}(X; s)_{s'} \approx \pi(s' \mid s)$. Indeed, Lemma H.8 shows that

$$\left|f_{\hat\theta}(X; s)_{s'} - \pi(s' \mid s)\right| \le \exp(-\Theta(\beta)) = T_{\text{eff}}^{-\Theta(1)}, \quad (23)$$

which implies the desired bound on the loss.

6. Causal Graphs with Multiple Parents

We next consider a generalization of Task 2.4. Let $\mathcal{G}$ be a directed acyclic graph over the vertex set $[T+1]$. For each node $i \in [T+1]$, we assume that the set of parent nodes $p(i) \subset [i-1]$ satisfy the property that either $p(i) = \emptyset$ or $|p(i)| = k$. If $p(i) \neq \emptyset$, we write $p(i) = \{p(i)_1, \dots, p(i)_k\}$, where $p(i)_1 < \dots < p(i)_k$. As before, let $\mathcal{R} = \{i \in [T+1] : p(i) = \emptyset\}$ be the root nodes. We additionally assume that $T + 1 \notin \mathcal{R}$.

We now consider k-parent transition tensors π: For any $a_1, \dots, a_k \in [S]$, $\pi(\cdot \mid a_1, \dots, a_k)$ is a probability distribution over $[S]$. Let P_π^k be a prior over such π. Each sequence is now generated as follows:

Task 6.1 (Graphs with Multiple Parents)**.**

1. Draw $\pi \sim P_\pi^k$.
2. For $i = 1, \dots, T+1$, if $p(i) = \emptyset$, sample $s_i \sim \mathrm{Unif}([S])$. Otherwise, sample $s_i \sim \pi(\cdot \mid s_{p(i)_1}, \dots, s_{p(i)_k})$
3. Return the input $x = s_{1:T}$ and the target $y = s_{T+1}$.

Example. One example of Task 6.1 is learning in-context n-grams. In an n-gram model, each token only depends on the prior $n-1$ tokens in the sequence. This n-gram model can be obtained by setting $k = n-1$, letting the root nodes be $\mathcal{R} = [n-1]$, and choosing the parent sets $p(i) = \{i-n+1, i-n+2, \dots, i-1\}$ for $i \geq n$. The conditional density $\mathbb{P}(s_{k+n} \mid s_{k+1:k+n-1})$ is then just the transition $\pi(s_{k+n} \mid s_{k+1}, \dots, s_{k+n-1})$; the goal is to estimate this transition in-context, by first learning that all sequences share the same n-gram causal structure.

Given a sequence $s_{1:T}$, a good estimate for the transition π is the empirical transition $\hat\pi_{s_{1:T}}(s' \mid a_1, \dots, a_k)$, defined as

$$\frac{\left|\{i : s_i = s', s_{p(i)_1} = a_1, \dots, s_{p(i)_k} = a_k\}\right|}{\left|\{i : s_{p(i)_1} = a_1, \dots, s_{p(i)_k} = a_k\}\right|} \qquad (24)$$

We explicitly construct a two-layer transformer with k heads in the first layer that approximately expresses this empirical transition.

Construction 6.2. There exists a two attention layer transformer $f_{\tilde\theta}$ with k heads such that

$$f_{\tilde\theta}(s_{1:T})_{s'} \approx \hat\pi_{s_{1:T}}(s' \mid s_{p(T+1)_1}, \dots, s_{p(T+1)_k}) \qquad (25)$$

Construction 6.2 is deferred to Appendix B. At a high level, the ℓth head in the first layer copies $p(i)_\ell$ to the residual stream of i, and copies $p(T+1)_\ell$ to the residual stream of T; the second attention head compares these tuples of parents, and thus attends to tokens i where $s_{p(i)_\ell} = s_{p(T+1)_\ell}$ for all $\ell \in [k]$.

In Figures 5 and 6, we show empirically that transformers trained on Task 6.1 for varying latent graphs $\mathcal{G}$ indeed converge to such a construction. The challenge, however, in analyzing the gradient descent dynamics is that there are multiple attention heads each of which attends to a different parent. The dynamics must thus break the symmetry between the multiple heads. Analyzing the optimization dynamics of a multi-head transformer for solving Task 6.1 is thus a very interesting direction for future work.

Impact Statement

This paper focuses on understanding and explaining existing methods and techniques. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgements

EN acknowledges support from a National Defense Science & Engineering Graduate Fellowship. AD acknowledges support from a NSF Graduate Research Fellowship. EN, AD, and JDL acknowledge support of the ARO under MURI Award W911NF-11-1-0304, the Sloan Research Fellowship, NSF CCF 2002272, NSF IIS 2107304, NSF CIF 2212262, ONR Young Investigator Award, and NSF CAREER Award 2144994.

References

Ahn, K., Cheng, X., Daneshmand, H., and Sra, S. Transformers learn to implement preconditioned gradient descent for in-context learning. *arXiv preprint arXiv:2306.00297*, 2023.

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.

Akyürek, E., Wang, B., Kim, Y., and Andreas, J. In-context language learning: Arhitectures and algorithms. *arXiv preprint arXiv:2401.12973*, 2024.

Bai, Y., Chen, F., Wang, H., Xiong, C., and Mei, S. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *arXiv preprint arXiv:2306.04637*, 2023.

Bietti, A., Cabannes, V., Bouchacourt, D., Jegou, H., and Bottou, L. Birth of a transformer: A memory viewpoint. *arXiv preprint arXiv:2306.00802*, 2023.

Boix-Adsera, E., Littwin, E., Abbe, E., Bengio, S., and Susskind, J. Transformers learn through gradual rank increase. *arXiv preprint arXiv:2306.07042*, 2023.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and Zhang, Q. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., and Mordatch, I. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.

Chow, C. and Liu, C. Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, 14(3):462–467, 1968.

Cohen, J. E., Iwasa, Y., Rautu, G., Beth Ruskai, M., Seneta, E., and Zbaganu, G. Relative entropy under mappings by stochastic matrices. *Linear Algebra and its Applications*, 179:211–235, 1993. ISSN 0024-3795. doi: https://doi.org/10.1016/0024-3795(93)90331-H. URL https://www.sciencedirect.com/science/article/pii/002437959390331H.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Edelman, B. L., Edelman, E., Goel, S., Malach, E., and Tsilivis, N. The evolution of statistical induction heads: In-context learning markov chains. *arXiv preprint arXiv:2402.11004*, 2024.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Friedman, D., Wettig, A., and Chen, D. Learning transformer programs. *arXiv preprint arXiv:2306.01128*, 2023.

Fu, D., Chen, T.-Q., Jia, R., and Sharan, V. Transformers learn higher-order optimization methods for in-context learning: A study with linear models. *arXiv preprint arXiv:2310.17086*, 2023.

Garg, S., Tsipras, D., Liang, P. S., and Valiant, G. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022.

Giannou, A., Rajput, S., Sohn, J.-y., Lee, K., Lee, J. D., and Papailiopoulos, D. Looped transformers as programmable computers. *arXiv preprint arXiv:2301.13196*, 2023.

Huang, Y., Cheng, Y., and Liang, Y. In-context convergence of transformers. *arXiv preprint arXiv:2310.05249*, 2023.

Jelassi, S., Sander, M., and Li, Y. Vision transformers provably learn spatial structure. *Advances in Neural Information Processing Systems*, 35:37822–37836, 2022.

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

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization, 2017.

Li, Y., Li, Y., and Risteski, A. How do transformers learn topic structure: Towards a mechanistic understanding. In *International Conference on Machine Learning*, pp. 19689–19729. PMLR, 2023.

Lu, H., Mao, Y., and Nayak, A. On the dynamics of training attention models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=1OCTOShAmqB.

Mahankali, A., Hashimoto, T. B., and Ma, T. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. *arXiv preprint arXiv:2307.03576*, 2023.

Nguyen, T. and Grover, A. Transformer neural processes: Uncertainty-aware meta learning via sequence modeling. In *International Conference on Machine Learning*, pp. 16569–16594. PMLR, 2022.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Reddy, G. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. *arXiv preprint arXiv:2312.03002*, 2023.

Snell, C., Zhong, R., Klein, D., and Steinhardt, J. Approximating how single head attention learns. *arXiv preprint arXiv:2103.07601*, 2021.

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*, 2023a.

Tian, Y., Wang, Y., Zhang, Z., Chen, B., and Du, S. Joma: Demystifying multilayer transformers via joint dynamics of mlp and attention. *arXiv preprint arXiv:2310.00535*, 2023b.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., and Vladymyrov, M. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pp. 35151–35174. PMLR, 2023.

Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021.

Zhang, R., Frei, S., and Bartlett, P. L. Trained transformers learn linear models in-context. *arXiv preprint arXiv:2306.09927*, 2023.

A. Disentangled Transformer Equivalence

Theorem A.1. *For any transformer* TF_θ *with any hidden dimension* d*, there exists a disentangled transformer* $\widetilde{\mathrm{TF}}_{\tilde\theta}$ *with the same depth and number of heads such that* $\mathrm{TF}_\theta(s_{1:T}) = \widetilde{\mathrm{TF}}_{\tilde\theta}(s_{1:T})$ *for any input sequence* $s_{1:T} \in [S]^T$*. Likewise, for any disentangled transformer* $\widetilde{\mathrm{TF}}_{\tilde\theta}$*, there exists a transformer* TF_θ *with the same depth and number of heads and with hidden dimension* $d^{(L)}$ *such that* $\mathrm{TF}_\theta(s_{1:T}) = \widetilde{\mathrm{TF}}_{\tilde\theta}(s_{1:T})$ *for any* $s_{1:T} \in [S]^T$*.*

Proof. Let

$$\theta = \{(Q_i^{(\ell)}, K_i^{(\ell)}, V_i^{(\ell)})_{\ell\in[L], i\in[m_\ell]}\} \cup \{E, P, W_O\} \quad \text{and} \quad \tilde\theta = \{A_i^{(\ell)}\}_{\ell\in[L], i\in[m_\ell]} \cup \{\tilde W_O\}.$$

Note that the reverse direction is clear as any disentangled transformer is also a transformer with hidden dimension d_L:

$$E = \begin{bmatrix} I_S \\ 0_{(d_L-S)\times S} \end{bmatrix} \in \mathbb{R}^{d_L\times S} \qquad P = \begin{bmatrix} 0_{S\times T} \\ I_T \\ 0_{(d_L-d)\times T} \end{bmatrix} \in \mathbb{R}^{d_L\times T} \qquad W_O = \tilde W_O$$

$$Q_i^{(\ell)} = \begin{bmatrix} A_i^{(\ell)} & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{d_L,d_L} \qquad K_i^{(\ell)} = \begin{bmatrix} I_{d_\ell} & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{d_L,d_L} \qquad V_i^{(\ell)} = \begin{bmatrix} 0_{i\cdot d_\ell\times d_\ell} & \\ I_{d_\ell} & 0_{d_L\times(d_L-d_\ell)} \\ 0_{(d_L-(i+1)\cdot d_\ell)\times d_\ell} & \end{bmatrix}.$$

We will prove that every transformer θ can be represented by a disentangled transformer $\tilde\theta$. We will begin by defining a sequence of matrices $Z^{(\ell)} \in \mathbb{R}^{d\times d_\ell}$ for $\ell \in \{0, \dots, L\}$. Let $Z^{(0)} := [E, P] \in \mathbb{R}^{d\times d_0}$, and for $\ell > 1$ let

$$Z^{(\ell)} := \begin{bmatrix} Z^{(\ell-1)} & V_1^{(\ell)} Z^{(\ell-1)} & \cdots & V_{m_\ell}^{(\ell)} Z^{(\ell-1)} \end{bmatrix} \in \mathbb{R}^{d\times d_\ell}.$$

Then we define

$$A_i^{(\ell)} := (Z^{(\ell-1)})^\top Q_i^{(\ell)} (V_i^{(\ell)})^\top Z^{(\ell-1)} \in \mathbb{R}^{d_{\ell-1}\times d_{\ell-1}} \quad \text{and} \quad \tilde W_O = W_O Z^{(L)}.$$

We will prove by induction that for any sequence $s_{1:T}$, $h^{(\ell)} = \tilde h^{(\ell)} (Z^{(\ell)})^\top$ for $\ell = 0, \dots, L$ where $\{h^{(\ell)}\}$ is the residual stream of TF_θ and $\{\tilde h^{(\ell)}\}$ is the residual stream of $\widetilde{\mathrm{TF}}_{\tilde\theta}$. First, when $\ell = 0$ we have that

$$h_i^{(0)} = E e_{s_i} + P e_i = \begin{bmatrix} E & P \end{bmatrix} \begin{bmatrix} e_{s_i} \\ e_i \end{bmatrix} = Z^{(0)} \tilde h_i^{(0)}.$$

Next, assume the result for $\ell - 1 \geq 0$. Then

$$\begin{aligned}
h^{(\ell)} &= h^{(\ell-1)} + \sum_{i=1}^{m_\ell} \mathrm{attn}\left(h^{(\ell-1)}; Q_i^{(\ell)} K_i^{(\ell)\top}\right) V_i^{(\ell)\top} \\
&= \tilde h^{(\ell-1)} (Z^{(\ell-1)})^\top + \sum_{i=1}^{m_\ell} \mathrm{attn}\Big(\tilde h^{(\ell-1)} (Z^{(\ell-1)})^\top; Q_i^{(\ell)} (K_i^{(\ell)})^\top\Big) (V_i^{(\ell)})^\top \\
&= \tilde h^{(\ell-1)} (Z^{(\ell-1)})^\top + \sum_{i=1}^{m_\ell} \mathrm{attn}\Big(\tilde h^{(\ell-1)}; (Z^{(\ell-1)})^\top Q_i^{(\ell)} (K_i^{(\ell)})^\top Z^{(\ell-1)}\Big) (Z^{(\ell-1)})^\top (V_i^{(\ell)})^\top \\
&= \tilde h^{(\ell-1)} (Z^{(\ell-1)})^\top + \sum_{i=1}^{m_\ell} \mathrm{attn}\Big(\tilde h^{(\ell-1)}; A_i^{(\ell)}\Big) (V_i^{(\ell)} Z^{(\ell-1)})^\top \\
&= \tilde h^{(\ell)} (Z^{(\ell)})^\top
\end{aligned}$$

which completes the induction. Therefore,

$$\mathrm{TF}_\theta(s_{1:T}) = h^{(L)} W_O^\top = \tilde h^{(L)} (Z^{(\ell)})^\top W_O^\top = h^{(L)} (W_O Z^{(\ell)})^\top = h^{(L)} \tilde W_O^\top = \widetilde{\mathrm{TF}}_\theta(s_{1:T})$$

which completes the proof. □

Example: Single Head Transformer. As an example, let us walk through the construction for a single-head transformer. A single layer attention-only transformer with only one head can be written as

$$\begin{aligned} h^{(0)} &= \mathrm{embed}(s_{1:T}; (E, P)) \\ \mathrm{TF}_\theta(s_{1:T}) &= \Big(h^{(0)} + \mathrm{attn}\Big(h^{(0)}; QK^\top\Big)V^\top\Big)W_O^\top. \end{aligned}$$

Recall that the input to the disentangled transformer is

$$\tilde{X} = \big[\tilde{x}_1, \ldots, \tilde{x}_T\big]^\top,$$

where $\tilde{x}_t = \big[e_{s_t}, e_t\big] \in \mathbb{R}^{d_0} = \mathbb{R}^{S+T}$. The input to the regular transformer, $h^{(0)}$, can then be written as

$$h^{(0)} = \tilde{X} \begin{bmatrix} E^\top \\ P^\top \end{bmatrix}.$$

Define $Z = \begin{bmatrix} E & P \end{bmatrix}$, so that $h^{(0)} = \tilde{X} Z^\top$. Set the weights $\tilde{A}, \tilde{W}_O$ of the disentangled transformer as

$$\tilde{A} = Z^\top Q K^\top Z \quad \text{and} \quad \tilde{W}_O = W_O \begin{bmatrix} Z & VZ \end{bmatrix}.$$

The output of the disentangled transformer is then

$$\begin{aligned} \widetilde{\mathrm{TF}}_{\tilde{\theta}}(s_{1:T}) &= \big[\tilde{X}, \quad \mathrm{attn}(\tilde{X}; \tilde{A})\big]\, \tilde{W}_O^\top \\ &= \big[\tilde{X}, \quad \mathrm{attn}(\tilde{X}; Z^\top Q K^\top Z)\big]\, \tilde{W}_O^\top \\ &= \Big[\tilde{X}, \quad \mathcal{S}\Big(\mathrm{MASK}\Big(\tilde{X} Z^\top Q K^\top Z \tilde{X}^\top\Big)\Big)\tilde{X}\Big] \begin{bmatrix} Z^\top \\ Z^\top V^\top \end{bmatrix} W_O^\top \\ &= \Big[\tilde{X} Z^\top, \quad \mathcal{S}\Big(\mathrm{MASK}\Big(\tilde{X} Z^\top Q K^\top Z \tilde{X}^\top\Big)\Big)\tilde{X} Z^\top\Big] \begin{bmatrix} I_d \\ V^\top \end{bmatrix} W_O^\top \\ &= \big[h^{(0)}, \quad \mathrm{attn}(h^{(0)}; QK^\top)\big] \begin{bmatrix} I_d \\ V^\top \end{bmatrix} W_O^\top \\ &= \Big(h^{(0)} + \mathrm{attn}(h^{(0)}; QK^\top)V^\top\Big)W_O^\top \\ &= \mathrm{TF}_\theta(s_{1:T}), \end{aligned}$$

as desired.

B. Multiple Parents Construction

We now present Construction 6.2.

Proof. Let $\tilde{X} \in \mathbb{R}^{T \times d}$ be the embedding of the sequence. Recall that the ℓth attention block is of the form

$$\mathrm{attn}(\tilde{X}; \widetilde{A}_\ell^{(1)}) := \mathcal{S}(\tilde{X} \widetilde{A}_\ell^{(1)} \tilde{X}^\top)\tilde{X} \in \mathbb{R}^{T \times d},$$

and

$$h_i^{(1)} = \begin{bmatrix} x_i \\ \mathrm{attn}(\tilde{X}; \widetilde{A}_1^{(1)})_i \\ \vdots \\ \mathrm{attn}(\tilde{X}; \widetilde{A}_k^{(1)})_i. \end{bmatrix} \in \mathbb{R}^{(k+1)d}.$$

The ℓth attention head performs two roles. For $i < T$, it copies $p(i)_\ell$ to the residual stream of i. Additionally, it copies $p(T+1)_\ell$ to the residual stream of T.

Formally, let $\widetilde{A}_\ell^{(1)}$ follow the same sparsity pattern as the construction in Construction 3.1, where only the position-position block $A_\ell^{(1)}$ is nonzero, and on this block let

$$(A_\ell^{(1)})_{ij} = \beta_1 \cdot \begin{cases} \mathbf{1}(j = p(i)_\ell) & i < T \\ \mathbf{1}(j = p(T+1)_\ell) & i = T \end{cases}.$$

Taking $\beta_1 \to \infty$, the output of this attention block is

$$\operatorname{attn}(\tilde{X}; \widetilde{A}_\ell^{(1)})_i = \begin{cases} \tilde{x}_{p(i)_\ell} & i \in \overline{\mathcal{R}} \setminus \{T\} \\ \tilde{x}_{p(T+1)_\ell} & i = T \end{cases}.$$

We let $\widetilde{A}^{(2)} \in \mathbb{R}^{(k+1)d \times (k+1)d}$ be the block diagonal matrix which compares the $\operatorname{attn}(\tilde{X}; \widetilde{A}_\ell^{(1)})_i$ components of the residual streams of $h_i^{(1)}$ to each other via their token embeddings.

Formally, we let

$$\widetilde{A}^{(2)} = \begin{bmatrix} 0_{d\times d} & 0_{d\times d} & 0_{d\times d} & \cdots & 0_{d\times d} \\ 0_{d\times d} & A_1^{(2)} & 0_{d\times d} & \cdots & 0_{d\times d} \\ 0_{d\times d} & 0_{d\times d} & A_2^{(2)} & \cdots & 0_{d\times d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0_{d\times d} & 0_{d\times d} & 0_{d\times d} & \cdots & A_k^{(2)} \end{bmatrix}$$

where each $A_k^{(2)} \in R^{d\times d}$ is

$$A_k^{(2)} = \beta_2 \begin{bmatrix} I_{S\times S} & 0_{S\times T} \\ 0_{T\times S} & 0_{T\times T} \end{bmatrix}.$$

We thus have, for $i \in \overline{\mathcal{R}} \setminus \{T\}$.

$$\begin{aligned} {h_i^{(1)}}^\top \widetilde{A}^{(2)} h_T^{(1)} &= \beta_2 \sum_{\ell=1}^{k} \Big(\operatorname{attn}(\tilde{X}; \widetilde{A}_\ell^{(1)})_i\Big)^\top A_k^{(2)} \operatorname{attn}(\tilde{X}; \widetilde{A}_\ell^{(1)})_T \\ &= \beta_2 \cdot \sum_{\ell=1}^{k} \mathbf{1}(s_{p(i)_\ell} = s_{p(T+1)_\ell}) \end{aligned}$$

Taking $\beta_2 \to \infty$, the softmax converges to a uniform distribution over tokens where ${h_i^{(1)}}^\top A^{(2)} h_T^{(1)}$ is maximized. These are the tokens i in which $s_{p(i)_\ell} = s_{p(T+1)_\ell}$ for all ℓ, along with the token T[2]. Thus

$$\mathcal{S}\Big(h^{(1)} \widetilde{A}^{(2)} h_T^{(1)}\Big)_i \approx \frac{\mathbf{1}_{i=T} + \mathbf{1}\big(s_{p(i)_1} = s_{p(T+1)_1}, \cdots, s_{p(i)_k} = s_{p(T+1)_k}\big)}{1 + \sum_{j<T} \mathbf{1}\big(s_{p(j)_1} = s_{p(T+1)_1}, \cdots, s_{p(j)_k} = s_{p(T+1)_k}\big)}.$$

Finally, choose W_O to output the token embedding of the x_i block of $h^{(1)}(X)_i$, so that $h^{(1)}(X) W_O = e_{s_i}$. We then have that

$$\begin{aligned} f_{\hat{\theta}}(s_{1:T})_{s'} &= \sum_i \mathbf{1}(s_i = s') \cdot \mathcal{S}\Big(h^{(1)} \widetilde{A}^{(2)} h_T^{(1)}\Big)_i \\ &\approx \frac{\mathbf{1}(s_T = s') + \sum_{i<T} \mathbf{1}\big(s_i = s', s_{p(i)_1} = s_{p(T+1)_1}, \cdots, s_{p(i)_k} = s_{p(T+1)_k}\big)}{1 + \sum_{j<T} \mathbf{1}\big(s_{p(j)_1} = s_{p(T+1)_1}, \cdots, s_{p(j)_k} = s_{p(T+1)_k}\big)} \\ &\approx \frac{\sum_i \mathbf{1}\big(s_i = s', s_{p(i)_1} = s_{p(T+1)_1}, \cdots, s_{p(i)_k} = s_{p(T+1)_k}\big)}{\sum_j \mathbf{1}\big(s_{p(j)_1} = s_{p(T+1)_1}, \cdots, s_{p(j)_k} = s_{p(T+1)_k}\big)} \\ &= \hat{\pi}_{s_{1:T}}\big(s' \mid s_{p(T+1)_1}, \ldots, s_{p(T+1)_k}\big), \end{aligned}$$

as desired. □

[2] It is possible for certain root nodes at the beginning of the sequence to be included, but this will be a vanishing fraction of tokens for typical sequences

C. Additional Experiments and Details

Single Parent Experiments: All single-parent experiments were run with a vocabulary size of $S = 10$, a sequence length of $T = 20$, a batch size of 1024, $\alpha = 0.1$, and learning rate $\eta = 0.3$. We initialize $\widetilde{A}^{(1)} = 0$, $\widetilde{A}^{(2)} = 0$, and $W_O = 0$.

In Figure 4, we repeat Figure 2(a) for the in-context learning graph in Figure 4(b), in addition to versions when the graph $\mathcal{G}$ comes from a Markov chain (Figure 4(a)) and when it is random graph (Figure 4(c)).

Multiple Parent Experiments: For experiment with multiple parents (Figure 6), we used $\alpha = 1$ and Adam (Kingma & Ba, 2017) with $\eta = 0.01$ but we initialized $\widetilde{A}^{(1)}_{ij}, \widetilde{A}^{(2)}_{ij} \sim N(0, \sigma^2)$ for $\sigma = 0.01$. This was necessary to break the symmetry between the heads.

Experiments with standard transformer architecture: We trained a two attention layer, decoder-based transformer of the form (2) on Task 2.4. We consider fixed position and token embeddings $P \in \mathbb{R}^{d \times T}$ and $E \in \mathbb{R}^{d \times S}$, with each column of P, E drawn i.i.d from $\mathcal{N}(0, \frac{1}{d} I_d)$. Additionally, between each attention layer, we add a one-hidden layer ReLU MLP with hidden width d. The model has one head per layer.

In Figure 7, we plot the average attention pattern of the first layer, averaged over a batch of 1024 sequences. We pick $S = 10, T = 20, d = 30$. The first layer attention pattern on a single sequence with embedding $h^{(0)}$ is given by

$$\mathcal{S}(\mathrm{MASK}(h^{(0)} A^{(1)} h^{(0)\top})) \in \mathbb{R}^{T \times T}.$$

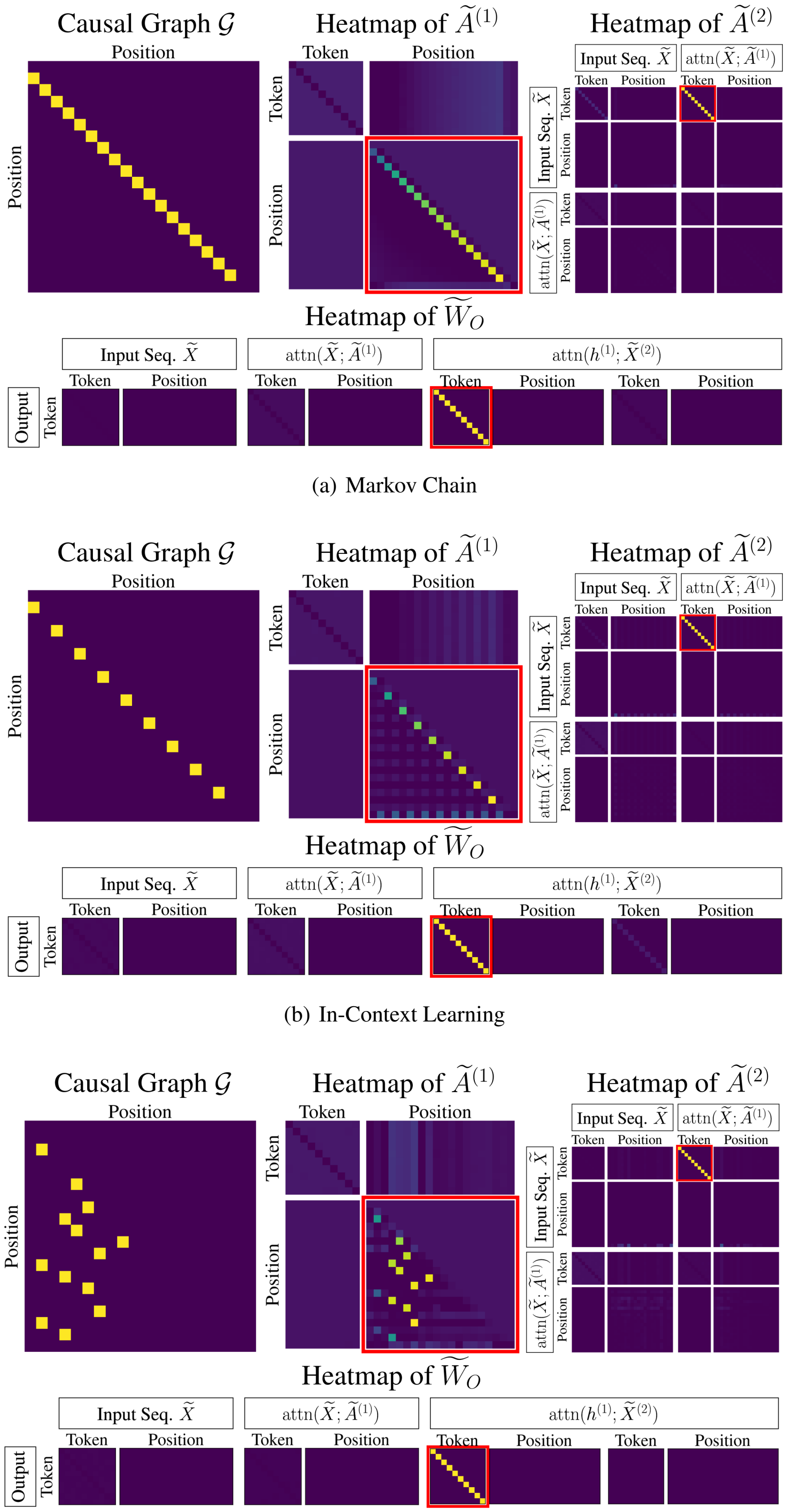
We observe that this average attention pattern is also approximately equal to the adjacency matrix of the graph $\mathcal{G}$.

Quantitative Comparison: We repeat Figure 4 for a set of 20 randomly generated causal graphs on $T = 20$ vertices and vocab size of $S = 3$. In each graph, each node i is a root node with probability $1/2$, and otherwise has its parent $p(i)$ drawn uniformly at random from the set $\{1, \ldots, i-1\}$. For each graph, one can compute the average first-layer attention weight from i to its parent $p(i)$, given by

$$\mathrm{avgattn} := \frac{1}{|\overline{\mathcal{R}}|} \sum_{i \in \overline{\mathcal{R}}} \mathcal{S}\Big(\mathrm{MASK}\Big(A^{(1)}\Big)\Big)_{i,p(i)}.$$

Over all 20 random graphs, avgattn has a mean of **0.837** and a standard deviation of **0.054**. In Figure 8, we plot the average value of $\mathcal{S}\big(\mathrm{MASK}\big(A^{(1)}\big)\big)_{i,p(i)}$ for each position i in the sequence. We observe that this attention weight is large across all positions in the sequence.

Experimental Details: Code for all the experiments can be found at https://github.com/eshnich/transformers-learn-causal-structure. All code was written in JAX (Bradbury et al., 2018), and run on a cluster of 10 NVIDIA RTX A6000 GPUs.

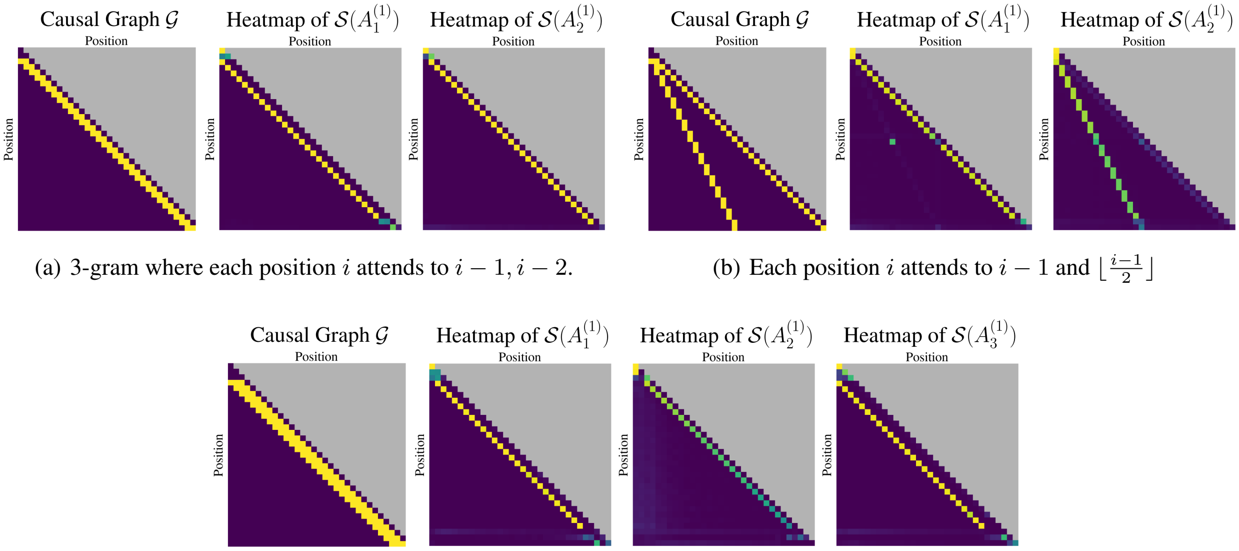


(a) Markov Chain

(b) In-Context Learning

(c) Random Causal Graph

Figure 4. $\widetilde{A}^{(1)}$ encodes the graph structure, for different latent graphs.

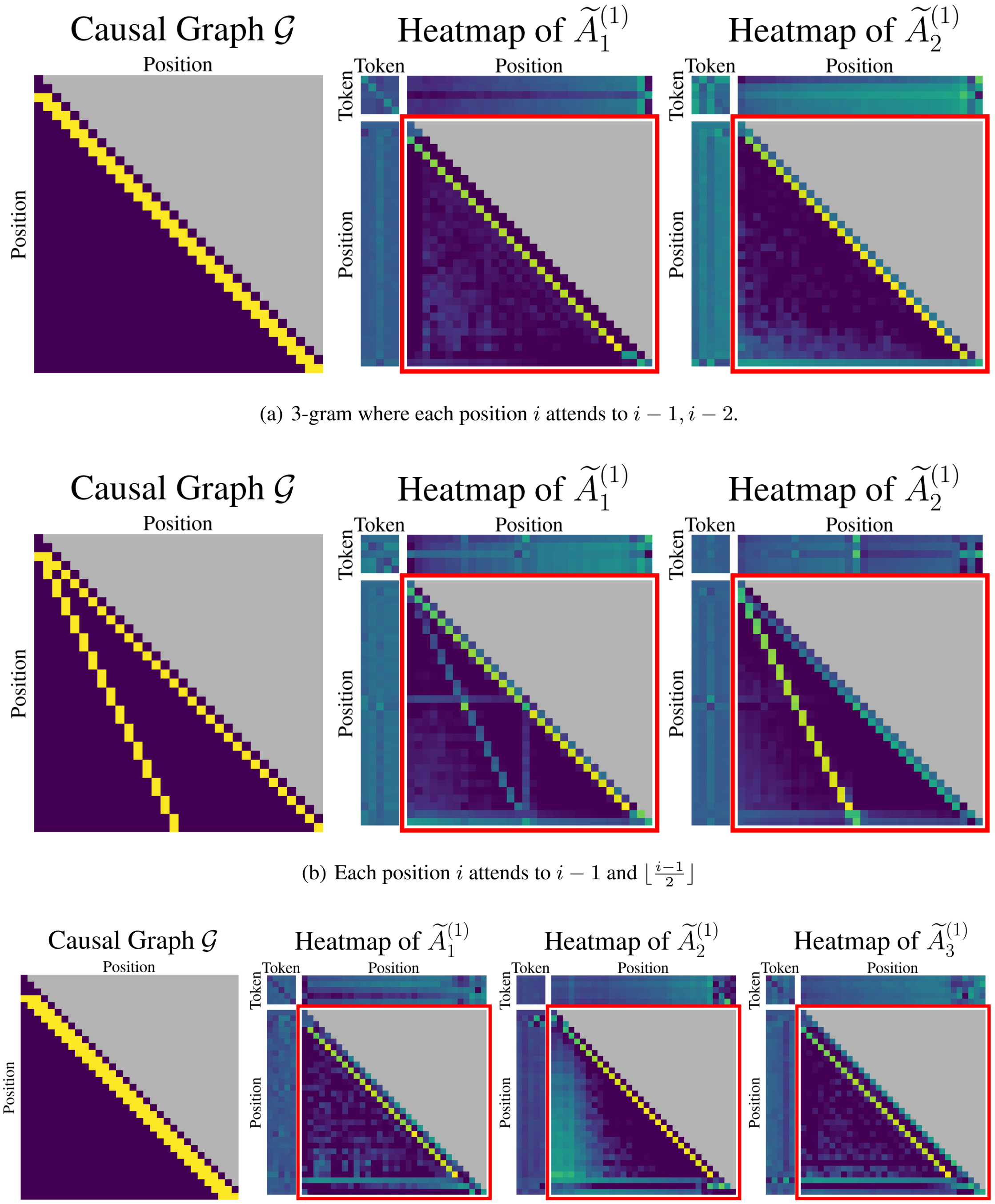


(a) 3-gram where each position i attends to $i-1, i-2$.

(b) Each position i attends to $i-1$ and $\lfloor \frac{i-1}{2} \rfloor$

(c) 4-gram where each position i attends to $i-1, i-2, i-3$.

Figure 5. **Multiple Parents:** We show three examples of trained transformers on Task 6.1 with $k = 2, 2, 3$ respectively. The left column shows the adjacency matrix of the causal graphs $\mathcal{G}$. To their right, we plot the attention patterns $\mathcal{S}(A_i^{(1)})$ for each head i where $A_i^{(1)}$ is the position-position component of $\widetilde{A}_i^{(1)}$. We see that each attention head learns a single set of parents in the causal graph $\mathcal{G}$, which agrees with Construction 6.2. See Figure 6 for plots of the full matrices $\widetilde{A}_i^{(1)}$.

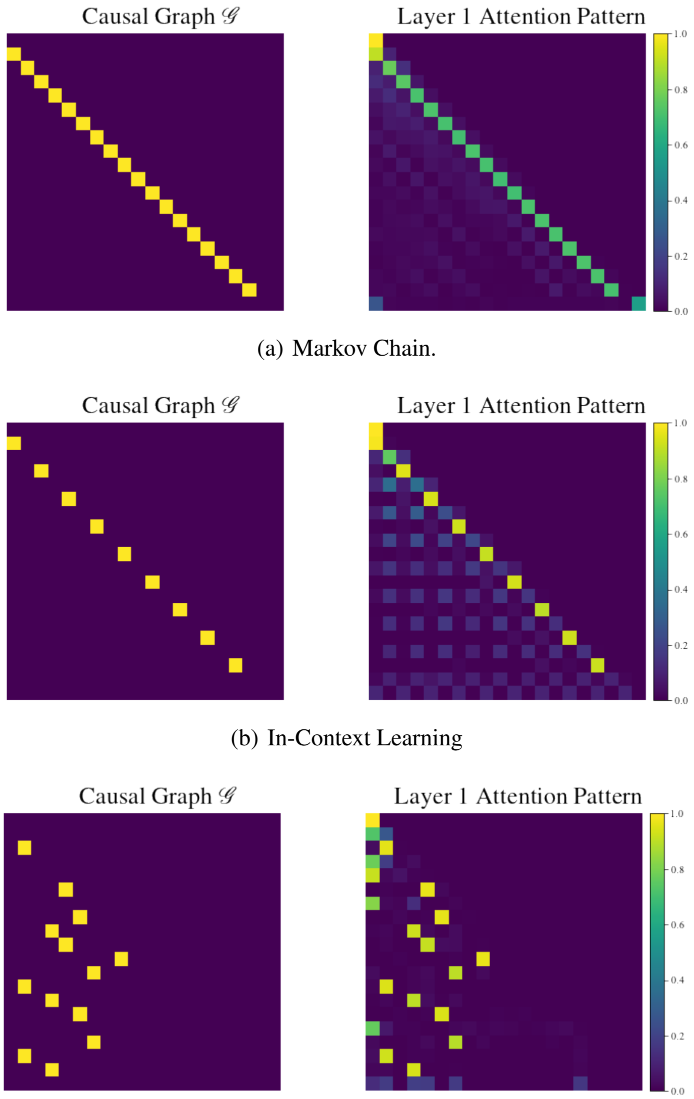


(a) 3-gram where each position i attends to $i-1, i-2$.

(b) Each position i attends to $i-1$ and $\lfloor \frac{i-1}{2} \rfloor$

(c) 4-gram where each position i attends to $i-1, i-2, i-3$

Figure 6. **Multiple Parents:** On the left, we plot the causal graph in the setting of Section 6 with $k = 2, 2, 3$ respectively. The first row corresponds to the 3-gram task in which each token depends on the previous 2. In the second row, each token at position i depends on the previous token and the token at position $\lfloor \frac{i-1}{2} \rfloor$. The third row corresponds to 4-gram in which each token depends on the previous 3 tokens. We train two-layer disentangled transformers on these tasks with k heads in each layer. On the right, we plot the first layer attention matrices, i.e. $\{\widetilde{A}_i^{(1)}\}_i$. We see that each attention head learns a single set of parents in the causal graph $\mathcal{G}$, which agrees with our Construction 6.2.



(a) Markov Chain.

(b) In-Context Learning

(c) Random Causal Graph

Figure 7. **Decoder-Based Transformer with MLPs:** In a two attention-layer, decoder-based transformer with MLPs, we observe that the average attention pattern on a sequence is approximately equal to the adjacency matrix of the causal graph $\mathcal{G}$. We remark that the random causal graph has the peculiar behavior that nodes with no parent seem to attend to the first token in the sequence.

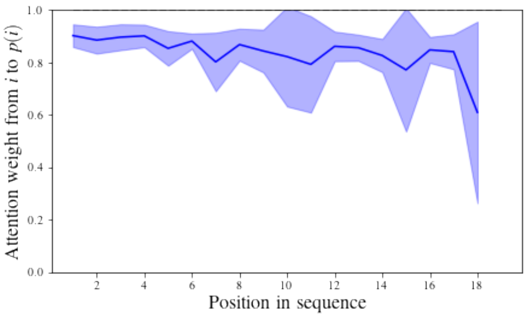


Figure 8. **Quantitative Comparison:** We plot the mean value of $\mathcal{S}(A^{(1)})_{i,p(i)}$ over all 20 graphs, as a function of the position i in the sequence. The shaded bars indicate one standard deviation. We observe that this average value is large (close to 1).

D. Analyzing the Dynamics

In this section we prove Theorem 4.4.

D.1. Proof of Lemma 3.2

Proof. The output of the first attention layer is

$$\mathrm{attn}(\tilde{X}; \widetilde{A}^{(1)}) = \mathcal{S}(\mathrm{MASK}(\tilde{X}\widetilde{A}^{(1)}\tilde{X}^\top)\tilde{X} = \mathcal{S}(\mathrm{MASK}(A^{(1)}))\tilde{X}.$$

Next, we have that

$$\begin{aligned} {h_T^{(1)}}^\top \widetilde{A}^{(2)} {h^{(1)}}^\top &= \tilde{x}_T^\top \begin{bmatrix} A^{(2)} & 0_{S\times T} \\ 0_{T\times S} & 0_{T\times T} \end{bmatrix} \mathrm{attn}(\tilde{X}; \widetilde{A}^{(1)})^\top \\ &= \tilde{x}_T^\top \begin{bmatrix} A^{(2)} & 0_{S\times T} \\ 0_{T\times S} & 0_{T\times T} \end{bmatrix} \tilde{X}^\top \mathcal{S}(\mathrm{MASK}(A^{(1)}))^\top \\ &= \overline{x}_T^\top A^{(2)} \overline{X}^\top \mathcal{S}(\mathrm{MASK}(A^{(1)}))^\top. \end{aligned}$$

Thus the output of the second attention layer is

$$\begin{aligned} \mathrm{attn}(h^{(1)}; \widetilde{A}^{(2)})_T &= {h^{(1)}}^\top \mathcal{S}\Big(h^{(1)} \Big(\widetilde{A}^{(2)}\Big)^\top h^{(T)} \Big) \\ &= {h^{(1)}}^\top \mathcal{S}\Big(\mathcal{S}(\mathrm{MASK}(A^{(1)}))\overline{X} {A^{(2)}}^\top \overline{x}_T \Big) \end{aligned}$$

Finally, the output is

$$\begin{aligned} \widetilde{\mathrm{TF}}_{\widetilde{\theta}}(s_{1:T}) &= \widetilde{W}_O^\top h_T^{(2)} \\ &= \big[I_S \quad 0_{S\times T} \mid 0_{S\times d} \big] \, \mathrm{attn}(h^{(1)}; \widetilde{A}^{(2)})_T \\ &= \big[I_S \quad 0_{S\times T} \mid 0_{S\times d} \big] \, {h^{(1)}}^\top \mathcal{S}\Big(\mathcal{S}(\mathrm{MASK}(A^{(1)}))\overline{X} {A^{(2)}}^\top \overline{x}_T \Big) \\ &= \overline{X}^\top \mathcal{S}\Big(\mathcal{S}(\mathrm{MASK}(A^{(1)}))\overline{X} {A^{(2)}}^\top \overline{x}_T \Big), \end{aligned}$$

as desired. □

D.2. Notation

We briefly introduce notation which will be used throughout the rest of the appendix. We let $X \in \mathbb{R}^{T\times S}$ be the token embedding of the sequence $s_{1:T}$. Additionally, for a lower triangular matrix $A \in \mathbb{R}^{T\times T}$, let $A_i \in \mathbb{R}^i$ denote the first i coordinates of the ith row of A. We overload notation so that $\mathcal{S}(A) \in \mathbb{R}^{T\times T}$ is the lower triangular matrix satisfying $\mathcal{S}(A)_i = \mathcal{S}(A_i)$; i.e, the softmax operation is applied row-wise to the first i coordinates of row i. Finally, we reparameterize ${A^{(2)}}^\top$ with $A^{(2)}$.

We can thus rewrite the reduced model f_θ as

$$f_\theta(s_{1:T}) = X^\top \mathcal{S}\Big(\mathcal{S}(A^{(1)}) X A^{(2)} x_T \Big).$$

We let $f_\theta(X; s)$ denote prediction of a transformer with embedding $X \in \mathbb{R}^{T\times S}$ conditioned on $s_T = s$, i.e

$$f_\theta(X; s) := X^\top \mathcal{S}\Big(\mathcal{S}(A^{(1)}) X A^{(2)} e_s \Big).$$

It is easy to see that the perturbed loss (11) can be written as

$$L(\theta) = -\frac{1}{S}\,\mathbb{E}_{\pi,X}\left[\sum_{s,s' \in [S]} \pi(s' \mid s) \log\left(f_\theta(X; s)_{s'} + \epsilon \right) \right],$$

where we use $\mathbb{E}_X$ and $\mathbb{E}_{s_{1:T}}$ interchangeably to represent expectation over the sequence $s_{1:T}$. We set the perturbation as $\epsilon = T_{\text{eff}}^{-1/2}$.

For notational convenience, we define

$$v_\theta(X; s) := \mathcal{S}(\mathcal{S}(A^{(1)})XA^{(2)}e_s),$$

so that $f_\theta(X; s) = X^\top v_\theta(X; s)$.

Let $\delta_s(X) \in \mathbb{R}^T$ be the vector where $\delta_s(X)_i = x_{i,s}$, and let $\hat{\mu}_X(s) := \frac{1}{T}\sum_{i=1}^T x_{i,s}$ be the empirical estimate of the frequency of s over the sequence X. We let $X_{\le i} \in \mathbb{R}^{i \times S}$ be the embedding of the first i tokens in the sequence, and let $\delta_s(X_{\le i}) \in \mathbb{R}^i$ be the indicator of s on these first i tokens.

Given a vector $v \in \mathbb{R}^k$, the operator $J_k : \mathbb{R}^k \to \mathbb{R}^{k \times k}$ is given by $J_k(v) = diag(v) - vv^\top$. J_k is the Jacobian of $\mathcal{S}$: $\nabla_u \mathcal{S}(u) = J_k(\mathcal{S}(u))$. We drop the subscript k when it is clear from context.

D.3. Heuristic Derivation of Lemma 5.3

During stage 1, the model can be rewritten as

$$f_\theta(X; s)_{s'} = e_{s'}^\top X^\top \mathcal{S}\Big(\beta_0 \mathcal{S}(A^{(1)})Xe_s\Big) = \delta_{s'}(X)^\top \mathcal{S}\Big(\beta_0 \mathcal{S}(A^{(1)})\delta_s(X)\Big).$$

When $\beta_0 \approx 0$, we can linearize the outer softmax as

$$\mathcal{S}(\beta_0 z) \approx \frac{1}{T}1_T + \beta_0 \cdot \left(\frac{1}{T}I_T - \frac{1}{T^2}1_T 1_T^\top\right)z,$$

and get that

$$\begin{aligned} f_\theta(X; s)_{s'} &\approx \frac{1}{T}\delta_{s'}(X)^\top 1_T + \frac{\beta_0}{T}\delta_{s'}(X)^\top \mathcal{S}(A^{(1)})\delta_s(X) - \frac{\beta_0}{T^2}\delta_{s'}(X)^\top 1_T \cdot 1_T^\top \mathcal{S}(A^{(1)})\delta_s(X) \\ &= \hat{\mu}_X(s') + \frac{\beta_0}{T}\Big(\delta_{s'}(X)^\top \mathcal{S}(A^{(1)})\delta_s(X) - \hat{\mu}_X(s') \cdot 1_T^\top \mathcal{S}(A^{(1)})\delta_s(X)\Big). \end{aligned} \tag{26}$$

First, observe that since $\beta_0 \approx 0$,

$$f_\theta(X; s)_{s'} \approx \hat{\mu}_X(s')$$

Next, taking the gradient of the approximation (26) with respect to $A_i^{(1)}$ yields

$$\nabla_{A_i^{(1)}} f_\theta(X; s)_{s'} \approx \frac{\beta_0}{T} J\Big(\mathcal{S}(A_i^{(1)})\Big)\delta_s(X_{\le i}) \cdot (x_{i,s'} - \hat{\mu}_X(s')).$$

Therefore by the chain rule, the population gradient is given by

$$\begin{aligned} \nabla_{A_i^{(1)}} L(\theta) &\approx -\frac{1}{S}\mathbb{E}_{\pi,X}\left[\sum_{s,s'} \frac{\pi(s' \mid s)}{f_\theta(X; s)_{s'} + \epsilon} \nabla_{A_i^{(1)}} f_\theta(X; s)_{s'}\right]. \\ &\approx -\frac{\beta_0}{ST} J\Big(\mathcal{S}(A_i^{(1)})\Big) \cdot \mathbb{E}_{\pi,X}\left[\sum_{s,s'} \frac{\pi(s' \mid s)}{\hat{\mu}_X(s')}(x_{i,s'} - \hat{\mu}_X(s'))\delta_s(X_{\le i})\right]. \end{aligned}$$

Letting $\hat{g}_i$ denote the term after the preconditioner, i.e $\hat{g}_i := \mathbb{E}_{\pi,X}\Big[\sum_{s,s'} \frac{\pi(s'|s)}{\hat{\mu}_X(s')}(x_{i,s'} - \hat{\mu}_X(s'))\delta_s(X_{\le i})\Big]$, we get that the

jth entry of $\hat{g}_i$, $\hat{g}_{i,j}$, is

$$\begin{aligned}
\hat{g}_{i,j} &= \mathbb{E}_{\pi,X}\left[\sum_{s,s'} \frac{\pi(s' \mid s)}{\hat{\mu}_X(s')}(x_{i,s'} - \hat{\mu}_X(s'))x_{j,s}\right] \\
&= \mathbb{E}_{\pi,X}\left[\sum_{s,s'} \frac{\pi(s' \mid s)}{\hat{\mu}_X(s')}x_{i,s'}x_{j,s} - \sum_{s,s'} \pi(s' \mid s)x_{j,s}\right] \\
&= \mathbb{E}_{\pi,X}\left[\sum_{s,s'} \frac{\pi(s' \mid s)}{\hat{\mu}_X(s')}x_{i,s'}x_{j,s}\right] - 1.
\end{aligned}$$

Conditioned on π, as the effective length of the sequence X grows large, due to our assumptions on P_π the sequence $x_1, \dots, x_T$ mixes, and thus $\hat{\mu}_X(s') \to \mu_\pi(s)$. As such,

$$\begin{aligned}
\hat{g}_{i,j} &\approx \mathbb{E}_{\pi}\left[\sum_{s,s'} \frac{\pi(s' \mid s)}{\mu_\pi(s')}\mathbb{E}_X[x_{i,s'}x_{j,s}]\right] - 1 \\
&= \mathbb{E}_{\pi}\left[\sum_{s,s'} \frac{\pi(s' \mid s)}{\mu_\pi(s')}\mathbb{P}_X[s_j = s, s_i = s']\right] - 1 \\
&= g_{i,j}.
\end{aligned}$$

D.4. Gradient Computations

Recall that $A^{(1)}_i \in \mathbb{R}^i$ is the ith row of $A^{(1)}$. Define the population gradients as

$$\begin{aligned}
G^{(1)}(A^{(1)}, A^{(2)})_i &:= \nabla_{A^{(1)}_i} L(\theta)\big|_{\theta=(A^{(1)},A^{(2)})} \\
G^{(2)}(A^{(1)}, A^{(2)}) &:= \nabla_{A^{(2)}} L(\theta)\big|_{\theta=(A^{(1)},A^{(2)})}.
\end{aligned}$$

The following lemma computes the population gradients:

Lemma D.1 (Population gradients)**.**

$$\begin{aligned}
G^{(1)}(A^{(1)}, A^{(2)})_i &= -\frac{1}{S} J(\mathcal{S}(A^{(1)}_i)) \sum_{s,s'} \mathbb{E}_{\pi,X}\left[\frac{\pi(s' \mid s)}{f_\theta(X;s)_{s'} + \epsilon}\delta_{s'}(X)^\top J(v_\theta(X;s))e_i \cdot X_{\le i}A^{(2)}e_s\right] \\
G^{(2)}(A^{(1)}, A^{(2)}) &= -\frac{1}{S}\sum_{s,s'} \mathbb{E}\left[\frac{\pi(s' \mid s)}{f_\theta(X;s)_{s'} + \epsilon} \cdot X^\top \mathcal{S}(A^{(1)})^\top J(v_\theta(X;s))\delta_{s'}(X)e_s^\top\right]
\end{aligned}$$

Proof. The model gradient with respect to $A^{(1)}_i$ is

$$\nabla_{A^{(1)}_i} f_\theta(X;s) = X^\top J(v_\theta(X;s))e_i \otimes J(\mathcal{S}(A^{(1)}_i))X_{\le i}A^{(2)}e_s$$

Therefore the loss gradient is given by

$$\begin{aligned}
G^{(1)}(A^{(1)}, A^{(2)})_i &= -\frac{1}{S}\sum_{s,s'} \mathbb{E}\left[\frac{\pi(s' \mid s)}{f_\theta(X;s)_{s'} + \epsilon}\nabla f_\theta(X;s)_{s'}\right] \\
&= -\frac{1}{S} J(\mathcal{S}(A^{(1)}_i)) \sum_{s,s'} \mathbb{E}\left[\frac{\pi(s' \mid s)}{f_\theta(X;s)_{s'} + \epsilon}\delta_{s'}(X)^\top J(v_\theta(X;s))e_i \cdot X_{\le i}A^{(2)}e_s\right].
\end{aligned}$$

Next, the model gradient of $A^{(2)}$ is

$$\nabla_{A^{(2)}} f_\theta(X;s)_{s'} = X^\top \mathcal{S}(A^{(1)})^\top J(v_\theta(X;s))\delta_{s'}(X)e_s^\top.$$

Thus

$$G^{(2)}(A^{(1)}, A^{(2)}) = -\frac{1}{S}\sum_{s,s'} \mathbb{E}\left[\frac{\pi(s' \mid s)}{f_\theta(X; s)_{s'} + \epsilon} \cdot X^\top \mathcal{S}(A^{(1)})^\top J(v_\theta(X; s))\delta_{s'}(X) e_s^\top\right]$$

□

D.5. Gradient of $A^{(1)}$ (Stage 1)

We show that during the first stage of training, $A^{(1)}$ converges to the adjacency matrix of the graph $\mathcal{G}$.

The first step is to show that a quantity called the "idealized gradient" approximately aligns with the adjacency matrix of $\mathcal{G}$. For a transition matrix π, define

$$g_{i,j}(\pi) := \sum_{s,s'} \frac{\pi(s' \mid s)}{\mu_\pi(s')} \cdot \mathbb{P}_X[s_i = s', s_j = s] - 1,$$

and let $g_{i,j} := \mathbb{E}_\pi[g_{i,j}(\pi)]$.

The following lemma shows that this idealized gradient is maximized at $j = p(i)$. The proof relies on the data processing inequality argument, and is deferred to Appendix G.1.

Lemma D.2 (Idealized gradient is aligned with $\mathcal{G}$)**.** *If $p(i) \neq \emptyset$, then*

$$g_{i,p(i)} \geq g_{i,j} + \frac{\gamma^3}{2S}$$

for all $j \in [i] \setminus p(i)$. Otherwise $g_{i,j} = 0$.

Next, we show that the true gradient with respect to $A^{(1)}$ can indeed be approximated by this idealized gradient, and hence the adjacency matrix of $\mathcal{G}$.

Lemma D.3 (True gradient of $A^{(1)}$ is aligned with $\mathcal{G}$ (Stage 1))**.** *Let $A^{(2)} = \beta_0 I$. There exist constants $c_{\gamma,S}, C_{\gamma,S}$ such that, if $\beta_0 \leq c_{\gamma,S} T_{\mathit{eff}}^{-3/2}$,*

- *If $p(i) = \emptyset$,*

  $$G^{(1)}(A^{(1)}, A^{(2)})_i = J(\mathcal{S}(A_i^{(1)}))v$$

  for v with $\|v\|_\infty \leq C_{\gamma,S}\frac{\beta_0}{T\sqrt{T_{\mathit{eff}}}}$.

- *If $p(i) \neq \emptyset$, then for any $j \neq p(i)$,*

  $$G^{(1)}(A^{(1)}, A^{(2)})_{i,p(i)} \leq G^{(1)}(A^{(1)}, A^{(2)})_{i,j} - \mathcal{S}(A_i^{(1)})_{p(i)}\Big(1 - \mathcal{S}(A_i^{(1)})_{p(i)}\Big) \cdot \frac{C_{\gamma,S}\beta_0}{T}.$$

Proof. First, see that

$$X_{\leq i} A^{(2)} e_s = \beta_0 X_{\leq i} e_s = \beta_0 \delta_s(X_{\leq i}).$$

Thus

$$G^{(1)}(A^{(1)}, A^{(2)})_i = -\beta_0 J(\mathcal{S}(A_i^{(1)})) \cdot \frac{1}{S}\sum_{s,s'} \mathbb{E}_{\pi,X}\left[\frac{\pi(s' \mid s)}{f_\theta(X; s)_{s'} + \epsilon}\delta_{s'}(X)^\top J(v_\theta(X; s)) e_i \cdot \delta_s(X_{\leq i})\right].$$

Let $\hat\theta := (A^{(1)}, 0)$, and define the quantities $g_i^*, \hat g_i$ by

$$g_i^* := T\sum_{s,s'} \mathbb{E}_{\pi,X}\left[\frac{\pi(s' \mid s)}{f_\theta(X; s)_{s'} + \epsilon}\delta_{s'}(X)^\top J(v_\theta(X; s)) e_i \cdot \delta_s(X_{\leq i})\right],$$

$$\hat g_i := T\sum_{s,s'} \mathbb{E}_{\pi,X}\left[\frac{\pi(s' \mid s)}{f_{\hat\theta}(X; s)_{s'} + \epsilon}\delta_{s'}(X)^\top J(v_{\hat\theta}(X; s)) e_i \cdot \delta_s(X_{\leq i})\right].$$

We remark that

$$G^{(1)}(A^{(1)}, A^{(2)})_i = -\frac{\beta_0}{ST} J(\mathcal{S}(A_i^{(1)}))g_i^*. \tag{27}$$

Since $\beta_0 \le c_{\gamma,S}\frac{1}{T_{\text{eff}}^{3/2}} \le 1$, by Lemma G.1 we have

$$\|\hat{g}_i - g_i^*\|_\infty \le 6S^2\epsilon^{-2}\beta_0 \le C_{\gamma,S}\frac{1}{\sqrt{T_{\text{eff}}}}.$$

It thus suffices to analyze $\hat{g}_i$. Note that $v_{\hat{\theta}}(X; s) = \frac{1}{T}1_T$. Therefore

$$f_{\hat{\theta}}(X; s)_{s'} = \frac{1}{T}1_T^\top \delta_{s'}(X) = \hat{\mu}_X(s').$$

and

$$\delta_{s'}(X)^\top J(v_{\hat{\theta}}(X; s))e_i = \delta_{s'}(X)^\top \left(\frac{1}{T}I_T - \frac{1}{T^2}1_T 1_T^\top\right)e_i = \frac{1}{T}(x_{i,s'} - \hat{\mu}_X(s')).$$

The jth entry of $\hat{g}_i$, $\hat{g}_{i,j}$, is thus equal to

$$\hat{g}_{i,j} = \sum_{s,s'} \mathbb{E}_{\pi,X}\left[\frac{\pi(s' \mid s)}{\hat{\mu}_X(s') + \epsilon}(x_{i,s'} - \hat{\mu}_X(s'))x_{j,s}\right].$$

By Lemma G.2, this is approximately equal to the idealized gradient $g_{i,j}$:

$$|\hat{g}_{i,j} - g_{i,j}| \le C_{\gamma,S}\frac{1}{\sqrt{T_{\text{eff}}}}.$$

We are now ready to prove the theorem. First, consider the case where $p(i) = \emptyset$. By Lemma D.2, $g_{i,j} = 0$, and thus

$$\left|g_{i,j}^*\right| \lesssim \frac{1}{\sqrt{T_{\text{eff}}}}$$

Since $G^{(1)}(A^{(1)}, A^{(2)})_i = -\frac{\beta_0}{ST}J(\mathcal{S}(A_i^{(1)}))g_i^*$, the claim follows.

Otherwise if $p(i) \ne \emptyset$, Lemma D.2 tells us that, for all $j \ne p(i)$,

$$g_{i,j}^* - g_{i,p(i)}^* \le g_{i,j} - g_{i,p(i)} + \left|g_{i,j} - g_{i,j}^*\right| + \left|g_{i,p(i)} - g_{i,p(i)}^*\right| \le -\frac{\gamma^3}{2S} + C_{\gamma,S}\frac{1}{\sqrt{T_{\text{eff}}}} \le -\frac{\gamma^3}{4S}.$$

Next, see that

$$G^{(1)}(A^{(1)}, A^{(2)})_{i,j} = -\frac{\beta_0}{ST}\left(\mathcal{S}(A_i^{(1)})_j g_{i,j}^* - \mathcal{S}(A_i^{(1)})^\top g_i^* \mathcal{S}(A_i^{(1)})_j\right).$$

Therefore for any $j \ne p(i)$, we can bound

$$\begin{aligned}
&G^{(1)}(A^{(1)}, A^{(2)})_{i,j} - G^{(1)}(A^{(1)}, A^{(2)})_{i,p(i)} \\
&= \frac{\beta_0}{ST}\left[\left(\mathcal{S}(A_i^{(1)})_{p(i)} - \mathcal{S}(A_i^{(1)})_j\right)\left(g_{i,p(i)}^* - \mathcal{S}(A_i^{(1)})^\top g_i^*\right) + \mathcal{S}(A_i^{(1)})_j(g_{i,p(i)}^* - g_{i,j}^*)\right] \\
&\ge \frac{\beta_0}{ST}\left[\left(\mathcal{S}(A_i^{(1)})_{p(i)} - \mathcal{S}(A_i^{(1)})_j\right)\left(1 - \mathcal{S}(A_i^{(1)})_{p(i)}\right)\frac{\gamma^3}{4S} + \mathcal{S}(A_i^{(1)})_j\frac{\gamma^3}{4S}\right] \\
&\ge \mathcal{S}(A_i^{(1)})_{p(i)}\left(1 - \mathcal{S}(A_i^{(1)})_{p(i)}\right) \cdot \frac{\beta_0\gamma^3}{4S^2T},
\end{aligned}$$

as desired. □

We can now analyze the gradient descent dynamics over multiple timesteps. First, we show that for most root nodes $i \in \mathcal{R}$, $A_i^{(1)}$ moves very little.

Lemma D.4. *Let $i \in \mathcal{R}$. Then*

$$\left|\mathcal{S}(A_i^{(1)}(\tau))_j - \frac{1}{i}\right| \lesssim \frac{\tau\eta_1\beta_0}{T\sqrt{T_{\mathit{eff}}} \cdot i^2}$$

for all $j \leq i$.

Proof. Let $r(A_i^{(1)}) = \max_j A_{i,j}^{(1)} - \min_j A_{i,j}^{(1)}$. We have that (where v is the vector from Lemma D.3),

$$\left\|G^{(1)}(A^{(1)}, A^{(2)})_i\right\|_\infty \leq \max_j \mathcal{S}(A_i^{(1)})_j \cdot \|v\|_\infty,$$

and thus

$$r(A_i^{(1)}(t+1)) \leq r(A_i^{(1)}(t)) + 2\eta_1 \max_j \mathcal{S}(A_i^{(1)}(t))_j \cdot \|v\|_\infty.$$

Fix $\omega \leq 1$. Assume there exists some $t \leq \tau$ such that $r(A_i^{(1)}(t)) > \log(1+\omega)$, and let t^* be the first such time t. We can always bound

$$\max_j \mathcal{S}(A_i^{(1)}(t))_j \leq \frac{\exp\Big(r(A_i^{(1)}(t))\Big)}{(i-1) + \exp\Big(r(A_i^{(1)}(t))\Big)},$$

and thus for $t < t^*$, $\max_j \mathcal{S}(A_i^{(1)}(t))_j \leq \frac{1+\omega}{i+\omega} \leq \frac{1+\omega}{i}$. Therefore

$$\log(1+\omega) < r(A_i^{(1)}(t^*)) \leq 2\tau\eta_1 \|v\|_\infty i^{-1} \cdot (1+\omega),$$

Bounding $\log(1+\omega) \geq \omega/2$ and $1+\omega \leq 2$, we get that

$$\omega \leq 8\tau\eta_1 \|v\|_\infty i^{-1} \lesssim \frac{\tau\eta_1\beta_0}{T\sqrt{T_{\text{eff}}} \cdot i}.$$

Additionally, when $r(A_i^{(1)}(t)) \leq \log(1+\omega)$, we have the bound

$$\frac{1}{i}(1-\omega) \leq \frac{1}{1+(1+\omega)(i-1)} \leq \mathcal{S}(A_i^{(1)}(t))_j \leq \frac{1}{i}(1+\omega).$$

Therefore

$$\left|\mathcal{S}(A_i^{(1)}(\tau))_j - \frac{1}{i}\right| \leq \frac{\omega}{i} \lesssim \frac{\tau\eta_1\beta_0}{T\sqrt{T_{\text{eff}}} \cdot i^2},$$

as desired. □

Next, we bound the time it takes until $\mathcal{S}\big(A^{(1)}(t)\big)_{i,p(i)} \approx 1$.

Lemma D.5. *Let $A^{(2)}(0) = \beta_0 I_S$, where $\beta_0 \leq c_{\gamma,S} \frac{1}{T_{\mathit{eff}}^{3/2}}$. There exists $\tau_1 \lesssim \eta_1^{-1}\beta_0^{-1}(T^2 + T\alpha^{-1})\log(T/\alpha)$ such that, for any $t \geq \tau_1$,*

$$\mathcal{S}\Big(A^{(1)}(t)\Big)_{i,p(i)} \geq 1 - \alpha.$$

for all i with $p(i) \neq \emptyset$.

Proof. By induction, one has that $A^{(1)}(t)_{i,p(i)} \geq A^{(1)}(t)_{i,j}$ throughout training. Thus $\mathcal{S}\big(A^{(1)}(t)\big)_{i,p(i)} \geq \frac{1}{T}$. Additionally, by Lemma D.3, one has that $\mathcal{S}\big(A^{(1)}(t)\big)_{i,p(i)}$ is increasing in t.

Fix i. Define $\Delta(t) = A^{(1)}(t)_{i,p(i)} - \max_{j \neq p(i)} A^{(1)}(t)_{i,j}$. One sees that

$$\mathcal{S}\Big(A^{(1)}(t)\Big)_{i,p(i)} \geq \frac{\exp(\Delta(t))}{T + \exp(\Delta(t))}.$$

Let $\tau^+(1/2)$ be the first time t at which $\mathcal{S}\big(A^{(1)}(t)\big)_{i,p(i)} > \frac{1}{2}$. For $t < \tau^+(1/2)$ we have $1 - \mathcal{S}\big(A^{(1)}(t)\big)_{i,p(i)} \geq \frac{1}{2}$, and thus by Lemma D.3,

$$\Delta(t+1) \geq \Delta(t) + \frac{C_{\gamma,S}\beta_0}{T^2}\eta_1.$$

Therefore $\Delta(\tau^+(1/2)) \gtrsim \frac{\beta_0\eta_1}{T^2}\tau^+(1/2)$. Assume that $\Delta(\tau^+(1/2)) \geq \log(2T)$. Then

$$\mathcal{S}\Big(A^{(1)}(\tau^+(1/2))\Big)_{i,p(i)} \geq \frac{\exp(\log(2T))}{T + \exp(\log(2T))} = \frac{2}{3},$$

a contradiction. Thus $\Delta(\tau^+(1/2)) \leq \log(2T)$, so $\tau^+(1/2) \lesssim T^2\eta_1^{-1}\beta_0^{-1}\log(2T)$.

Let $\tau^+(\alpha)$ be the first time at which $\mathcal{S}(A^{(1)}(\tau^+(\alpha))_{i,p(i)} < 1 - \alpha$. For $\tau^+(1/2) \leq t < \tau^+(\alpha)$, we then have

$$\Delta(t+1) \geq \Delta(t) + \frac{C_{\gamma,S}\beta_0\alpha}{T}\eta_1,$$

and thus if $\tau^+(\alpha) - \tau^+(1/2) \gtrsim T\alpha^{-1}\beta_0^{-1}\log(T/\alpha)$,

$$\Delta(\tau^+(\alpha)) \geq \frac{C_{\gamma,S}\beta_0\alpha}{T}\eta_2(\tau^+(\alpha) - \tau^+(1/2)) \geq \log\left(\frac{T}{\alpha}\right)$$

Then

$$\mathcal{S}\Big(A^{(1)}(\tau^+(\alpha))_{i,p(i)}\Big) \geq \frac{\exp(\log(T/\alpha))}{T + \exp(\log(T/\alpha))} = \frac{\frac{1}{\alpha}}{1 + \frac{1}{\alpha}} \geq 1 - \alpha,$$

a contradiction. Thus $\tau^+(\alpha) - \tau^+(1/2) \lesssim T\alpha^{-1}\beta_0^{-1}\log(T/\alpha)$, and so $\tau^+(\alpha) \lesssim T^2\eta_1^{-1}\beta_0^{-1}\log(2T) + T\alpha^{-1}\beta_0^{-1}\log(T/\alpha) \lesssim \eta_1^{-1}\beta_0^{-1}(T^2 + T\alpha^{-1})\log(T/\alpha)$, as desired. □

Combining the previous two lemmas, the following corollary tells us the value of $A^{(1)}$ after stage 1 of the algorithm.

Corollary D.6 (Ouptut of stage 1)**.** *Let $\beta_0 \leq c_{\gamma,S}\frac{1}{T_{\mathit{eff}}^{3/2}}$, and set $\tau_1 = C_{\gamma,S}\eta_1^{-1}\beta_0^{-1}T^2\log(T)$ for appropriately chosen constants $c_{\gamma,S}, C_{\gamma,S}$. Then:*

- *If $i \in \overline{\mathcal{R}}$,*

$$1 - \mathcal{S}\Big(A^{(1)}(\tau_1)\Big)_{i,p(i)} \lesssim T^{-1},$$

- *If $i \in \mathcal{R}$,*

$$\sup_{j \in [i]}\left|\mathcal{S}(A_i^{(1)}(\tau_1))_j - \frac{1}{i}\right| \lesssim \min\left(1, \frac{T\log T}{\sqrt{T_{\mathit{eff}}}\cdot i^2}\right).$$

Proof. This follows directly from plugging in $\tau = \tau_1$ into Lemma D.4 and selecting $\alpha = \Theta(T^{-1})$ in Lemma D.5. □

D.6. Gradient of $A^{(2)}$ (Stage 2)

First, we observe that the population dynamics of $A^{(2)}$ possess a certain symmetry:

Lemma D.7. *For all time, $A^{(2)} = \beta_0 I_S + \beta(I_S - \frac{1}{S}1_S 1_S^\top)$ for some scalar β.*

Proof. If $A^{(2)} = \beta_1 I_S + \beta_2 1_S 1_S^\top$ (all diagonals are equal and all off-diagonals are equal), then by symmetry the gradient is also of this form. Additionally, see that

$$\begin{aligned}
1_S^\top G^{(2)}(A^{(1)}, A^{(2)}) &= -\frac{1}{S}\sum_{s,s'} \mathbb{E}_{\pi,X}\left[\frac{\pi(s' \mid s)}{f_\theta(X;s)_{s'} + \epsilon} \cdot 1_S^\top X^\top \mathcal{S}(A^{(1)})^\top J(v_\theta(X;s))\delta_{s'}(X) e_s^\top\right] \\
&= -\frac{1}{S}\sum_{s,s'} \mathbb{E}_{\pi,X}\left[\frac{\pi(s' \mid s)}{f_\theta(X;s)_{s'} + \epsilon} \cdot 1_T^\top \mathcal{S}(A^{(1)})^\top J(v_\theta(X;s))\delta_{s'}(X) e_s^\top\right] \\
&= -\frac{1}{S}\sum_{s,s'} \mathbb{E}_{\pi,X}\left[\frac{\pi(s' \mid s)}{f_\theta(X;s)_{s'} + \epsilon} \cdot 1_T^\top J(v_\theta(X;s))\delta_{s'}(X) e_s^\top\right] \\
&= 0,
\end{aligned}$$

since $J(v_\theta(X;s))1_T = 0$. Therefore $G^{(2)}(A^{(1)}, A^{(2)}) = \beta \cdot \left(I_S - \frac{1}{S}1_S 1_S^\top\right)$ for some scalar β. Since we initialize $A^{(2)} = \beta_0 I$, throughout training $A^{(2)}$ is of the form $A^{(2)} = \beta_0 I_S + \beta(I_S - \frac{1}{S}1_S 1_S^\top)$. □

Throughout the rest of the proof, we let $\beta(t)$ be the scalar such that

$$A^{(2)}(t) = \beta(t) I_S - (\beta(t) - \beta_0)\frac{1}{S}1_S 1_S^\top.$$

The goal of this section is to show that when $A^{(1)}$ approximates the adjacency matrix of $\mathcal{G}$, $\beta(t)$ will grow large. Since the gradient descent update for $A^{(2)}$ is

$$A^{(2)}(t+1) = A^{(2)}(t) - \eta_2 G^{(2)}(A^{(1)}(t), A^{(2)}(t)),$$

the update for $\beta(t)$ is

$$\beta(t+1) = \beta(t) - \eta_2 \cdot \frac{1}{S-1}\operatorname{Tr}\Big(G^{(2)}(A^{(1)}(t), A^{(2)}(t))\Big).$$

As such, we define the quantity $\Delta_\beta(\theta)$ by

$$\Delta_\beta(\theta) := \frac{1}{S-1}\operatorname{Tr}\Big(G^{(2)}(A^{(1)}, A^{(2)})\Big).$$

Finally, for notational convenience, let $A_*^{(1)}$ be the $T \times T$ matrix such that

$$\mathcal{S}\Big(A_*^{(1)}\Big)_{ij} = \begin{cases} \mathbf{1}(j = p(i)) & \text{if } i \in \overline{\mathcal{R}} \\ \mathcal{S}(A^{(1)}(\tau_1))_{i,j} & \text{if } i \in \mathcal{R} \end{cases}.$$

$A_*^{(1)}$ encodes the adjacency matrix of $\mathcal{G}$ on nodes i where $p(i) \neq \emptyset$.

Lemma D.8 (Stage 2)**.** *Let $\theta = (A^{(1)}, A^{(2)})$, where $A^{(1)} = A^{(1)}(\tau_1)$ is the output of stage 1, and $A^{(2)} = \beta I_S - (\beta - \beta_0)\frac{1}{S}1_S 1_S^\top$ for $\beta \geq 0$. If β satisfies*

$$\exp(\beta) \leq \exp(\beta^*) := C_{\gamma,S} T_{\mathit{eff}}^{1/12} \log^{-1/6} T,$$

then

$$1 \geq -\Delta_\beta(\theta) \geq \frac{1}{4}\gamma^8 S^{-6} e^{-2\beta} > 0.$$

Proof. Note that $XA^{(2)}e_s = \beta X e_s - (\beta - \beta_0)\frac{1}{S}1_T$. Since the row sums of $\mathcal{S}(A^{(1)})$ are 1,

$$\mathcal{S}(A^{(1)})XA^{(2)}e_s = \beta\mathcal{S}(A^{(1)})Xe_s - \frac{\beta-\beta_0}{S}1_T,$$

and thus

$$v_\theta(X;s) = \mathcal{S}(\beta\mathcal{S}(A^{(1)})Xe_s).$$

Define $z_\theta(X;s) = \mathcal{S}(A^{(1)})Xe_s$. We have that

$$\begin{aligned}
-\Delta_\beta(\theta) &= -\frac{1}{S-1}\operatorname{Tr}\left[G^{(2)}(A^{(1)},A^{(2)})\right]\\
&= \frac{1}{S(S-1)}\sum_{s,s'}\mathbb{E}_{\pi,X}\left[\frac{\pi(s'\mid s)}{f_\theta(X;s)_{s'}+\epsilon}\delta_{s'}(X)^\top J(v_\theta(X;s))\mathcal{S}(A^{(1)})Xe_s\right]\\
&= \frac{1}{S(S-1)}\sum_{s,s'}\mathbb{E}_{\pi,X}\left[\frac{\pi(s'\mid s)}{f_\theta(X;s)_{s'}+\epsilon}\delta_{s'}(X)^\top J(\mathcal{S}(\beta z_\theta(X;s)))z_\theta(X;s)\right]\\
&= \frac{1}{S(S-1)}\sum_{s,s'}\mathbb{E}_{\pi,X}\left[\frac{\pi(s'\mid s)}{\delta_{s'}(X)^\top\mathcal{S}(\beta z_\theta(X;s))+\epsilon}\delta_{s'}(X)^\top J(\mathcal{S}(\beta z_\theta(X;s)))z_\theta(X;s)\right]
\end{aligned}$$

We first show the upper bound. We can write

$$\begin{aligned}
\delta_{s'}(X)^\top J(\mathcal{S}(\beta z_\theta(X;s)))z_\theta(X;s) &\le \sum_i \delta_{s'}(X)_i\mathcal{S}(\beta z_\theta(X;s))_i z_\theta(X;s)_i\\
&\le \sum_i \delta_{s'}(X)_i\mathcal{S}(\beta z_\theta(X;s))_i\\
&= \delta_{s'}(X)^\top\mathcal{S}(\beta z_\theta(X;s)),
\end{aligned}$$

since $0 \le z_\theta(X;s)_i \le 1$. Therefore

$$-\Delta_\beta(\theta) \le \frac{1}{S(S-1)}\sum_{s,s'}\mathbb{E}_{\pi,X}\left[\pi(s'\mid s)\right] = \frac{1}{S-1} \le 1.$$

We next move to the lower bound. Define $\tilde z(X;s) := \mathcal{S}(A_*^{(1)})Xe_s$. We have that

$$\tilde z(X;s)_i = \begin{cases} x_{p(i),s} & \text{if } i \notin \mathcal{R}\\ z_\theta(X;s)_i & \text{if } i \in \mathcal{R}\end{cases}. \tag{28}$$

First, we will aim to replace $z_\theta(X;s)$ with $\tilde z(X;s)$. Indeed, when $p(i) \ne \emptyset$,

$$|\tilde z(X;s)_i - z_\theta(X;s)_i| = \left|\left(\mathcal{S}(A^{(1)})_i - \mathcal{S}(A_*^{(1)})_i\right)^\top\delta_s(X)\right| \le \left\|\mathcal{S}(A^{(1)})_i - \mathcal{S}(A_*^{(1)})_i\right\|_1 \lesssim T^{-1},$$

by Corollary D.6. Thus $\|\tilde z(X;s) - z_\theta(X;s)\|_\infty \lesssim T^{-1}$.

Define

$$q_{s'}(z) := \frac{\delta_{s'}(X)^\top J(\mathcal{S}(\beta z))z}{\delta_{s'}(X)^\top\mathcal{S}(\beta z)+\epsilon},$$

so that

$$-\Delta_\beta(\theta) = \frac{1}{S(S-1)}\sum_{s,s'}\mathbb{E}_{\pi,X}[\pi(s'\mid s)q_{s'}(z_\theta(X;s))].$$

By Lemma H.1, we have that

$$|q_{s'}(z_\theta(X;s)) - q_{s'}(\tilde{z}(X;s))| \lesssim (1+\beta)\|z_\theta(X;s) - \tilde{z}(X;s)\|_\infty \lesssim (1+\beta)T^{-1},$$

and thus

$$\left| -\Delta_\beta(\theta) - \frac{1}{S(S-1)}\sum_{s,s'} \mathbb{E}_{\pi,X}[\pi(s' \mid s) q_{s'}(\tilde{z}(X;s))] \right| \lesssim (1+\beta)T^{-1}.$$

Next, plugging in the definition of $q_{s'}$, we get

$$\begin{aligned}
&\frac{1}{S(S-1)}\sum_{s,s'} \mathbb{E}_{\pi,X}[\pi(s' \mid s) q_{s'}(\tilde{z}(X;s))] \\
&= \frac{1}{S(S-1)}\sum_{s,s'} \mathbb{E}_{\pi,X}\left[\pi(s' \mid s)\frac{\delta_{s'}(X)^\top\left(\operatorname{diag}(\mathcal{S}(\beta\tilde{z}(X;s))) - \mathcal{S}(\beta\tilde{z}(X;s))\mathcal{S}(\beta\tilde{z}(X;s))^\top\right)\tilde{z}(X;s)}{\delta_{s'}(X)^\top \mathcal{S}(\beta\tilde{z}(X;s)) + \epsilon}\right] \\
&\geq \frac{1}{S(S-1)}\sum_{s,s'} \mathbb{E}_{\pi,X}\left[\pi(s' \mid s)\cdot\left(\frac{\sum_i x_{i,s'}\mathcal{S}(\beta\tilde{z}(X;s))_i \tilde{z}(X;s)_i}{\epsilon + \sum_i x_{i,s'}\mathcal{S}(\beta\tilde{z}(X;s))_i} - \sum_i \mathcal{S}(\beta\tilde{z}(X;s))_i\tilde{z}(X;s)_i\right)\right]. \qquad (29)
\end{aligned}$$

Our next goal is to replace the term in the parentheses in (29) with something independent of X, where the concentration holds as T_{eff} grows large. Indeed, define the quantities $E^{(1)}_{s,s'}(X), E^{(2)}_{s,s'}(X), E^{(3)}_{s}(X)$ by

$$E^{(1)}_{s,s'}(X) := \sum_i x_{i,s'}\mathcal{S}(\beta\tilde{z}(X;s))_i\tilde{z}(X;s)_i \qquad (30)$$

$$E^{(2)}_{s,s'}(X) := \sum_i x_{i,s'}\mathcal{S}(\beta\tilde{z}(X;s))_i \qquad (31)$$

$$E^{(3)}_{s}(X) := \sum_i \mathcal{S}(\beta\tilde{z}(X;s))_i\tilde{z}(X;s)_i, \qquad (32)$$

so that

$$\frac{1}{S(S-1)}\sum_{s,s'} \mathbb{E}_{\pi,X}[\pi(s' \mid s) q_{s'}(\tilde{z}(X;s))] \geq \frac{1}{S(S-1)}\sum_{s,s'} \mathbb{E}_{\pi,X}\left[\pi(s' \mid s)\cdot\left(\frac{E^{(1)}_{s,s'}(X)}{\epsilon + E^{(2)}_{s,s'}(X)} - E^{(3)}_s(X)\right)\right].$$

Let $r = \frac{|\mathcal{R}|}{T}$. One can make the approximation

$$\frac{E^{(1)}_{s,s'}(X)}{\epsilon + E^{(2)}_{s,s'}(X)} \approx \frac{(1-r)e^\beta\mu_\pi(s)\pi(s' \mid s) + re^{\beta\mu_\pi(s)}\mu_\pi(s)\mu_\pi(s')}{(1-r)(e^\beta - 1)\mu_\pi(s)\pi(s' \mid s) + (1-r)\mu_\pi(s') + re^{\beta\mu_\pi(s)}\mu_\pi(s')} \qquad (33)$$

$$E^{(3)}_s(X) \approx \frac{(1-r)e^\beta\mu_\pi(s) + re^{\beta\mu_\pi(s)}\mu_\pi(s)}{(1-r)(e^\beta - 1)\mu_\pi(s) + (1-r) + re^{\beta\mu_\pi(s)}}. \qquad (34)$$

This motivates defining the following idealized gradient:

$$\begin{aligned}
\hat{g}(\beta) := \frac{1}{S(S-1)}\sum_s \mathbb{E}_\pi\Big[\mu_\pi(s)\cdot\Big(&\sum_{s'}\frac{(1-r)e^\beta\pi(s' \mid s)^2 + re^{\beta\mu_\pi(s)}\mu_\pi(s')\pi(s' \mid s)}{(1-r)(e^\beta - 1)\mu_\pi(s)\pi(s' \mid s) + (1-r)\mu_\pi(s') + re^{\beta\mu_\pi(s)}\mu_\pi(s')} \\
&- \frac{(1-r)e^\beta + re^{\beta\mu_\pi(s)}}{(1-r)(e^\beta - 1)\mu_\pi(s) + (1-r) + re^{\beta\mu_\pi(s)}}\Big)\Big]
\end{aligned}$$

Indeed, the approximations in (33) and (34) can be made rigorous: by Lemma H.6 and Lemma H.7, we have that

$$\begin{aligned}
\left|\frac{1}{S(S-1)}\sum_{s,s'} \mathbb{E}_{\pi,X}\left[\pi(s' \mid s)\cdot\left(\frac{E^{(1)}_{s,s'}(X)}{\epsilon + E^{(2)}_{s,s'}(X)} - E^{(3)}_s(X)\right)\right] - \hat{g}(\beta)\right| &\lesssim (1+\beta)\cdot\frac{\log^{1/2} T}{T_{\text{eff}}^{1/4}} \\
&\lesssim e^\beta\frac{\log^{1/2} T}{T_{\text{eff}}^{1/4}}.
\end{aligned}$$

Finally, it suffices to show that $\hat{g}(\beta) \geq 0$. Define the function $h_s : \mathbb{R} \to \mathbb{R}$ by

$$h_s(z) = \frac{(1-r)e^{\beta}z^2 + re^{\beta\mu_\pi(s)}z}{(1-r)(e^{\beta}-1)\mu_\pi(s)z + (1-r) + re^{\beta\mu_\pi(s)}} - \frac{(1-r)e^{\beta} + re^{\beta\mu_\pi(s)}}{(1-r)(e^{\beta}-1)\mu_\pi(s) + (1-r) + re^{\beta\mu_\pi(s)}}.$$

Simplifying the formula for $\hat{g}(\beta)$, we see that it can be written in terms of this h_s:

$$\hat{g}(\beta) = \frac{1}{S(S-1)}\sum_s \mathbb{E}_\pi\left[\mu_\pi(s)\cdot\left(\sum_{s'}\mu_\pi(s')h_s\left(\frac{\pi(s' \mid s)}{\mu_\pi(s')}\right)\right)\right].$$

Furthermore, h_s is convex, and so $\hat{g}(\beta)$ is actually a linear combination of h_s-divergences and is hence nonnegative. The following lemma relates the h_s-divergence to the χ^2-divergence in order to get a quantitative lower bound on $\hat{g}(\beta)$ away from 0. The proof is deferred to Appendix G.1.

Lemma D.9. $\hat{g}(\beta) \geq \frac{1}{2}\gamma^8 S^{-6}e^{-2\beta} > 0.$

To conclude, when $\beta \leq \beta^*$,

$$\left|-\Delta_\beta(\theta) - \frac{1}{S(S-1)}\sum_{s,s'}\mathbb{E}[\pi(s' \mid s)q(\tilde{z}(X;s))]\right| \lesssim e^{\beta}T^{-1} \leq \frac{1}{8}\gamma^8 S^{-6}e^{-2\beta}$$

$$\left|\frac{1}{S(S-1)}\sum_{s,s'}\mathbb{E}_{\pi,X}\left[\pi(s' \mid s)\cdot\left(\frac{E^{(1)}_{s,s'}(X)}{\epsilon + E^{(2)}_{s,s'}(X)} - E^{(3)}_s(X)\right)\right] - \hat{g}(\beta)\right| \lesssim e^{\beta}\frac{\log^{1/2}T}{T_{\text{eff}}^{1/4}} \leq \frac{1}{8}\gamma^8 S^{-6}e^{-2\beta},$$

and thus

$$-\Delta_\beta(\theta) \geq \frac{1}{4}\gamma^8 S^{-6}e^{-2\beta},$$

as desired. □

Lemma D.10 (Dynamics of $A^{(2)}$)**.** *Let $A^{(1)}(\tau_1)$ be the output of stage 1 of Algorithm 1, and let $\eta_2 \leq 1$. There exists $\tau_2 \lesssim_{\gamma,S} e^{2\beta^*}\beta^*\eta_2^{-1}$ such that*

$$1 + \beta^* \geq \beta(\tau_1 + \tau_2) \geq \beta^*.$$

Proof. If $\beta(t) \leq \beta^*$, then by Lemma D.8

$$\beta(t+1) \geq \beta(t) + \eta_2 \cdot \frac{1}{4}\gamma^8 S^{-6}e^{-2\beta(t)} \geq \beta(t) + \eta_2 \cdot \frac{1}{4}\gamma^8 S^{-6}e^{-2\beta^*}.$$

Assume that $\beta(\tau_1 + t) < \beta^*$ for all $t \leq \mathcal{T} := 4S^6\gamma^{-8}e^{2\beta^*}\beta^*\eta_2^{-1}$. Then

$$\beta(\tau_1 + \mathcal{T}) \geq \frac{1}{4}\gamma^8 S^{-6}e^{-2\beta^*}\mathcal{T}\eta_2 = \beta^*,$$

a contradiction. Therefore $\beta(\tau_1 + \tau_2) \geq \beta^*$ for some $\tau_2 \leq \mathcal{T} \lesssim e^{2\beta^*}\beta^*\eta_2^{-1}$. Finally, by Lemma D.8, $\beta(t+1) \leq \beta(t) + 1$, and thus letting τ_2 be the smallest such time we have $1 + \beta^* \geq \beta(\tau_1 + \tau_2) \geq \beta^*$. □

D.7. Proof of Theorem 4.4

Proof of Theorem 4.4. Pick $\beta_0 \leq c_{\gamma,S}\frac{1}{T_{\text{eff}}^{3/2}}$, and set $\tau_1 = C_{\gamma,S}\eta_1^{-1}\beta_0^{-1}T^2\log(T)$ for constants $c_{\gamma,S}, C_{\gamma,S}$ chosen appropriately. By Corollary D.6, the output of stage 1 satisfies

$$1 - \mathcal{S}\Big(A^{(1)}(\tau_1)\Big)_{i,p(i)} \lesssim T^{-1}.$$

for $i \in \overline{\mathcal{R}}$.

Next, by Lemma D.10 there exists $\tau_2 = \tilde{O}\left(T_{\text{eff}}^{1/6}\eta_2^{-1}\right)$ such that $\beta(\tau_1 + \tau_2) \ge \beta^*$.

It now suffices to bound the loss of the predictor $\hat{\theta}$. We have

$$\begin{aligned}\left|L(\hat{\theta}) - L^*\right| &\le \mathbb{E}_{\pi,X}\left[\frac{1}{S}\sum_{s,s'} \pi(s' \mid s) \cdot \left|\log\left(f_{\hat{\theta}}(X;s)_{s'} + \epsilon\right) - \log \pi(s' \mid s)\right|\right] \\ &= \mathbb{E}_{\pi}\left[\frac{1}{S}\sum_{s,s'} \pi(s' \mid s)\mathbb{E}_X\left[\left|\log\left(f_{\hat{\theta}}(X;s)_{s'} + \epsilon\right) - \log \pi(s' \mid s)\right|\right]\right]\end{aligned}$$

For $A, B > 0$, one has the bound

$$|\log A - \log B| \le \frac{|A - B|}{\min(A,B)}.$$

Therefore

$$\begin{aligned}&\left|\log\left(f_{\hat{\theta}}(X;s)_{s'} + \epsilon\right) - \log \pi(s' \mid s)\right| \\ &\le \left(\left|f_{\hat{\theta}}(X;s)_{s'} - \pi(s' \mid s)\right| + \epsilon\right) \cdot \frac{1}{\min(f_{\hat{\theta}}(X;s)_{s'} + \epsilon, \pi(s' \mid s))} \\ &\lesssim \left(\left|f_{\hat{\theta}}(X;s)_{s'} - \pi(s' \mid s)\right| + \epsilon\right)\left(\mathbf{1}_{f_{\hat{\theta}}(X;s)_{s'} \ge \frac{\gamma^3}{4S}} + \epsilon^{-1}\mathbf{1}_{f_{\hat{\theta}}(X;s)_{s'} < \frac{\gamma^3}{4S}}\right),\end{aligned}$$

and thus by Cauchy

$$\begin{aligned}&\mathbb{E}_X\left|\log\left(f_{\hat{\theta}}(X;s)_{s'} + \epsilon\right) - \log \pi(s' \mid s)\right| \\ &\lesssim \left(\left(\mathbb{E}_X\left|f_{\hat{\theta}}(X;s)_{s'} - \pi(s' \mid s)\right|^2\right)^{1/2} + \epsilon\right)\left(1 + \epsilon^{-1}\mathbb{P}_X\left(f_{\hat{\theta}}(X;s)_{s'} < \frac{\gamma^3}{4S}\right)\right) \\ &\lesssim \left(\left(\mathbb{E}_X\left|f_{\hat{\theta}}(X;s)_{s'} - \pi(s' \mid s)\right|^2\right)^{1/2} + \epsilon\right)\left(1 + \epsilon^{-1}T_{\text{eff}}^{-1}\right) \\ &\lesssim \left(\mathbb{E}_X\left|f_{\hat{\theta}}(X;s)_{s'} - \pi(s' \mid s)\right|^2\right)^{1/2} + \epsilon,\end{aligned}$$

where the bound $\mathbb{P}_X\left(f_{\hat{\theta}}(X;s)_{s'} < \frac{\gamma^3}{4S}\right) \lesssim T_{\text{eff}}^{-1}$ follows from Lemma H.8. Altogether, applying Lemma H.8 again, we get

$$\begin{aligned}\left|L(\hat{\theta}) - L^*\right| &\lesssim \left(\mathbb{E}_X\left|f_{\hat{\theta}}(X;s)_{s'} - \pi(s' \mid s)\right|^2\right)^{1/2} + \epsilon \\ &\le \frac{e^{\beta^*}\log^{1/2} T}{T_{\text{eff}}^{1/4}} + e^{-\beta^*\gamma/2} + \epsilon \\ &\lesssim \frac{\log^{1/3} T}{T_{\text{eff}}^{1/6}} + \left(\frac{\log^{1/6} T}{T_{\text{eff}}^{1/12}}\right)^{\gamma/2} \\ &\lesssim \left(\frac{\log^2 T}{T_{\text{eff}}}\right)^{\gamma/24}.\end{aligned}$$

□

E. Markov Chain Preliminaries

Given a Markov chain π with stationary measure μ_π, we define the normalized and centered transition matrix $B_\pi \in \mathbb{R}^{S \times S}$ by:

$$(B_\pi)_{s,s'} := \sqrt{\frac{\mu_\pi(s)}{\mu_\pi(s')}}[\pi(s'|s) - \mu_\pi(s')].$$

An immediate consequence is that

$$(B_\pi^k)_{s,s'} := \sqrt{\frac{\mu_\pi(s)}{\mu_\pi(s')}}[\pi^k(s'|s) - \mu_\pi(s')]$$

which allows for the decomposition

$$\pi^k(s'|s) = \mu_\pi(s') + (B_\pi^k)_{s,s'}\sqrt{\frac{\mu_\pi(s')}{\mu_\pi(s)}}.$$

We also observe that

$$\|B_\pi\|_F^2 = \sum_{s,s'}\left(\frac{\mu_\pi(s)\pi(s' \mid s)^2}{\mu_\pi(s')} - \mu_\pi(s')\mu_\pi(s)\right) = \sum_{s,s'}\frac{\mu_\pi(s)\pi(s' \mid s)^2}{\mu_\pi(s')} - 1. \tag{35}$$

Definition E.1 (Spectral Gap)**.** We say that a Markov chain π with stationary measure μ_π has a spectral gap of $1 - \lambda(\pi)$ where $\lambda(\pi) := \|B_\pi\|_2$.

Lemma E.2. *Let* $\min_{s,s'} \pi(s \mid s') \geq \gamma/S$. *Then* $\lambda(\pi) \leq 1 - \gamma/S$.

Proof. By Lemma E.3, we can write

$$\pi = \frac{\gamma}{S}1\mu_\pi^\top + (1-\gamma)Q$$

for another stochastic matrix Q. One then sees that $\pi^\top Q = \pi$. Therefore

$$\pi - 1\mu_\pi^\top = (1 - \gamma/S)(\pi - 1\mu_\pi^\top),$$

so

$$\|\pi - 1\mu_\pi^\top\|_{\mu_\pi} = (1-\gamma/S)\|Q - 1\mu_\pi^\top\|_{\mu_\pi} \leq 1 - \gamma/S.$$

Therefore $\lambda(\pi) \leq 1 - \gamma/S$. □

Lemma E.3. *Let* $\min_{s,s'} \pi(s \mid s') \geq \gamma/S$. *Then* $\min_s \mu_\pi(s) \geq \gamma/S$.

Proof. Since $\mu_\pi(s)$ is stationary,

$$\begin{aligned}\mu_\pi(s') &= \sum_s \pi(s' \mid s)\mu_\pi(s) \\ &\geq \sum_s \gamma/S \cdot \mu_\pi(s) \\ &= \gamma/S,\end{aligned}$$

as desired. □

Lemma E.4. *Let* $\min_{s,s'} \pi(s \mid s') \geq \gamma/S$. *Then*

$$\min_{j\neq k} \mathrm{TV}(\pi(\cdot \mid j), \pi(\cdot \mid k)) \leq 1 - \gamma.$$

Proof. Write

$$\begin{aligned}\mathrm{TV}(\pi(\cdot \mid j), \pi(\cdot \mid k)) &= \frac{1}{2}\sum_s |\pi(s \mid j) - \pi(s \mid k)| \\ &= \frac{1}{2}\sum_s (\pi(s \mid j) + \pi(s \mid k) - 2\min\{\pi(s \mid j), \pi(s \mid k)\}) \\ &\leq 1 - \gamma.\end{aligned}$$

□

Lemma E.5. $\|B_\pi\|_F^2 \geq \gamma^2/S$

Proof. By definition

$$\|B_\pi\|_F^2 = \sum_{s,s'} \frac{\pi(s' \mid s)^2 \mu_\pi(s)}{\mu_\pi(s')} - 1,$$

and thus

$$\begin{aligned}
\|B_\pi\|_F^2 &= \sum_{s'} \frac{1}{\mu_\pi(s')} \left(\sum_s \pi(s' \mid s)^2 \mu_\pi(s) - \mu_\pi(s')^2 \right) \\
&\geq \sum_{s,s'} \mu_\pi(s) (\pi(s' \mid s) - \mu_\pi(s'))^2 \\
&\geq \frac{\gamma^2}{S} \sum_s \mu_\pi(s) \\
&= \frac{\gamma^2}{S}.
\end{aligned}$$

□

Lemma E.6 ((Cohen et al., 1993), Theorem 3.1)**.** *Let π be a stochastic matrix such that $\max_s \pi(s' \mid s) > 0$ for all s'. Then, for any f-divergence D_f and probability vectors x, y,*

$$D_f(\pi \circ x || \pi \circ y) \leq \alpha(\pi) D_f(x||y),$$

where the contraction coefficient $\alpha(\pi)$ is defined as

$$\alpha(\pi) := \max_{j \neq k} \mathrm{TV}(\pi(\cdot \mid j), \pi(\cdot \mid k)) = \frac{1}{2} \max_{j \neq k} \|\pi(\cdot \mid j) - \pi(\cdot \mid k)\|_1.$$

F. Concentration

Definition F.1 (Graph Distance)**.** Let $\mathcal{G}$ be the directed acyclic graph in Section 2.2. Let $\overline{G}$ denote the undirected version of $\mathcal{G}$. Then we define $d(i,j)$ to be length of the shortest path between i, j in $\mathcal{G}$. If i, j are not connected in G then $d(i,j) := \infty$.

Definition F.2 (Effective Sequence Length)**.** For $\lambda \in (0,1)$, we define the effective sequence length $T_{\text{eff}}(\lambda)$ by:

$$T_{\text{eff}}(\lambda) := \frac{T^2}{\sum_{i,j=1}^T \lambda^{d(i,j)}}.$$

This formula for $T_{\text{eff}}(\lambda)$ is closely related to the definition of T_{eff} (Definition 4.3):

Lemma F.3. *Decompose $\mathcal{G} = \bigcup_{i=1}^k \mathcal{T}_i$ where $\mathcal{T}_i$ are disjoint trees. Let L_i denote the number of leaves of tree $\mathcal{T}_i$ for $i = 1, \ldots, k$. Then,*

$$T_{\mathit{eff}}(\lambda) \geq \frac{T(1-\lambda)}{\max_{i=1}^k L_i} =: (1-\lambda) T_{\mathit{eff}}$$

Proof. Note that $T_{\text{eff}}(\lambda)^{-1}$ naturally decomposes to a sum within each tree as $d(i,j) := \infty$ when i and j are not connected:

$$\begin{aligned}
\frac{1}{T_{\text{eff}}(\lambda)} &= \frac{1}{T^2} \sum_{l=1}^k \sum_{i,j \in \mathcal{T}_l} \lambda^{d(i,j)} \\
&= \frac{1}{T^2} \sum_{l=1}^k \sum_{i,j \in \mathcal{T}_l} \lambda^{d(i,j)} \\
&= \frac{1}{T^2} \sum_{l=1}^k \sum_{i \in \mathcal{T}_l} \sum_{k \geq 0} \#\{j \in \mathcal{T}_l \ : \ d(i,j) = k\} \lambda^k.
\end{aligned}$$

Now note that for a fixed node i, each path from i to j with $d(i,j) = k$ can be lengthened to a path that reaches a leaf. Furthermore, for each leaf there can be only one such j. Therefore, $\#\{j \in \mathcal{T}_l \ : \ d(i,j) = k\} \le L_l$. Plugging this in gives:

$$\begin{aligned}\frac{1}{T_{\text{eff}}(\lambda)} &\le \frac{1}{T^2} \sum_{l=1}^{k} |\mathcal{T}_l| L_l \sum_{k \ge 0} \lambda^k \\ &= \frac{\sum_{l=1}^{k} |\mathcal{T}_l| L_l}{T^2(1-\lambda)} \\ &\le \frac{\max_l T_l}{T(1-\lambda)}\end{aligned}$$

which completes the proof. □

Throughout the remainder of this section, the only assumption we place on π is that $\min \pi(s' \mid s) \ge \gamma/S$. Defining $\lambda := 1 - \gamma/S$, we have that $\lambda \ge \lambda(\pi)$ by Lemma E.2, and thus $T_{\text{eff}}(\lambda)^{-1} \lesssim T_{\text{eff}}^{-1}$.

Lemma F.4. *For any π and any $i, j < T$,*

$$|\mathbb{P}_X[x_j = s, x_i = s'] - \mu_\pi(s)\mu_\pi(s')| \le \sqrt{\mu_\pi(s)\mu_\pi(s')}\lambda(\pi)^{d(i,j)}.$$

Proof. Let k be the closest common parent of i, j so that $d(k,i) + d(k,j) = d(i,j)$ and there exist directed paths from k to i and k to j in $\mathcal{G}$. Then,

$$\begin{aligned}&\mathbb{P}[s_j = s, s_i = s'] - \mu_\pi(s)\mu_\pi(s') \\ &= \operatorname{Cov}[x_{j,s} x_{i,s'}] \\ &= \mathbb{E}[(x_{j,s} - \mu_\pi(s))(x_{i,s'} - \mu_\pi(s'))] \\ &= \sum_{s_k \in [S]} \mu_\pi(s_k)(\pi^{d(k,j)}(s'|s_k) - \mu_\pi(s'))(\pi^{d(k,i)}(s|s_k) - \mu_\pi(s)) \\ &= \sum_{s_k \in [S]} \mu_\pi(s_k)\left((B_\pi^{d(k,j)})_{s_k,s}\sqrt{\frac{\mu_\pi(s)}{\mu_\pi(s_k)}}\right)\left((B_\pi^{d(k,i)})_{s_k,s'}\sqrt{\frac{\mu_\pi(s')}{\mu_\pi(s_k)}}\right) \\ &= \sqrt{\mu_\pi(s)\mu_\pi(s')} \sum_{s_k \in [S]} (B_\pi^{d(k,j)})_{s_k,s}(B_\pi^{d(k,i)})_{s_k,s'} \\ &= \sqrt{\mu_\pi(s)\mu_\pi(s')}[(B_\pi^{d(k,j)})^\top (B_\pi^{d(k,i)})]_{s,s'}.\end{aligned}$$

Therefore taking absolute values gives:

$$\begin{aligned}|\mathbb{P}[s_j = s, s_i = s'] - \mu_\pi(s)\mu_\pi(s')| &\le \sqrt{\mu_\pi(s)\mu_\pi(s')}\|B_\pi\|^{d(k,j)+d(k,i)} \\ &\le \sqrt{\mu_\pi(s)\mu_\pi(s')}\lambda(\pi)^{d(i,j)}.\end{aligned}$$

□

Lemma F.5. *For any subset $I \subset [T-1]$, define*

$$\hat{\mu}_{X_I}(s) := \frac{1}{|I|}\sum_{i \in I} x_{i,s}.$$

Then,

$$\mathbb{E}_X[\hat{\mu}_{X_I}(s)] = \mu_\pi(s) \quad \textit{and} \quad \mathbb{E}_X[(\hat{\mu}_{X_I}(s) - \mu_\pi(s))^2] \le \frac{\mu_\pi(s) T^2}{T_{\textit{eff}}(\lambda)|I|^2}.$$

Note that Lemma F.5 is excluding the token x_T as it is resampled from $\operatorname{Unif}([S])$.

Proof. The first claim follows from the fact that $\mathbb{E}[x_{i,s}] = \mu_\pi(s)$ as the sequence X is initialized from μ_π. Then,

$$\begin{aligned}
\mathbb{E}_X[(\hat{\mu}_{X_I}(s) - \mu_\pi(s))^2] &= \frac{1}{|I|^2} \sum_{i,j \in I} \mathbb{E}_X[x_{i,s}x_{j,s} - \mu_\pi(s)^2] \\
&\le \frac{\mu_\pi(s)}{|I|^2} \sum_{i,j \in I} \lambda^{d(i,j)} \\
&\le \frac{\mu_\pi(s)}{|I|^2} \sum_{i,j=1}^{T-1} \lambda^{d(i,j)} \\
&= \frac{\mu_\pi(s)(T-1)^2}{T_{\text{eff}}(\lambda)|I|^2}
\end{aligned}$$

which completes the proof. □

Corollary F.6.

$$\mathbb{E}_X[(\hat{\mu}_X(s) - \mu_\pi(s))^2] \lesssim \frac{1}{T_{\textit{eff}}(\lambda)}.$$

Proof. One can write

$$\hat{\mu}_X(s) = \frac{T-1}{T}\hat{\mu}_{X_{[T-1]}}(s) + \frac{1}{T}x_{T,s}.$$

Thus

$$\mathbb{E}_X[(\hat{\mu}_X(s) - \mu_\pi(s))^2] \le \left(\frac{T-1}{T}\right)^2 \mathbb{E}_X\left[\left(\hat{\mu}_{X_{[T-1]}}(s) - \mu_\pi(s)\right)^2\right] + \frac{1}{T^2} \lesssim \frac{1}{T_{\text{eff}}(\lambda)}.$$

□

Lemma F.7. *For any subset $I \subset [T-1]$ such that $p(i) \neq \emptyset$ for all $i \in I$, define*

$$\hat{c}_{X_I}(s, s') := \frac{1}{|I|}\sum_{i \in I} x_{p(i),s}x_{i,s}.$$

Then,

$$\mathbb{E}_X[\hat{c}_{X_I}(s,s')] = \mu_\pi(s)\pi(s'|s) \quad \textit{and} \quad \mathbb{E}_X[(\hat{\mu}_{X_I}(s) - \mu_\pi(s)\pi(s'|s))^2] \lesssim \frac{T^2}{T_{\textit{eff}}(\lambda)|I|^2}.$$

Proof. The first result follows from linearity of expectation and the fact that the Markov process is stationary. Then,

$$\begin{aligned}
&\mathbb{E}_X[(\hat{\mu}_{X_I}(s) - \mu_\pi(s)\pi(s'|s))^2] \\
&= \frac{1}{|I|^2}\sum_{i,j \in I} \mathbb{E}[x_{p(i),s}x_{i,s'}x_{p(j),s}x_{j,s'}] - \mu_\pi(s)^2\pi(s'|s)^2.
\end{aligned}$$

There are three possibilities for the dependency graph of i, j. First, if $i = j$ the expression in the sum is equal to $\mu_\pi(s)\pi(s'|s)(1 - \mu_\pi(s)\pi(s'|s))$. Next, if i, j are independent conditioned on $p(i), p(j)$, we get

$$\begin{aligned}
&\mathbb{E}[x_{p(i),s}x_{i,s'}x_{p(j),s}x_{j,s'}] - \mu_\pi(s)^2\pi(s'|s)^2 \\
&= \pi(s'|s)^2(\mathbb{E}[x_{p(i),s}x_{p(j),s}] - \mu_\pi(s)^2) \\
&\le \mu_\pi(s)\pi(s'|s)^2\lambda^{d(p(i),p(j))}.
\end{aligned}$$

Finally, if i, j are dependent conditioned on $p(i), p(j)$ it means that either there is a directed path from i to $p(j)$ or a direct path from j to $p(i)$ in the directed graph $\mathcal{G}$. Without loss of generality, we can assume that there is a directed path from j to $p(i)$. Then we have:

$$\begin{aligned}
&\mu_\pi(s)\pi(s'|s)\pi^{d(j,p(i))}(s|s')\pi(s'|s) - \mu_\pi(s)^2\pi(s'|s)^2 \\
&= \mu_\pi(s)\pi(s'|s)^2\left[\pi^{d(j,p(i))}(s|s') - \mu_\pi(s)\right] \\
&\leq \sqrt{\mu_\pi(s)\mu_\pi(s')}\pi(s'|s)^2\lambda^{d(j,p(i))}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_X[(\hat{\mu}_{X_I}(s) - \mu_\pi(s)\pi(s'|s))^2] \\
&\lesssim \frac{1}{|I|^2}\sum_{i,j\in I}\lambda^{d(i,j)} \\
&\lesssim \frac{T^2}{T_{\text{eff}}(\lambda)|I|^2}.
\end{aligned}$$

□

Lemma F.8. *For any subset $I \subset [T-1]$ such that $p(i) \neq \emptyset$ for all $i \in I$,*

$$\mathbb{E}_X\left[\left(\frac{1}{|I|}\sum_{i\in I}x_{p(i),s} - \mu_\pi(s)\right)^2\right] \lesssim \frac{T^2}{T_{\mathit{eff}}(\lambda)|I|^2}.$$

Proof. As above, we will directly compute the second moment:

$$\begin{aligned}
\frac{1}{|I|^2}\sum_{i,j\in I}x_{p(i),s}x_{p(j),s} - \mu_\pi(s)^2 &\leq \frac{\mu_\pi(s)}{|I|^2}\sum_{i,j\in I}\lambda^{d(p(i),p(j))} \\
&\leq \frac{\mu_\pi(s)}{|I|^2}\sum_{i,j\in I}\lambda^{d(i,j)-2} \\
&\leq \frac{\mu_\pi(s)}{\lambda^2|I|^2}\sum_{i,j\in T}\lambda^{d(i,j)} \\
&\leq \frac{T^2\mu_\pi(s)}{T_{\text{eff}}(\lambda)\lambda^2|I|^2}.
\end{aligned}$$

□

G. Lemmas for Stage 1

G.1. Strong Data Processing Inequality

We briefly recall the definition of the χ^2 divergence between two probability distributions on state space $\mathcal{X}$:

$$\chi^2(P||Q) := \sum_{x\in\mathcal{X}}\frac{P(x)^2}{Q(x)} - 1,$$

along with the χ^2 mutual information between two random variables Y, Z

$$I_{\chi^2}(Y;Z) = \sum_{y,z\in\mathcal{X}}\frac{P(Y=y,Z=z)^2}{P(Y=y)P(Z=z)} - 1$$

Proof of Lemma D.2. First we consider the case where i and j are in separate trees. If $i \neq T$, then $\mathbb{P}_X[s_i = s', s_j = s] = \mu_\pi(s)\mu_\pi(s')$, and thus

$$g_{i,j}(\pi) = \sum_{s,s'} \pi(s' \mid s)\mu_\pi(s) - 1 = 0.$$

We note that this subsumes the case where i is a root note, since that necessarily implies that j is in a different tree. Otherwise when $i = T$,

$$g_{i,j}(\pi) = \frac{1}{S}\sum_{s,s'} \frac{\pi(s' \mid s)\mu_\pi(s)}{\mu_\pi(s')} - 1 = \frac{1}{S}\sum_{s'} \frac{\mu_\pi(s')}{\mu_\pi(s')} - 1 = 0.$$

Next, assume that i and j are in the same tree. When $j = p(i)$, we have

$$\begin{aligned} g_{i,p(i)}(\pi) &= \sum_{s,s'} \frac{\pi(s' \mid s)}{\mu_\pi(s')} \cdot \mathbb{P}_X[s_i = s', s_j = s] - 1 \\ &= \sum_{s,s'} \frac{\pi(s' \mid s)^2 \mu_\pi(s)}{\mu_\pi(s')} - 1 \\ &= \|B_\pi\|_F^2, \end{aligned}$$

where the last equality is (35).

If $j \neq p(i)$ and $j \neq i$, then by AM-GM:

$$\begin{aligned} g_{i,j}(\pi) &= \sum_{s,s'} \frac{\pi(s' \mid s)}{\mu_\pi(s')} \cdot \mathbb{P}_X[s_i = s', s_j = s] - 1 \\ &\le \frac{1}{2}\sum_{s,s'} \frac{\mu_\pi(s)\pi(s' \mid s)^2}{\mu_\pi(s')} + \frac{1}{2}\sum_{s,s'} \frac{\mathbb{P}_X[s_i = s', s_j = s]^2}{\mu_\pi(s)\mu_\pi(s')} - 1 \\ &= \frac{1}{2}\|B_\pi\|_F^2 + \frac{1}{2} I_{\chi^2}(s_i; s_j \mid \pi). \end{aligned}$$

We see that the χ^2-mutual information can be rewritten as

$$I_{\chi^2}(s_i; s_j \mid \pi) = \sum_{s'} \mu_\pi(s') \cdot \chi^2(\mathbb{P}_X[s_j = \cdot \mid s_i = s'] \,||\, \mu_\pi).$$

Let $p(i,j)$ be the least common ancestor of i and j. Let x be the probability distribution defined by $x = \mathbb{P}_X\left[s_{p(i,j)} = \cdot \mid s_i = s'\right]$. The distribution $\pi^{d(j,p(i,j))} \circ x$ is

$$\begin{aligned} (\pi^{d(j,p(i,j))} \circ x)(s) &= \sum_{s^*} \pi^{d(j,p(i,j))}(s \mid s^*) \cdot x(s^*) \\ &= \sum_{s^*} \mathbb{P}_X[s_j = s \mid s_{p(i,j)} = s^*] \cdot \mathbb{P}_X\left[s_{p(i,j)} = s^* \mid s_i = s'\right] \\ &= \mathbb{P}_X[s_j = s \mid s_i = s'], \end{aligned}$$

where the last line uses the fact that s_i and s_j are conditionally independent given $p(i,j)$.

Applying Lemma E.6, we thus have

$$\chi^2(\mathbb{P}_X[s_j = \cdot \mid s_i = s'] \,||\, \mu_\pi) \le \alpha(\pi)^{d(j,p(i,j))} \cdot \chi^2\left(\mathbb{P}_X\left[s_{p(i,j)} = \cdot \mid s_i = s'\right] \,||\, \mu_\pi\right).$$

Therefore

$$\begin{aligned}
I_{\chi^2}(s_i; s_j \mid \pi) &\le \alpha(\pi)^{d(j,p(i,j))} \sum_{s'} \mu_\pi(s') \cdot \chi^2\big(\mathbb{P}_X\big[s_{p(i,j)} = \cdot \mid s_i = s'\big] \,||\, \mu\big) \\
&= \alpha(\pi)^{d(j,p(i,j))} \cdot I_{\chi^2}(s_{p(i,j)}; s_i \mid \pi) \\
&= \alpha(\pi)^{d(j,p(i,j))} \sum_{s} \mu_\pi(s) \cdot \chi^2\big(\mathbb{P}_X\big[s_i = \cdot \mid s_{p(i,j)} = s\big] \,||\, \mu\big) \\
&= \alpha(\pi)^{d(j,p(i,j))} \sum_{s} \mu_\pi(s) \cdot \chi^2\Big(\pi^{d(i,p(i,j))}(\cdot \mid s) || \, \mu\Big).
\end{aligned}$$

Since $i > j$, $d(i, p(i,j)) \ge 1$, and thus we can apply Lemma E.6 to get

$$\chi^2\Big(\pi^{d(i,p(i,j))}(\cdot \mid s) || \, \mu\Big) \le \alpha(\pi)^{d(i,p(i,j))-1} \cdot \chi^2(\pi(\cdot \mid s) || \, \mu).$$

Altogether,

$$\begin{aligned}
I_{\chi^2}(s_i; s_j \mid \pi) &\le \alpha(\pi)^{d(j,p(i,j))+d(i,p(i,j))-1} \sum_{s} \mu_\pi(s) \cdot \chi^2(\pi(\cdot \mid s) || \, \mu) \\
&= \alpha(\pi)^{d(i,j)-1} \cdot \left(\sum_{s,s'} \frac{\pi(s' \mid s)^2 \mu_\pi(s)}{\mu_\pi(s')} - 1 \right) \\
&= \alpha(\pi)^{d(i,j)-1} \|B_\pi\|_F^2.
\end{aligned}$$

For $j \ne p(i), d(i,j) \ge 2$, so

$$\begin{aligned}
g_{i,j}(\pi) &\le \frac{1}{2}\Big(\alpha(\pi)^{d(i,j)-1} + 1\Big) \|B_\pi\|_F^2 \\
&\le \frac{1}{2}(\alpha(\pi) + 1) \|B_\pi\|_F^2.
\end{aligned}$$

and thus

$$g_{i,p(i)}(\pi) - g_{i,j}(\pi) \ge \frac{1 - \alpha(\pi)}{2} \cdot \|B_\pi\|_F^2.$$

By Assumption 4.1 and Lemma E.4, we have $1 - \alpha(\pi) \ge \gamma$ and $\|B_\pi\|_F^2 \ge \gamma^2/S$. Therefore

$$g_{i,p(i)}(\pi) - g_{i,j}(\pi) \ge \frac{\gamma^3}{2S}.$$

Finally, when $j = i$, we have

$$g_{i,i}(\pi) = \sum_s \pi(s \mid s) - 1.$$

Therefore

$$g_{i,i} = \sum_s \mathbb{E}[\pi(s \mid s)] - 1 = 0.$$

Therefore $g_{i,p(i)} - g_{i,i} \ge \gamma^2/S \ge \frac{\gamma^3}{2S}$. □

G.2. Auxiliary Dynamics Lemmas

Lemma G.1. *Let $\theta = (A^{(1)}, \beta_0 I_S), \hat{\theta} = (A^{(1)}, 0)$, for $\beta_0 \le 1$. Define $g_i^*, \hat{g}_i \in \mathbb{R}^i$ by*

$$\begin{aligned}
g_i^* &:= T \sum_{s,s'} \mathbb{E}\left[\frac{\pi(s' \mid s)}{f_\theta(X; s)_{s'} + \epsilon} \delta_{s'}(X)^\top J(v_\theta(X; s)) e_i \cdot \delta_s(X_{\le i}) \right], \\
\hat{g}_i &:= T \sum_{s,s'} \mathbb{E}\left[\frac{\pi(s' \mid s)}{f_{\hat{\theta}}(X; s)_{s'} + \epsilon} \delta_{s'}(X)^\top J(v_{\hat{\theta}}(X; s)) e_i \cdot \delta_s(X_{\le i}) \right].
\end{aligned}$$

Then $\|g_i^* - \hat{g}_i\|_\infty \le 3S^2\epsilon^{-2}(e^{\beta_0} - 1)$

Proof. We can bound

$$\begin{aligned}
&\left|\frac{1}{f_\theta(X;s)_{s'}+\epsilon}\delta_{s'}(X)^\top J(v_\theta(X;s))e_i - \frac{1}{f_{\hat\theta}(X;s)_{s'}+\epsilon}\delta_{s'}(X)^\top J(v_{\hat\theta}(X;s))e_i\right| \\
&\quad\le \left|\frac{1}{f_\theta(X;s)_{s'}+\epsilon} - \frac{1}{f_{\hat\theta}(X;s)_{s'}+\epsilon}\right|\left|\delta_{s'}(X)^\top J(v_{\hat\theta}(X;s))e_i\right| \\
&\quad+ \frac{1}{f_{\hat\theta}(X;s)_{s'}+\epsilon}\left|\delta_{s'}(X)^\top\big(J(v_\theta(X;s)) - J(v_{\hat\theta}(X;s))\big)e_i\right|.
\end{aligned}$$

First, see that

$$\begin{aligned}
\left|f_\theta(X;s)_{s'} - f_{\hat\theta}(X;s)_{s'}\right| &= \left|\delta_{s'}(X)^\top\big(v_\theta(X;s) - v_{\hat\theta}(X;s)\big)\right| \\
&\le \left\|v_\theta(X;s) - v_{\hat\theta}(X;s)\right\|_1,
\end{aligned}$$

since $\|\delta_{s'}(X)\|_\infty \le 1$. Next, we have

$$v_\theta(X;s) = \mathcal{S}\Big(\beta_0\cdot\mathcal{S}(A^{(1)})X^\top I_S e_s\Big)) = \mathcal{S}\Big(\beta_0\cdot\mathcal{S}(A^{(1)})\delta_s(X)\Big)).$$

Since $\mathcal{S}(A^{(1)})\delta_s(X)$ has entries in $[0,1]$, we can bound each entry of $v_\theta(X;s)$ as

$$\frac{1}{(T-1)e^{\beta_0}+1} \le v_\theta(X;s)_i \le \frac{e^{\beta_0}}{e^{\beta_0}+(T-1)},$$

and thus

$$\left|v_\theta(X;s)_i - v_{\hat\theta}(X;s)_i\right| = \left|v_\theta(X;s)_i - \frac{1}{T}\right| \le \frac{e^{\beta_0}}{e^{\beta_0}+(T-1)} - \frac{1}{T} \le \frac{e^{\beta_0}-1}{T}.$$

Thus

$$\left|f_\theta(X;s)_{s'} - f_{\hat\theta}(X;s)_{s'}\right| \le e^{\beta_0} - 1. \tag{36}$$

Next, see that

$$\delta_{s'}(X)^\top J(v_\theta(X;s))e_i = v_\theta(X;s)_i[x_{i,s'} - f_\theta(X;s)_{s'}],$$

and thus

$$\begin{aligned}
&\left|\delta_{s'}(X)^\top\big(J(v_\theta(X;s)) - J(v_{\hat\theta}(X;s))\big)e_i\right| \\
&\quad\le \left|v_\theta(X;s)_i - v_{\hat\theta}(X;s)_i\right|\left|x_{i,s'} - f_\theta(X;s)_{s'}\right| + v_{\hat\theta}(X;s)_i\left|f_\theta(X;s)_{s'} - f_{\hat\theta}(X;s)_{s'}\right| \\
&\quad\le \frac{2(e^{\beta_0}-1)}{T}.
\end{aligned}$$

Altogether, we have the bound

$$\left|\frac{1}{f_\theta(X;s)_{s'}+\epsilon}\delta_{s'}(X)^\top J(v_\theta(X;s))e_i - \frac{1}{f_{\hat\theta}(X;s)_{s'}+\epsilon}\delta_{s'}(X)^\top J(v_{\hat\theta}(X;s))e_i\right| \le \frac{3(e^{\beta_0}-1)}{\epsilon^2 T}.$$

Therefore

$$\begin{aligned}
&\|g_i^* - \hat g_i\| \\
&\le T\sum_{s,s'}\mathbb{E}_{\pi,X}\left[\pi(s'\mid s)\left|\frac{1}{f_\theta(X;s)_{s'}+\epsilon}\delta_{s'}(X)^\top J(v_\theta(X;s))e_i - \frac{1}{f_{\hat\theta}(X;s)_{s'}+\epsilon}\delta_{s'}(X)^\top J(v_{\hat\theta}(X;s))e_i\right|\right] \\
&\le T\sum_{s,s'}E_{\pi,X}\left[\pi(s'\mid s)\cdot\frac{3(e^{\beta_0}-1)}{\epsilon^2 T}\right] \\
&\le 3S\epsilon^{-2}(e^{\beta_0}-1) \\
&\le 6S\epsilon^{-2}\beta_0,
\end{aligned}$$

since $e^z - 1 \le 2z$ for $z\in[0,1]$. □

G.3. Concentration

Lemma G.2. *For any $s, s' \in \mathcal{S}$ and any π with spectral gap $1 - \lambda(\pi) \geq 1 - \lambda$ (see Definition E.1) and $\mu_\pi(s') \geq \gamma/S$, there exists a sufficiently large constant $C_{\gamma,S}$ such that if $\epsilon \geq C_{\gamma,S} T_{\mathit{eff}}^{-1/2}$ and $i \geq j$,*

$$\left| \mathbb{E}_X \left[\frac{(x_{i,s'} - \hat{\mu}_X(s'))x_{j,s}}{\hat{\mu}_X(s') + \epsilon} \right] - \left(\frac{\mathbb{P}_X[s_i = s', s_j = s]}{\mu_\pi(s')} - \mathbb{P}_X[s_j = s] \right) \right| \lesssim \frac{1}{\sqrt{T_{\mathit{eff}}}}.$$

Proof.

$$\begin{aligned} E_\pi(s, s') &:= \mathbb{E}_X \left[\frac{(x_{i,s'} - \hat{\mu}_X(s'))x_{j,s}}{\hat{\mu}_X(s') + \epsilon} \right] - \frac{\mathbb{P}_X[s_i = s', s_j = s]}{\mu_\pi(s')} + \mathbb{P}_X[s_j = s] \\ &= \mathbb{E}_X \left[\frac{x_{i,s'} x_{j,s}}{\hat{\mu}_X(s') + \epsilon} \right] - \frac{\mathbb{P}_X[s_i = s', s_j = s]}{\mu_\pi(s')} - \mathbb{E}_X \left[\frac{\hat{\mu}_X(s')}{\hat{\mu}_X(s') + \epsilon} x_{j,s} \right] + \mathbb{P}_X[s_j = s]. \end{aligned}$$

$E_\pi(s, s')$ can be rewritten as:

$$\begin{aligned} E_\pi(s, s') &= \mathbb{E}_X \left[\frac{x_{i,s'} x_{j,s}}{\hat{\mu}_X(s') + \epsilon} - \frac{x_{i,s'} x_{j,s}}{\mu_\pi(s')} - \frac{\hat{\mu}_X(s')}{\hat{\mu}_X(s') + \epsilon} x_{j,s} + x_{j,s} \right] \\ &= \mathbb{E}_X \left[\frac{x_{i,s'} x_{j,s}}{\hat{\mu}_X(s') + \epsilon} - \frac{x_{i,s'} x_{j,s}}{\mu_\pi(s')} + \frac{\epsilon x_{j,s}}{\hat{\mu}_X(s') + \epsilon} \right] \\ &= \mathbb{E}_X \left[\frac{x_{i,s'} x_{j,s} [\mu_\pi(s') - \hat{\mu}_X(s') - \epsilon] + \epsilon x_{j,s} \mu_\pi(s')}{(\hat{\mu}_X(s') + \epsilon)\mu_\pi(s')} \right] \end{aligned}$$

Note that the inside of the expectation is upper bounded by $O(\epsilon^{-1})$. Therefore by the triangle inequality we have

$$\begin{aligned} |E_\pi(s, s')| &\leq \mathbb{E}_X \left[\frac{x_{i,s'} x_{j,s} |\hat{\mu}_X(s') - \mu_\pi(s')| + \epsilon [x_{i,s'} x_{j,s} + \mu_\pi(s') x_{j,s}]}{(\hat{\mu}_X(s') + \epsilon)\mu_\pi(s')} \right] \\ &= \mathbb{E}_X \left[\frac{x_{i,s'} x_{j,s} |\hat{\mu}_X(s') - \mu_\pi(s')| + \epsilon [x_{i,s'} x_{j,s} + \mu_\pi(s') x_{j,s}]}{(\hat{\mu}_X(s') + \epsilon)\mu_\pi(s')} \mathbf{1}_{\hat{\mu}_X(s') > \frac{\mu_\pi(s')}{2}} \right] \\ &\quad + \mathbb{E}_X \left[\frac{x_{i,s'} x_{j,s} |\hat{\mu}_X(s') - \mu_\pi(s')| + \epsilon [x_{i,s'} x_{j,s} + \mu_\pi(s') x_{j,s}]}{(\hat{\mu}_X(s') + \epsilon)\mu_\pi(s')} \mathbf{1}_{\hat{\mu}_X(s') \leq \frac{\mu_\pi(s')}{2}} \right] \\ &\lesssim \mathbb{E}_X \left[\frac{x_{i,s'} x_{j,s} |\hat{\mu}_X(s') - \mu_\pi(s')| + \epsilon [x_{i,s'} x_{j,s} + \mu_\pi(s') x_{j,s}]}{\mu_\pi(s')^2} \right] \\ &\quad + \epsilon^{-1} \mathbb{P}_X \left[\hat{\mu}_X(s') \leq \frac{\mu_\pi(s')}{2} \right] \\ &\lesssim \sqrt{\mathbb{E}[(\hat{\mu}_X(s') - \mu_\pi(s'))^2]} + \epsilon + \frac{1}{\epsilon T_{\text{eff}}} \\ &\lesssim \frac{1}{\sqrt{T_{\text{eff}}}} + \epsilon, \end{aligned}$$

where the last inequality follows from Corollary F.6. □

H. Lemmas for Stage 2

H.1. Idealized Gradient

Proof of Lemma D.9. Recall

$$h_s(z) = \frac{(1-r)e^\beta z^2 + r e^{\beta \mu_\pi(s)} z}{(1-r)(e^\beta - 1)\mu_\pi(s) z + (1-r) + r e^{\beta \mu_\pi(s)}} - \frac{(1-r)e^\beta + r e^{\beta \mu_\pi(s)}}{(1-r)(e^\beta - 1)\mu_\pi(s) + (1-r) + r e^{\beta \mu_\pi(s)}}.$$

and

$$\hat{g}(\beta) = \frac{1}{S(S-1)} \sum_s \mathbb{E}_\pi \left[\mu_\pi(s) \cdot \left(\sum_{s'} \mu_\pi(s') h_s\left(\frac{\pi(s' \mid s)}{\mu_\pi(s')} \right) \right) \right].$$

For a function $h(z) = \frac{Az^2+Bz}{Cz+D}$, one has

$$h''(z) = \frac{2D(AD - BC)}{(Cz + D)^3}.$$

Thus for $z \in [0, S\gamma^{-1}]$,

$$\begin{aligned}
h_s''(z) &= \frac{2\big(1 - r + re^{\beta\mu_\pi(s)}\big) \cdot (1 - r) \cdot \big(e^\beta(1 - r) + re^{\beta\mu_\pi(s)+\beta} - r(e^\beta - 1)\mu_\pi(s)e^{\beta\mu_\pi(s)}\big)}{\big((1 - r)(e^\beta - 1)\mu_\pi(s)z + (1 - r) + re^{\beta\mu_\pi(s)}\big)^3} \\
&\geq \frac{2(1 - r)^2 e^\beta}{\big((1 - r)(e^\beta - 1)\mu_\pi(s)S\gamma^{-1} + (1 - r) + re^{\beta\mu_\pi(s)}\big)^3} \\
&\geq \frac{2(1 - r)^2 e^\beta}{\big(S\gamma^{-1}e^\beta\big)^3} \\
&\geq 2\gamma^5 S^{-3} e^{-2\beta}.
\end{aligned}$$

Therefore for $z \in [0, S\gamma^{-1}]$,

$$h_s(z) \geq h_s'(1)(z - 1) + \gamma^5 S^{-3} e^{-2\beta} \cdot (z - 1)^2.$$

Note that $\frac{\pi(s'|s)}{\mu_\pi(s')} \leq \frac{S}{\gamma}$. Therefore

$$h_s\left(\frac{\pi(s' \mid s)}{\mu_\pi(s')}\right) \geq h_s'(1)\left(\frac{\pi(s' \mid s)}{\mu_\pi(s')} - 1\right) + \gamma^5 S^{-3} e^{-2\beta} \cdot \left(\frac{\pi(s' \mid s)}{\mu_\pi(s')} - 1\right)^2$$

and thus

$$\begin{aligned}
\left(\sum_{s'} \mu_\pi(s') h_s\left(\frac{\pi(s' \mid s)}{\mu_\pi(s')}\right)\right) &\geq \gamma^5 S^{-3} e^{-2\beta} \sum_{s'} \mu_\pi(s')\left(\frac{\pi(s' \mid s)}{\mu_\pi(s')} - 1\right)^2 \\
&= \gamma^5 S^{-3} e^{-2\beta} \chi^2(\pi(\cdot \mid s) || \mu_\pi).
\end{aligned}$$

Altogether,

$$\begin{aligned}
\hat{g}(\beta) &= \frac{1}{S(S-1)} \sum_s \mathbb{E}_\pi\left[\mu_\pi(s) \cdot \left(\sum_{s'} \mu_\pi(s') h_s\left(\frac{\pi(s' \mid s)}{\mu_\pi(s')}\right)\right)\right] \\
&\geq \gamma^5 S^{-5} e^{-2\beta} \mathbb{E}_\pi\left[\sum_s \mu_\pi(s) \chi^2((\pi(\cdot \mid s) || \mu)\right] \\
&= \gamma^5 S^{-5} e^{-2\beta} \mathbb{E}_\pi\left[\|B_\pi\|_F^2\right] \\
&= \frac{1}{2}\gamma^8 S^{-6} e^{-2\beta}.
\end{aligned}$$

□

H.2. Auxiliary Dynamics Lemmas

Lemma H.1. *Define*

$$q_{s'}(z) = \frac{\delta_{s'}(X)^\top J(\mathcal{S}(\beta z))z}{\delta_{s'}(X)^\top \mathcal{S}(\beta z) + \epsilon},$$

Then $\sup_{z \in [0,1]^T} \left\|\nabla q_{s'}(z)\right\|_1 \leq 10(1 + \beta)$.

Proof. We have that

$$\nabla_z q_{s'}(z) = \frac{J(\beta z)\delta_{s'}(X) + \beta \nabla J(\mathcal{S}(\beta z))(\delta_{s'}(X), z)}{\delta_{s'}(X)^\top \mathcal{S}(\beta z) + \epsilon} - \frac{\delta_{s'}(X)^\top J(\beta z) z \cdot \beta J(\beta z)\delta_{s'}(X)}{(\delta_{s'}(X)^\top \mathcal{S}(\beta z) + \epsilon)^2}.$$

First, by Lemma H.2,

$$\|J(\beta z)\delta_{s'}(X)\|_1 \le 2\delta_{s'}(X)^\top \mathcal{S}(\beta z).$$

Next, by Lemma H.3,

$$\left\|\nabla J(\mathcal{S}(\beta z))(\delta_{s'}(X), z)\right\|_1 \le 2\mathcal{S}(\beta z)^\top(\delta_{s'}(X) \odot z) + 4\mathcal{S}(\beta z)^\top \delta_{s'}(X)\mathcal{S}(\beta z)^\top z \le 6\mathcal{S}(\beta z)^\top \delta_{s'}(X),$$

where the last inequality uses the fact that z has entries in $[0, 1]$. Finally,

$$\left|\delta_{s'}(X)^\top J(\beta z) z \cdot J(\beta z)\delta_{s'}(X)\right| \le \|J(\beta z)\delta_{s'}(X)\|_1 \cdot \|J(\beta z)\delta_{s'}(X)\|_1 \cdot \|z\|_\infty \le 4(\delta_{s'}(X)^\top \mathcal{S}(\beta z))^2.$$

Altogether,

$$\|\nabla_z q_{s'}(z)\|_1 \le \frac{(2 + 6\beta)\delta_{s'}(X)^\top \mathcal{S}(\beta z)}{\delta_{s'}(X)^\top \mathcal{S}(\beta z) + \epsilon} + \frac{4\beta(\delta_{s'}(X)^\top \mathcal{S}(\beta z))^2}{(\delta_{s'}(X)^\top \mathcal{S}(\beta z) + \epsilon)^2} \le 2 + 10\beta.$$

□

Lemma H.2. *Let u be a vector with nonnegative entries. Then $\|J(\mathcal{S}(v))u\|_1 \le 2\mathcal{S}(v)^\top u$*

Proof.

$$\|J(\mathcal{S}(v))u\|_1 = \sum_i \left|\mathcal{S}(v)_i(u_i - \mathcal{S}(v)^\top u)\right| \le \sum_i \mathcal{S}(v)_i u_i + \mathcal{S}(v)^\top u \cdot \sum_i \mathcal{S}(v)_i = 2\mathcal{S}(v)^\top u.$$

□

Lemma H.3. *Recall that $J(s) = diag(s) - ss^\top$. Then $\nabla_v J(\mathcal{S}(v)) \in \mathbb{R}^{d\times d\times d}$ satisfies*

$$\|\nabla J(\mathcal{S}(v))(u, w)\|_1 \le 2\mathcal{S}(v)^\top (u \odot w) + 4\mathcal{S}(v)^\top u\mathcal{S}(v)^\top w.$$

for nonnegative vectors u, w.

Proof. See that

$$J(\mathcal{S}(v))(u, w) = u^\top \mathrm{diag}(\mathcal{S}(v))w - \mathcal{S}(v)^\top u\mathcal{S}(v)^\top w = \mathcal{S}(v)^\top(u \odot w) - \mathcal{S}(v)^\top u\mathcal{S}(v)^\top w.$$

Taking the gradient, and noting that $\nabla_v \mathcal{S}(v) = J(v)$, we get

$$\nabla J(\mathcal{S}(v))(u, w) = J(\mathcal{S}(v))(u \odot w) - \mathcal{S}(v)^\top w \cdot J(\mathcal{S}(v))u - \mathcal{S}(v)^\top u \cdot J(\mathcal{S}(v))w.$$

Since $u \odot w$ is also a nonnegative vector, we get that

$$\|\nabla J(\mathcal{S}(v))(u, w)\|_1 \le 2\mathcal{S}(v)^\top (u \odot w) + 4\mathcal{S}(v)^\top u\mathcal{S}(v)^\top w.$$

□

H.3. Concentration

Lemma H.4. *For any nonzero scalars A_1, A_2, B_1, B_2,*

$$\left|\frac{A_1}{B_1} - \frac{A_2}{B_2}\right| \le \frac{1}{|B_2|}\left(\left|\frac{A_1}{B_1}\right| \cdot |B_1 - B_2| + |A_1 - A_2|\right).$$

Proof.

$$\begin{aligned}\left|\frac{A_1}{B_1} - \frac{A_2}{B_2}\right| &\le \left|\frac{A_1}{B_1} - \frac{A_1}{B_2}\right| + \left|\frac{A_1}{B_2} - \frac{A_2}{B_2}\right| \\ &= |A_1|\left|\frac{1}{B_1} - \frac{1}{B_2}\right| + \frac{1}{|B_2|}|A_1 - A_2| \\ &= \frac{|A_1||B_1 - B_2|}{|B_1 B_2|} + \frac{|A_1 - A_2|}{|B_2|} \\ &= \frac{1}{|B_2|}\left(\left|\frac{A_1}{B_1}\right| \cdot |B_1 - B_2| + |A_1 - A_2|\right).\end{aligned}$$

□

For the following lemmas, let $\hat{\theta} = (\hat{A}^{(1)}, \hat{A}^{(2)})$ be the output of Algorithm 1. Define

$$\mathcal{S}\left(A_*^{(1)}\right)_{ij} = \begin{cases} \mathbf{1}(j = p(i)) & p(i) \neq \emptyset \\ \hat{A}_{i,j}^{(1)} & p(i) = \emptyset \end{cases}.$$

and let $\tilde{z}(X; s) := \mathcal{S}(A_*^{(1)}) X e_s$.

Lemma H.5. *For $i \in \mathcal{R}$,*

$$\mathbb{E}_X\left[\left|\tilde{z}(X; s)_i - \hat{\mu}_{X_{\le i}}(s)\right|^2\right] \lesssim \min\left(1, \frac{T^2 \log^2 T}{T_{\mathit{eff}} \cdot i^2}\right).$$

Proof. By Corollary D.6

$$\left|\mathcal{S}(A_*^{(1)})_{i,j} - \frac{1}{i}\right| \lesssim \frac{T \log T}{T_{\mathrm{eff}}^{1/2} i^2}.$$

Therefore

$$\begin{aligned}\left|\tilde{z}(X; s)_i - \hat{\mu}_{X_{\le i}}(s)\right| &= \left|\left(\mathcal{S}(A_*^{(1)})_i - \frac{1}{i}\mathbf{1}_i\right) \cdot \delta_s(X_{\le i})\right| \\ &\le \left\|\mathcal{S}(A_*^{(1)})_i - \frac{1}{i}\mathbf{1}_i\right\|_1 \\ &\lesssim \frac{T \log T}{T_{\mathrm{eff}}^{1/2} i}.\end{aligned}$$

Finally,

$$\mathbb{E}_X\left[\left|\hat{\mu}_{X_{\le i}}(s) - \mu_\pi(s)\right|^2\right] \lesssim \frac{\mu_\pi(s) T^2}{T_{\mathrm{eff}}(\lambda) i^2}.$$

Altogether,

$$\mathbb{E}_X\left[\left|\tilde{z}(X; s)_i - \hat{\mu}_{X_{\le i}}(s)\right|^2\right] \lesssim \frac{T^2 \log^2 T}{T_{\mathrm{eff}} \cdot i^2},$$

and the conclusion follows as $\tilde{z}(X; s)_i, \hat{\mu}_{X_{\le i}}(s) \in [0, 1]$. □

Lemma H.6. *Define*

$$E_s^{(3)}(X) := \sum_i \mathcal{S}(\beta \tilde{z}(X;s))_i \tilde{z}(X;s)_i$$

Then

$$\mathbb{E}_X\left[\left|E_s^{(3)}(X) - \frac{(1-r)e^\beta \mu_\pi(s) + re^{\beta\mu_\pi(s)}\mu_\pi(s)}{(1-r)(e^\beta-1)\mu_\pi(s) + (1-r) + re^{\beta\mu_\pi(s)}}\right|^2\right] \lesssim (1+\beta^2)\cdot\frac{\log T}{T_{\mathit{eff}}^{1/2}}.$$

Proof. Plugging in the formula for $\tilde{z}(X;s)$ (28), we get that

$$\begin{aligned}
E_s^{(3)}(X) &= \frac{\sum_i \exp\left(\beta\tilde{z}(X;s)_i\right)\tilde{z}(X;s)_i}{\sum_i \exp\left(\beta\tilde{z}(X;s)_i\right)} \\
&= \frac{e^\beta \sum_{i\in\overline{\mathcal{R}}} x_{p(i),s} + \sum_{i\in\mathcal{R}} e^{\beta\tilde{z}(X;s)_i}\tilde{z}(X;s)_i}{(e^\beta-1)\sum_{i\in\overline{\mathcal{R}}} x_{p(i),s} + \left|\overline{\mathcal{R}}\right| + \sum_{i\in\mathcal{R}} e^{\beta\tilde{z}(X;s)_i}}
\end{aligned}$$

We define the error terms

$$\begin{aligned}
\mathcal{E}_1(X) &:= \frac{1}{T}\sum_{i\in\overline{\mathcal{R}}} x_{p(i),s} - (1-r)\mu_\pi(s) \\
\mathcal{E}_2(X) &:= \frac{1}{T}\sum_{i\in\mathcal{R}} e^{\beta\tilde{z}(X;s)_i}\tilde{z}(X;s)_i - re^{\beta\mu_\pi(s)}\mu_\pi(s) \\
\mathcal{E}_3(X) &:= \frac{1}{T}\sum_{i\in\mathcal{R}} e^{\beta\tilde{z}(X;s)_i} - re^{\beta\mu_\pi(s)}.
\end{aligned}$$

Then

$$E_s^{(3)}(X) = \frac{(1-r)e^\beta\mu_\pi(s) + re^{\beta\mu_\pi(s)}\mu_\pi(s) + e^\beta\mathcal{E}_1(X) + \mathcal{E}_2(X)}{(1-r)(e^\beta-1)\mu_\pi(s) + (1-r) + re^{\beta\mu_\pi(s)} + (e^\beta-1)\mathcal{E}_1(X) + \mathcal{E}_3(X)}$$

Thus applying Lemma H.4, we get that

$$\begin{aligned}
&\left|E_s^{(3)}(X) - \frac{(1-r)e^\beta\mu_\pi(s) + re^{\beta\mu_\pi(s)}\mu_\pi(s)}{(1-r)(e^\beta-1)\mu_\pi(s) + (1-r) + re^{\beta\mu_\pi(s)}}\right| \\
&\quad\le \left|E_s^{(3)}(X)\right|\cdot\frac{\left(\left|e^\beta\mathcal{E}_1(X)\right| + \mathcal{E}_3(X)\right) + \left(\left|e^\beta\mathcal{E}_1(X)\right| + \mathcal{E}_2(X)\right)}{(1-r)(e^\beta-1)\mu_\pi(s) + (1-r) + re^{\beta\mu_\pi(s)}} \\
&\quad\lesssim (1-r)^{-1}\cdot\left(|\mathcal{E}_1(X)| + e^{-\beta}|\mathcal{E}_2(X)| + e^{-\beta}|\mathcal{E}_3(X)|\right),
\end{aligned}$$

since $\left|E_s^{(3)}(X)\right| \le 1$ and $(1-r)(e^\beta-1)\mu_\pi(s) + (1-r) + re^{\beta\mu_\pi(s)} \ge (1-r)e^\beta\gamma$.

First, by Lemma F.8, we have

$$\mathbb{E}\left[\mathcal{E}_1(X)^2\right] \lesssim \frac{1}{T_{\text{eff}}}.$$

Next, we bound $\mathcal{E}_2$:

$$|\mathcal{E}_2(X)| \le \frac{1}{T}\sum_{i\in\mathcal{R}}\left|e^{\beta\tilde{z}(X;s)}\tilde{z}(X;s) - e^{\beta\mu_\pi(s)}\mu_\pi(s)\right| \le \frac{1}{T}\sum_{i\in\mathcal{R}}(1+\beta)e^\beta|\tilde{z}(X;s)_i - \mu_\pi(s)|,$$

and thus by Lemma H.5

$$\begin{aligned}\mathbb{E}_X\left[\mathcal{E}_2(X)^2\right] &\le \frac{(1+\beta)^2 e^{2\beta}}{T}\sum_{i\in\mathcal{R}}\mathbb{E}|\tilde{z}(X;s)_i - \mu_\pi(s)|^2 \\ &\lesssim \frac{(1+\beta)^2 e^{2\beta}}{T}\sum_i \min\left(1, \frac{T^2\log^2 T}{T_{\text{eff}}\cdot i^2}\right) \\ &= \frac{(1+\beta)^2 e^{2\beta}}{T}\left(\frac{T\log T}{T_{\text{eff}}^{1/2}} + \sum_{i > \frac{T\log T}{T_{\text{eff}}^{1/2}}} \frac{T^2\log^2 T}{T_{\text{eff}}\cdot i^2}\right) \\ &\lesssim \frac{(1+\beta)^2 e^{2\beta}\log T}{T_{\text{eff}}^{1/2}}.\end{aligned}$$

Next, we bound $\mathcal{E}_3$.

$$|\mathcal{E}_3(X)| \le \frac{1}{T}\sum_{i\in\mathcal{R}}\left|e^{\beta\tilde{z}(X;s)_i} - e^{\beta\mu_\pi(s)}\right| \le \frac{1}{T}\sum_{i\in\mathcal{R}}\beta e^{\beta}|\tilde{z}(X;s)_i - \mu_\pi(s)|,$$

so by an identical calculation to as for $\mathcal{E}_2$,

$$\mathbb{E}_X\left[\mathcal{E}_3(X)^2\right] \lesssim \frac{\beta^2 e^{2\beta}\log T}{T_{\text{eff}}^{1/2}}.$$

Altogether,

$$\begin{aligned}&\mathbb{E}\left[\left|E_s^{(3)}(X) - \frac{(1-r)e^{\beta}\mu_\pi(s) + re^{\beta\mu_\pi(s)}\mu_\pi(s)}{(1-r)(e^\beta - 1)\mu_\pi(s) + (1-r) + re^{\beta\mu_\pi(s)}}\right|^2\right] \\ &\lesssim (1-r)^{-2}\left(\mathbb{E}\left[\mathcal{E}_1(X)^2\right] + e^{-2\beta}\mathbb{E}\left[\mathcal{E}_2(X)^2\right] + e^{-2\beta}\mathbb{E}\left[\mathcal{E}_3(X)^2\right]\right) \\ &\lesssim (1+\beta^2)\cdot\frac{\log T}{T_{\text{eff}}^{1/2}},\end{aligned}$$

where the last inequality also relies on Assumption 4.2. □

Lemma H.7. *Define*

$$\begin{aligned}E_{s,s'}^{(1)}(X) &:= \sum_i x_{i,s'}\mathcal{S}(\beta\tilde{z}(X;s))_i\tilde{z}(X;s)_i \\ E_{s,s'}^{(2)}(X) &:= \sum_i x_{i,s'}\mathcal{S}(\beta\tilde{z}(X;s))_i,\end{aligned}$$

Then

$$\begin{aligned}&\mathbb{E}\left[\left|\frac{E_{s,s'}^{(1)}(X)}{E_{s,s'}^{(2)}(X)+\epsilon} - \frac{(1-r)e^{\beta}\mu_\pi(s)\pi(s'\mid s) + re^{\beta\mu_\pi(s)}\mu_\pi(s')\mu_\pi(s)}{(1-r)(e^\beta - 1)\mu_\pi(s)\pi(s'\mid s) + (1-r)\mu_\pi(s') + re^{\beta\mu_\pi(s)}\mu_\pi(s')}\right|^2\right] \\ &\lesssim (1+\beta^2)\cdot\frac{\log T}{T_{\mathit{eff}}^{1/2}}.\end{aligned}$$

Proof. Plugging in the formula for $\tilde{z}(X;s)$ (28), we have that

$$\frac{E_{s,s'}^{(1)}(X)}{E_{s,s'}^{(2)}(X)+\epsilon} = \frac{e^{\beta}\sum_{i\in\overline{\mathcal{R}}} x_{i,s'}x_{p(i),s} + \sum_{i\in\mathcal{R}} e^{\beta\tilde{z}(X;s)_i}\tilde{z}(X;s)_i x_{i,s'}}{(e^\beta - 1)\sum_{i\in\overline{\mathcal{R}}} x_{i,s'}x_{p(i),s} + \sum_{i\in\overline{\mathcal{R}}} x_{i,s'} + \sum_{i\in\mathcal{R}} e^{\beta\tilde{z}(X;s)_i}x_{i,s'} + \epsilon\sum_i e^{\beta\tilde{z}(X;s)_i}}$$

We define the error terms

$$
\begin{aligned}
\mathcal{E}_4(X) &:= \frac{1}{T}\sum_{i \in \overline{\mathcal{R}}} x_{p(i),s} x_{i,s'} - (1-r)\mu_\pi(s)\pi(s' \mid s) \\
\mathcal{E}_5(X) &:= \frac{1}{T}\sum_{i \in \overline{\mathcal{R}}} x_{i,s'} - (1-r)\mu_\pi(s') \\
\mathcal{E}_6(X) &:= \frac{1}{T}\sum_{i \in \mathcal{R}} \left(e^{\beta \tilde{z}(X;s)_i}\tilde{z}(X;s)_i x_{i,s'} - e^{\beta\mu_\pi(s)}\mu_\pi(s)\mu_\pi(s')\right) \\
\mathcal{E}_7(X) &= \frac{1}{T}\sum_{i \in \mathcal{R}} \left(e^{\beta \tilde{z}(X;s)_i} x_{i,s'} - e^{\beta\mu_\pi(s)}\mu_\pi(s')\right).
\end{aligned}
$$

Then

$$
\frac{E^{(1)}_{s,s'}(X)}{E^{(2)}_{s,s'}(X) + \epsilon} =
$$

$$
\frac{(1-r)e^{\beta}\mu_\pi(s)\pi(s' \mid s) + re^{\beta\mu_\pi(s)}\mu_\pi(s)\mu_\pi(s') + e^{\beta}\mathcal{E}_4(X) + \mathcal{E}_6(X)}{(1-r)[(e^{\beta}-1)\mu_\pi(s)\pi(s' \mid s) + \mu_\pi(s')] + re^{\beta\mu_\pi(s)}\mu_\pi(s') + (e^{\beta}-1)\mathcal{E}_4(X) + \mathcal{E}_5(X) + \mathcal{E}_7(X) + \frac{\epsilon}{T}\sum_i e^{\beta\tilde{z}(X;s)_i}}.
$$

Therefore by Lemma H.4,

$$
\begin{aligned}
&\left|\frac{E^{(1)}_{s,s'}(X)}{E^{(2)}_{s,s'}(X) + \epsilon} - \frac{(1-r)e^{\beta}\mu_\pi(s)\pi(s' \mid s) + re^{\beta\mu_\pi(s)}\mu_\pi(s')\mu_\pi(s)}{(1-r)(e^{\beta}-1)\mu_\pi(s)\pi(s' \mid s) + (1-r)\mu_\pi(s') + re^{\beta\mu_\pi(s)}\mu_\pi(s')}\right| \\
&\le \left|\frac{E^{(1)}_{s,s'}(X)}{E^{(2)}_{s,s'}(X) + \epsilon}\right| \cdot \frac{\left(e^{\beta}|\mathcal{E}_4(X)| + |\mathcal{E}_5(X)| + |\mathcal{E}_7(X)| + e^{\beta}\epsilon\right) + e^{\beta}|\mathcal{E}_4(x)| + |\mathcal{E}_6(X)|}{(1-r)(e^{\beta}-1)\mu_\pi(s)\pi(s' \mid s) + (1-r)\mu_\pi(s') + re^{\beta\mu_\pi(s)}\mu_\pi(s')} \\
&\lesssim |\mathcal{E}_4(X)| + e^{-\beta}|\mathcal{E}_5(X)| + e^{-\beta}|\mathcal{E}_6(X)| + e^{-\beta}|\mathcal{E}_7(X)| + \epsilon,
\end{aligned}
$$

where the last step uses $(1-r)(e^{\beta}-1)\mu_\pi(s)\pi(s' \mid s) + (1-r)\mu_\pi(s') + re^{\beta\mu_\pi(s)}\mu_\pi(s') \ge (1-r)e^{\beta}\gamma^2$ along with Assumption 4.2 and the convention that $\lesssim$ subsumes γ^{-1} terms, and also that $\left|\frac{E^{(1)}_{s,s'}(X)}{E^{(2)}_{s,s'}(X)+\epsilon}\right| \le 1$.

We can use Lemma F.7 to bound $\mathcal{E}_4$:

$$
\begin{aligned}
\mathbb{E}[\mathcal{E}_4(X)^2] &= \frac{\left|\overline{\mathcal{R}}\right|^2}{T^2}\,\mathbb{E}[(\hat{c}_{X_{\overline{\mathcal{R}}}}(s,s') - \mu_\pi(s)\pi(s' \mid s)^2] \\
&\lesssim \frac{\left|\overline{\mathcal{R}}\right|^2}{T^2} \cdot \frac{\mu_\pi(s')T^2}{T_{\text{eff}}(\lambda)\left|\overline{\mathcal{R}}\right|^2} \\
&\lesssim \frac{1}{T_{\text{eff}}(\lambda)}.
\end{aligned}
$$

Next, we use Lemma F.5 to bound $\mathcal{E}_5$:

$$
\begin{aligned}
\mathbb{E}[\mathcal{E}_5(X)^2] &= \frac{\left|\overline{\mathcal{R}}\right|^2}{T^2}\,\mathbb{E}[(\hat{\mu}_{X_{\overline{\mathcal{R}}}}(s') - \mu_\pi(s'))^2] \\
&\lesssim \frac{\left|\overline{\mathcal{R}}\right|^2}{T^2} \cdot \frac{\mu_\pi(s')T^2}{T_{\text{eff}}(\lambda)\left|\overline{\mathcal{R}}\right|^2} \\
&\lesssim \frac{1}{T_{\text{eff}}(\lambda)}.
\end{aligned}
$$

Next, we bound $\mathcal{E}_6$:

$$\begin{aligned}|\mathcal{E}_6(X)| &\le \frac{1}{T}\sum_{i\in\mathcal{R}}\left|x_{i,s'}\left(e^{\beta\tilde{z}(X;s)}\tilde{z}(X;s) - e^{\beta\mu_\pi(s)}\mu_\pi(s)\right)\right| + \frac{1}{T}\left|e^{\beta\mu_\pi(s)}\mu_\pi(s)\sum_{i\in\mathcal{R}}(x_{i,s'} - \mu_\pi(s'))\right| \\ &\le \frac{1}{T}\sum_{i\in\mathcal{R}}(1+\beta)e^{\beta}|\tilde{z}(X;s) - \mu_\pi(s)| + \frac{1}{T}e^{\beta}\left|\sum_{i\in\mathcal{R}}(x_{i,s'} - \mu_\pi(s'))\right|.\end{aligned}$$

The first term can be bounded equivalently as to was done for $\mathcal{E}_2$, and thus

$$\mathbb{E}\left[\left(\frac{1}{T}\sum_{i\in\mathcal{R}}(1+\beta)e^{\beta}|\tilde{z}(X;s) - \mu_\pi(s)|\right)^2\right] \lesssim \frac{(1+\beta)^2e^{2\beta}\log T}{T_{\text{eff}}^{1/2}}.$$

In the second term, since $x_{i,s'} - \mu_\pi(s')$ are independent and mean 0 for all $i \neq T$,

$$\begin{aligned}\mathbb{E}\left[\left(\frac{1}{T}e^{\beta}\left|\sum_{i\in\mathcal{R}}(x_{i,s'} - \mu_\pi(s'))\right|\right)^2\right] &= \frac{e^{2\beta}}{T^2}\sum_{i\in\mathcal{R}}\mathbb{E}\left[(x_{i,s'} - \mu_\pi(s'))^2\right] \\ &\lesssim \frac{e^{2\beta}}{T}.\end{aligned}$$

Altogether

$$\mathbb{E}[|\mathcal{E}_6(X)|^2] \lesssim \frac{(1+\beta)^2e^{2\beta}\log T}{T_{\text{eff}}^{1/2}}.$$

Finally, we bound $\mathcal{E}_7$:

$$\begin{aligned}|\mathcal{E}_7(X)| &\le \frac{1}{T}\sum_{i\in\mathcal{R}}\left|x_{i,s'}\left(e^{\beta\tilde{z}(X;s)} - e^{\beta\mu_\pi(s)}\right)\right| + \frac{1}{T}\left|e^{\beta\mu_\pi(s)}\sum_{i\in\mathcal{R}}(x_{i,s'} - \mu_\pi(s'))\right| \\ &\le \frac{1}{T}\sum_{i\in\mathcal{R}}\beta e^{\beta}|\tilde{z}(X;s) - \mu_\pi(s)| + \frac{1}{T}e^{\beta}\left|\sum_{i\in\mathcal{R}}(x_{i,s'} - \mu_\pi(s'))\right|.\end{aligned}$$

Thus via an identical calculation as $\mathcal{E}_6$,

$$\mathbb{E}[|\mathcal{E}_7(X)|^2] \lesssim \frac{(1+\beta)^2e^{2\beta}\log T}{T_{\text{eff}}^{1/2}}.$$

Altogether,

$$\begin{aligned}&\mathbb{E}\left[\left|\frac{E^{(1)}_{s,s'}(X)}{E^{(2)}_{s,s'}(X)+\epsilon} - \frac{(1-r)e^{\beta}\mu_\pi(s)\pi(s' \mid s) + re^{\beta\mu_\pi(s)}\mu_\pi(s')\mu_\pi(s)}{(1-r)(e^{\beta}-1)\mu_\pi(s)\pi(s' \mid s) + (1-r)\mu_\pi(s') + re^{\beta\mu_\pi(s)}\mu_\pi(s')}\right|^2\right] \\ &\lesssim (1+\beta^2)\cdot\frac{\log T}{T_{\text{eff}}^{1/2}}.\end{aligned}$$

□

Lemma H.8. *Let $\hat{\theta} = \left(\hat{A}^{(1)}, \hat{A}^{(2)}\right)$ be the output of Algorithm 1, where $\hat{A}^{(2)} = (\beta_0 + \beta(\tau_1+\tau_2))I_S - \frac{\beta(\tau_1+\tau_2)}{S}1_S1_S^\top$. Then*

$$\mathbb{E}_X\left[\left|f_{\hat{\theta}}(X;s)_{s'} - \pi(s' \mid s)\right|^2\right] \lesssim (1+\beta^{*2})\cdot\frac{\log T}{T_{\mathit{eff}}^{1/2}} + e^{-\beta^*\gamma}.$$

and

$$\mathbb{P}\left[f_{\hat{\theta}}(X;s)_{s'} \le \frac{\gamma^3}{4S^2}\right] \lesssim \frac{1}{T_{\mathit{eff}}}.$$

Proof. First, by Lemma D.10, $1+\beta^* \ge \beta(\tau_1+\tau_2) \ge \beta^*$. For notational convenience, let $\beta = \beta(\tau_1+\tau_2)$ Recall the definitions

$$\mathcal{S}\left(A_*^{(1)}\right)_{ij} = \begin{cases} \mathbf{1}(j = p(i)) & p(i) \neq \emptyset \\ \hat{A}^{(1)}_{i,j} & p(i) = \emptyset \end{cases}.$$

and

$$\tilde{z}(X;s) = \mathcal{S}(A_*^{(1)})\delta_s(X) = \begin{cases} x_{p(i),s} & \text{if } i \notin \mathcal{R} \\ z_{\hat{\theta}}(X;s)_i & \text{if } i \in \mathcal{R} \end{cases}.$$

By Corollary D.6 $\|z_{\hat{\theta}}(X;s) - \tilde{z}(X;s)\|_\infty \lesssim T^{-1}$. Letting $f(z) = \delta^\top \mathcal{S}(\beta z)$, we see that $\|\nabla_z f(z)\|_1 = \beta\|J(\mathcal{S}(\beta z))\delta\|_1 \le 2\beta$, and thus

$$\left|f_{\hat{\theta}}(X;s)_{s'} - \delta_{s'}(X)^\top \mathcal{S}(\beta\tilde{z}(X;s))\right| \lesssim \beta T^{-1}.$$

Next, we have that

$$\delta_{s'}(X)^\top \mathcal{S}(\beta\tilde{z}(X;s)) = \frac{\sum_i x_{i,s'}\exp(\beta\tilde{z}(X;s)_i)}{\sum_i \exp(\beta\tilde{z}(X;s)_i)},$$

and thus

$$\begin{aligned}
&\delta_{s'}(X)^\top \mathcal{S}(\beta\tilde{z}(X;s)) \\
&= \frac{(e^\beta - 1)\sum_{i\in\overline{\mathcal{R}}} x_{p(i),s}x_{i,s'} + \sum_{i\in\overline{\mathcal{R}}} x_{i,s'} + \sum_{i\in\mathcal{R}} x_{i,s'}e^{\beta\tilde{z}(X;s)_i}}{(e^\beta-1)\sum_{i\in\overline{\mathcal{R}}} x_{p(i),s} + \left|\overline{\mathcal{R}}\right| + \sum_{i\in\mathcal{R}} e^{\beta\tilde{z}(X;s)_i}} \\
&= \frac{(1-r)(e^\beta-1)\mu_\pi(s)\pi(s' \mid s) + (1-r)\mu_\pi(s') + r\mu_\pi(s')e^{\beta\mu_\pi(s)} + (e^\beta-1)\mathcal{E}_4(X) + \mathcal{E}_5(X) + \mathcal{E}_7(X)}{(1-r)(e^\beta-1)\mu_\pi(s) + (1-r) + re^{\beta\mu_\pi(s)} + (e^\beta-1)\mathcal{E}_1(X) + \mathcal{E}_3(X)}
\end{aligned}$$

Therefore by Lemma H.4,

$$\begin{aligned}
&\left|\delta_{s'}(X)^\top \mathcal{S}(\beta\tilde{z}(X;s)) - \frac{(1-r)(e^\beta-1)\mu_\pi(s)\pi(s' \mid s) + (1-r)\mu_\pi(s') + r\mu_\pi(s')e^{\beta\mu_\pi(s)}}{(1-r)(e^\beta-1)\mu_\pi(s) + (1-r) + re^{\beta\mu_\pi(s)}}\right| \\
&\quad\le \frac{\left|\delta_{s'}(X)^\top \mathcal{S}(\beta\tilde{z}(X;s))\right| \cdot \left(e^\beta|\mathcal{E}_4(X)| + |\mathcal{E}_5(X)| + |\mathcal{E}_7(X)|\right) + e^\beta|\mathcal{E}_1(x)| + |\mathcal{E}_3(X)|}{(1-r)(e^\beta-1)\mu_\pi(s) + (1-r) + re^{\beta\mu_\pi(s)}} \\
&\quad\lesssim |\mathcal{E}_1(x)| + e^{-\beta}|\mathcal{E}_3(X)| + |\mathcal{E}_4(X)| + e^{-\beta}|\mathcal{E}_5(X)| + e^{-\beta}|\mathcal{E}_7(X)|,
\end{aligned}$$

where the last inequality uses Assumption 4.2. Next, see that

$$\begin{aligned}
&\left|\frac{(1-r)(e^\beta-1)\mu_\pi(s)\pi(s' \mid s) + (1-r)\mu_\pi(s') + r\mu_\pi(s')e^{\beta\mu_\pi(s)}}{(1-r)(e^\beta-1)\mu_\pi(s) + (1-r) + re^{\beta\mu_\pi(s)}} - \pi(s' \mid s)\right| \\
&\quad\le \frac{\left|(1-r)\mu_\pi(s') + r\mu_\pi(s')e^{\beta\mu_\pi(s)} + (1-r)\pi(s' \mid s) + re^{\beta\mu_\pi(s)}\pi(s' \mid s)\right|}{(1-r)(e^\beta-1)\mu_\pi(s) + (1-r) + re^{\beta\mu_\pi(s)}} \\
&\quad\lesssim \frac{e^{\beta\mu_\pi(s)}}{(1-r)e^\beta\gamma} \\
&\quad\lesssim e^{\beta(\mu_\pi(s)-1)} \\
&\quad\lesssim e^{-\beta\frac{\gamma(S-1)}{S}} \\
&\quad\lesssim e^{-\beta\gamma/2}
\end{aligned}$$

Altogether, we get

$$\left|f_{\hat{\theta}}(X;s)_{s'} - \pi(s' \mid s)\right| \lesssim \frac{\beta}{T} + |\mathcal{E}_4(X)| + |\mathcal{E}_1(x)| + e^{-\beta}|\mathcal{E}_3(X)| + e^{-\beta}|\mathcal{E}_5(X)| + e^{-\beta}|\mathcal{E}_7(X)| + e^{-\beta\gamma/2},$$

and thus

$$\mathbb{E}_X\left[\left|f_{\hat{\theta}}(X;s)_{s'} - \pi(s' \mid s)\right|^2\right] \lesssim (1+\beta^2)\cdot\frac{\log T}{T_{\text{eff}}^{1/2}} + e^{-\beta\gamma} \lesssim (1+\beta^{*2})\cdot\frac{\log T}{T_{\text{eff}}^{1/2}} + e^{-\beta^*\gamma}$$

Next, we need to bound $\mathbb{P}_X\left[f_{\hat{\theta}}(X;s)_{s'} \le \frac{\gamma^3}{4S^2}\right]$. We start by bounding the probability $\delta_{s'}(X)^\top \mathcal{S}(\beta\tilde{z}(X;s))$ is small. We have the naive bound

$$\begin{aligned}
\delta_{s'}(X)^\top \mathcal{S}(\beta\tilde{z}(X;s)) &= \frac{\sum_i x_{i,s'}\exp(\beta\tilde{z}(X;s)_i)}{\sum_i \exp(\beta\tilde{z}(X;s)_i)} \\
&\ge \frac{e^\beta \sum_{i\in\overline{\mathcal{R}}} x_{p(i),s}x_{i,s'}}{e^\beta\cdot T} \\
&= \frac{1}{T}\sum_{i\in\overline{\mathcal{R}}} x_{p(i),s}x_{i,s'} \\
&= (1-r)\hat{c}_{X_{\overline{\mathcal{R}}}}(s,s').
\end{aligned}$$

By Markov's inequality and Lemma F.7,

$$\begin{aligned}
\mathbb{P}_X\left[\hat{c}_{X_{\overline{\mathcal{R}}}}(s,s') \le \frac{\gamma^2}{2S^2}\right] &\le \mathbb{P}_X\left[\left|\hat{c}_{X_{\overline{\mathcal{R}}}}(s,s') - \mu_\pi(s)\pi(s'\mid s)\right| \ge \frac{\gamma^2}{2S^2}\right] \\
&\le \frac{2S^2}{\gamma^2}\,\mathbb{E}_X\left[\left|c_{X_{\overline{\mathcal{R}}}}(s,s') - \mu_\pi(s)\pi(s'\mid s)\right|^2\right] \\
&\lesssim \frac{T^2}{\left|\overline{\mathcal{R}}\right|^2 T_{\text{eff}}} \\
&\lesssim \frac{1}{T_{\text{eff}}}.
\end{aligned}$$

Therefore

$$\mathbb{P}_X\left[\delta_{s'}(X)^\top \mathcal{S}(\beta\tilde{z}(X;s)) \le \frac{\gamma^3}{2S^2}\right] \le \mathbb{P}_X\left[\delta_{s'}(X)^\top \mathcal{S}(\beta\tilde{z}(X;s)) \le (1-r)\frac{\gamma^2}{2S^2}\right] \lesssim \frac{1}{T_{\text{eff}}}.$$

To conclude, on the event that $\delta_{s'}(X)^\top \mathcal{S}(\beta\tilde{z}(X;s)) > \frac{\gamma^3}{2S^2}$, we have

$$\begin{aligned}
f_{\hat{\theta}}(X;s)_{s'} &> \frac{\gamma^3}{2S^2} - \left|f_{\hat{\theta}}(X;s)_{s'} - \delta_{s'}(X)^\top \mathcal{S}(\beta\tilde{z}(X;s))\right| \\
&\ge \frac{\gamma^3}{2S^2} - O\big(\beta T^{-1}\big) \\
&\ge \frac{\gamma^3}{4S^2},
\end{aligned}$$

since $\beta \le 1 + \beta^* \lesssim T$. Altogether,

$$\mathbb{P}_X\left[f_{\hat{\theta}}(X;s)_{s'} \le \frac{\gamma^3}{4S^2}\right] \lesssim \frac{1}{T_{\text{eff}}}.$$

□

H.4. Proof of Theorem 4.5

Proof. By Lemma H.8, we get that

$$\mathbb{E}_X\left[\left(f(X;s)_{s'} - \pi(s'\mid s)\right)^2\right] \lesssim_{\gamma,S} \frac{\log T}{T_{\text{eff}}^{\Theta_\gamma(1)}}$$

Therefore by Markov's inequality,

$$\mathbb{P}_X\left[\left(f(X;s)_{s'} - \pi(s' \mid s)\right)^2 \geq 100S^2 \cdot \mathbb{E}_X\left[\left(f(X;s)_{s'} - \pi(s' \mid s)\right)^2\right]\right] \leq \frac{1}{100S^2}.$$

Union bounding, with probability 0.99 we have

$$\sup_{s,s'} \left|f(X;s)_{s'} - \pi(s' \mid s)\right| \leq 100S^2 \cdot \mathbb{E}_X\left[\left(f(X;s)_{s'} - \pi(s' \mid s)\right)^2\right] \lesssim_{\gamma,S} \frac{\log T}{T_{\text{eff}}^{\Theta_\gamma(1)}},$$

as desired. □

I. Finite Sample Analysis

Our theory focuses on the case of gradient descent on the population loss (11). It is relatively straightforward to extend our analysis to the finite sample setting. In this case, we are given a dataset of N prompts of length T:

$$\mathcal{D} = \{s_{1:T}^{(n)}\}_{n \in [N]}.$$

Each sequence $s_{1:T}^{(n)}$ is generated via the procedure in Task 2.4, with transition matrix $\pi^{(i)} \sim P_\pi$. Let $X^{(n)} \in \mathbb{R}^{T \times S}$ be the embedding of $s_{1:T}^{(n)}$.

We now consider running gradient descent on the finite sample loss $\hat{L}$:

$$\hat{L}(\theta) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{s' \in [S]} \pi^{(n)}(s' \mid s_T^{(n)}) \log\left(f_\theta(s_{1:T}^{(n)}) + \epsilon\right). \tag{37}$$

Below, we present a sketch of the extension of the analysis of our main theorem to this finite sample setting.

I.1. Stage 1

The crux of Stage 1 is Lemma D.3, where in (27) we show that

$$\nabla_{A_i^{(1)}} L(\theta) = -\frac{\beta_0}{ST} J\left(\mathcal{S}\left(A_i^{(1)}\right)\right) g_i^*,$$

where the vector $g_i^* \in \mathbb{R}$ is defined by

$$g_i^* := T\sum_{s,s'} \mathbb{E}_{\pi,X}\left[\frac{\pi(s' \mid s)}{f_\theta(X;s)_{s'} + \epsilon}\delta_{s'}(X)^\top J(v_\theta(X;s))e_i \cdot \delta_s(X_{\leq i})\right].$$

In particular, we show that g_i^* satisfies the property that

$$g_{i,j}^* - g_{i,p(i)}^* \leq -\frac{\gamma^3}{4S}$$

for $i \in \overline{\mathcal{R}}$ and $\left|g_{i,j}^*\right| \lesssim T_{\text{eff}}^{-1/2}$ for $i \in \mathcal{R}$. As a step towards proving this, we let $\hat{\theta} = (A^{(1)}, 0)$, define the quantity $\hat{g}_i$ by

$$\hat{g}_i := T\sum_{s,s'} \mathbb{E}_{\pi,X}\left[\frac{\pi(s' \mid s)}{f_\theta(X;s)_{s'} + \epsilon}\delta_{s'}(X)^\top J(v_\theta(X;s))e_i \cdot \delta_s(X_{\leq i})\right],$$

and show that

$$\hat{g}_{i,j} - \hat{g}_{i,p(i)} \leq -\frac{\gamma^3}{4S} \text{ for } i \in \overline{\mathcal{R}}$$
$$|\hat{g}_{i,j}| \lesssim T_{\text{eff}}^{-1/2} \text{ for } i \in \mathcal{R}.$$

The empirical gradient can be written as

$$\nabla_{A_i^{(1)}} \hat{L}(\theta) = -\frac{\beta_0}{ST} J\Big(\mathcal{S}\Big(A_i^{(1)}\Big)\Big) g_i^{\text{emp}},$$

where

$$g_i^{\text{emp}} = \frac{ST}{N} \sum_{n=1}^{N} \sum_{s'} \frac{\pi^{(n)}(s' \mid s_T^{(n)})}{f_\theta(X^{(n)}; s_T^{(n)})_{s'} + \epsilon} \delta_{s'}(X^{(n)})^\top J(v_\theta(X^{(n)}; s_T^{(n)})) e_i \cdot \delta_{s_T^{(n)}}(X_{\le i}^{(n)}).$$

As in the population setting, we define $\hat{g}_i^{\text{emp}}$ to be

$$\hat{g}_i^{\text{emp}} = \frac{ST}{N} \sum_{n=1}^{N} \sum_{s'} \frac{\pi^{(n)}(s' \mid s_T^{(n)})}{f_{\hat{\theta}}(X^{(n)}; s_T^{(n)})_{s'} + \epsilon} \delta_{s'}(X^{(n)})^\top J(v_{\hat{\theta}}(X^{(n)}; s_T^{(n)})) e_i \cdot \delta_{s_T^{(n)}}(X_{\le i}^{(n)}).$$

By Lemma G.1, we get that $\|g_i^{\text{emp}} - \hat{g}_i^{\text{emp}}\|_\infty \le \frac{C_{\gamma,S}}{\sqrt{T_{\text{eff}}}}$. It thus suffices to show that $\|\hat{g}_i^{\text{emp}} - \hat{g}_i\|_\infty$ is small, which is given by the following lemma:

Lemma I.1. *For any $\delta > 0$*

$$\|\hat{g}_i^{\text{emp}} - \hat{g}_i\|_\infty \le \frac{C_{\gamma,S} \log\left(\frac{T}{\delta}\right)}{\sqrt{N}}$$

with probability $1 - \delta$.

Proof. First, see that $\hat{g}_{i,j}^{\text{emp}}$ can be written as

$$\hat{g}_{i,j}^{\text{emp}} = \frac{1}{N} \sum_{n=1}^{N} \underbrace{S \sum_{s'} \frac{\pi^{(n)}(s' \mid s_T^{(n)})}{\hat{\mu}_{X^{(n)}}(s') + \epsilon} (x_{i,s'}^{(n)} - \hat{\mu}_{X^{(n)}}(s')) x_{j,s_T^{(n)}}^{(n)}}_{=:Z^{(n)}}.$$

Define the event $\mathcal{A}^{(n)}$ as

$$\mathcal{A}^{(n)} = \bigcup_{s' \in [S]} \{\hat{\mu}_{X^{(n)}}(s') \le \frac{1}{2} \mu_{\pi^{(n)}}(s')\}.$$

By a union bound, $\mathbb{P}(\mathcal{A}^{(n)}) \le S/T_{\text{eff}}$. We can naively bound $\left|Z^{(n)}\right| \le \epsilon^{-1} S$, and on the complement of $\mathcal{A}^{(n)}$ (denoted by $\overline{\mathcal{A}^{(n)}}$) we can bound $\left|Z^{(n)}\right| \lesssim S\gamma^{-1}$. Therefore we can concentrate $\frac{1}{N}\sum_n Z^{(n)}$ as:

$$\begin{aligned} \left|\frac{1}{N}\sum_n Z^{(n)} - \mathbb{E}[Z]\right| &\le \left|\frac{1}{N}\sum_n Z^{(n)} \mathbf{1}(\overline{\mathcal{A}^{(n)}}) - \mathbb{E}[Z^{(n)} \mathbf{1}(\overline{\mathcal{A}^{(n)}})]\right| + \left|\frac{1}{N}\sum_n Z^{(n)} \mathbf{1}(\mathcal{A}^{(n)}) - \mathbb{E}[Z^{(n)} \mathbf{1}(\mathcal{A}^{(n)})]\right| \\ &\lesssim \frac{S\gamma^{-1} \log(T/\delta)}{\sqrt{N}} + \epsilon^{-1} \cdot \frac{1}{N} \sum_n \mathbf{1}(\mathcal{A}^{(n)}) + \epsilon^{-1} \mathbb{P}(\mathcal{A}^{(n)}), \end{aligned}$$

with probability $1 - \frac{\delta}{2T^2}$ by Hoeffding's inequality on the $Z^{(n)} \mathbf{1}(\mathcal{A}^{(n)})$. Next, we see that the $\mathbf{1}(\mathcal{A}^{(n)})$ are Bernoulli random variables with mean $\mathbb{P}(\mathcal{A}^{(n)}) \le S/T_{\text{eff}}$ and standard deviation at most $\sqrt{\mathbb{P}(\mathcal{A}^{(n)})/N}$. Therefore with probability $1 - \frac{\delta}{T^2}$, one has

$$\begin{aligned} \left|\frac{1}{N}\sum_n Z^{(n)} - \mathbb{E}[Z]\right| &\lesssim \frac{S\gamma^{-1} + \epsilon^{-1}\sqrt{S/T_{\text{eff}}}}{\sqrt{N}} \log(T/\delta) \\ &= \frac{C_{\gamma,S} \log(T/\delta)}{\sqrt{N}}, \end{aligned}$$

since $\epsilon = \Theta(T_{\text{eff}}^{-1/2})$. Union bounding over $i, j \in [T]$ yields the desired result □

Combining everything together, we get that for $N \gtrsim C_{\gamma,S} T_{\text{eff}} \log T$, the quantities g_i^{emp} satisfy

$$g_{i,j}^{\text{emp}} - g_{i,p(i)}^{\text{emp}} \le -\frac{\gamma^3}{4S} \text{ for } i \in \overline{\mathcal{R}}$$
$$\left|g_{i,j}^{\text{emp}}\right| \lesssim T_{\text{eff}}^{-1/2} \text{ for } i \in \mathcal{R}.$$

with high probability, and thus Stage 1 succeeds on the empirical loss with high probability.

I.2. Stage 2

One challenge in directly using the population analysis for stage 2 is that the finite-sample update no longer preserves symmetry, and hence we do not have that $A^{(2)} = \beta_0 I_S + \beta(I_S - \frac{1}{S}1_S 1_S^\top)$ throughout the entirety of stage 2. Instead, we will consider taking only a single large gradient step with learning rate η_2, on an independent dataset of N prompts.

During the first step of stage 2, the population gradient is

$$\nabla_{A^{(2)}} L(\theta) = -\beta^{\text{pop}} \cdot \left(I_S - \frac{1}{S} 1_S 1_S^\top\right),$$

where, by Lemma D.8,

$$1 \ge \beta^{\text{pop}} \ge C_{\gamma,S}^{-1} > 0.$$

The following lemma concentrates the population gradient to the empirical gradient at $\theta(\tau_1)$:

Lemma I.2.

$$\left\|\nabla_{A^{(2)}} \hat{L}(\theta) - \nabla_{A^{(2)}} L(\theta)\right\|_\infty \lesssim \frac{C_{\gamma,S}}{\sqrt{N}}.$$

Proof. The finite sample gradient can be written as

$$\nabla_{A^{(2)}} \hat{L}(\theta) = -\frac{1}{N} \sum_{n=1}^N \underbrace{\sum_{s'} \frac{\pi^{(n)}(s' \mid s_T^{(n)})}{f_\theta(X^{(n)}; s_T^{(n)})_{s'} + \epsilon} \cdot {X^{(n)}}^\top \mathcal{S}(A^{(1)})^\top J(v_\theta(X^{(n)}; s_T^{(n)})) \delta_{s'}(X^{(n)}) e_{s_T^{(n)}}^\top}_{M^{(n)}}.$$

Let $\hat{\theta} = (A^{(1)}(\tau_1), 0)$. By (36), we have

$$\left|f_\theta(X^{(n)}; s_T^{(n)})_{s'} - f_{\hat{\theta}}(X^{(n)}; s_T^{(n)})_{s'}\right| \lesssim e^{\beta_0 - 1} \lesssim T_{\text{eff}}^{-1/2}.$$

Additionally, $f_{\hat{\theta}}(X^{(n)}; s_T^{(n)})_{s'} = \hat{\mu}_{X^{(n)}}(s')$. We next see that we can bound

$$\begin{aligned}
&\left|e_s^\top {X^{(n)}}^\top \mathcal{S}(A^{(1)})^\top J(v_\theta(X^{(n)}; s_T^{(n)})) \delta_{s'}(X^{(n)})\right| \\
&\le 2\left\|\mathcal{S}(A^{(1)}) X^{(n)} e_s\right\|_\infty \left\|\delta_{s'}(X^{(n)})\right\|_\infty \\
&\le 2.
\end{aligned}$$

Therefore on $\mathcal{A}^{(n)}$, each entry of $M^{(n)}$ can be bounded in absolute value by $2\epsilon^{-1}$, and on $\overline{\mathcal{A}^{(n)}}$ each entry can be bounded by some $C_{\gamma,S}$. Therefore by an identical concentration argument as in Lemma I.1, with high probability we get that

$$\left\|\nabla_{A^{(2)}} \hat{L}(\theta) - \nabla_{A^{(2)}} L(\theta)\right\|_\infty \lesssim \frac{C_{\gamma,S}}{\sqrt{N}}.$$

□

After one gradient step with learning rate η_2, $A^{(2)}(\tau_1+1)$ is equal to

$$A^{(2)}(\tau_1+1) = \beta_0 I_S + \eta_2 \beta^{\mathrm{pop}} \cdot \left(I_S - \frac{1}{S} 1_S 1_S^\top \right) + R,$$

where the error matrix R satisfies $\|R\|_\infty \lesssim \frac{\eta_2}{\sqrt{N}}$.

Next, define the parameter vector θ^{pop} as $\theta^{\mathrm{pop}} = (A^{(1)}(\tau_1), A^{(2)}_{\mathrm{pop}})$, where $A^{(2)}_{\mathrm{pop}} = \beta_0 I_S + \eta_2 \beta^{\mathrm{pop}} \cdot \left(I_S - \frac{1}{S} 1_S 1_S^\top \right)$ is the result of the population update. We can bound the error between the finite-sample predictor and population predictor as

$$\begin{aligned} \left| f_{\theta(\tau_1+1)}(X;s)_{s'} - f_{\theta^{\mathrm{pop}}}(X;s)_{s'} \right| &= \left| \delta_{s'}^\top \left(\mathcal{S}(\mathcal{S}(A^{(1)})XA^{(2)}(\tau_1+1)e_s) - \mathcal{S}(\mathcal{S}(A^{(1)})XA^{(2)}_{\mathrm{pop}}e_s) \right) \right| \\ &\le \|\mathcal{S}(\mathcal{S}(A^{(1)})XA^{(2)}(\tau_1+1)e_s) - \mathcal{S}(\mathcal{S}(A^{(1)})XA^{(2)}_{\mathrm{pop}}e_s)\|_\infty \\ &\le \|\mathcal{S}(A^{(1)})XA^{(2)}(\tau_1+1)e_s - \mathcal{S}(A^{(1)})XA^{(2)}_{\mathrm{pop}}e_s\|_\infty \\ &\le \|A^{(2)}(\tau_1+1) - A^{(2)}_{\mathrm{pop}}\|_\infty \\ &\lesssim \frac{\eta_2}{\sqrt{N}}. \end{aligned}$$

Now if we choose η_2 so that $\eta_2 \beta^{\mathrm{pop}} = \beta^* = \Theta(\log T_{\text{eff}})$, then by Lemma H.8, we get that θ^{pop} satisfies

$$\mathbb{E}_X \left[|f_{\theta^{\mathrm{pop}}}(X;s)_{s'} - \pi(s' \mid s)|^2 \right] \lesssim T_{\text{eff}}^{-c\gamma}.$$

Therefore

$$\begin{aligned} \mathbb{E}_X \left[\left| f_{\theta(\tau_1+1)}(X;s)_{s'} - \pi(s' \mid s) \right|^2 \right] &\lesssim T_{\text{eff}}^{c\gamma} + \frac{\eta_2^2}{N} \\ &\lesssim T_{\text{eff}}^{-c\gamma} + \frac{\log^2 T_{\text{eff}}}{N} \\ &\lesssim T_{\text{eff}}^{-c\gamma}, \end{aligned}$$

as long as $N \gtrsim T_{\text{eff}}^{c\gamma}$. Therefore the output of running gradient descent on the finite sample loss, $f_{\theta(\tau_1+1)}$, achieves small population loss.

Altogether, both Stage 1 and Stage 2 succeed on the finite-sample loss as long as $N \gtrsim T_{\text{eff}} \log T$.