# Transformers Provably Learn Algorithmic Solutions for Graph Connectivity, But Only with the Right Data

**Qilin Ye** [*] [♠] [♣] **Deqing Fu** [*] [♠] **Robin Jia** [♠] **Vatsal Sharan** [♠]

## Abstract

Transformers often fail to learn generalizable algorithms, instead relying on brittle heuristics. Using graph connectivity as a testbed, we explain this phenomenon both theoretically and empirically. We consider a simplified Transformer architecture, the Disentangled Transformer, and prove that an $L$-layer model can compute connectivity in graphs with diameters up to $3^L$, implementing an algorithm equivalent to computing powers of the adjacency matrix. By analyzing training dynamics, we prove that whether the model learns this strategy hinges on whether most training instances are within this model capacity. Within-capacity graphs (diameter $\leq 3^L$) drive the learning of the algorithmic solution while beyond-capacity graphs drive the learning of a simple heuristic based on node degrees. Finally, we empirically show that restricting training data to stay within a model's capacity makes both standard and Disentangled Transformers learn the exact algorithm.

## 1 Introduction

Large language models (LLMs) based on the Transformer architecture have demonstrated remarkable capabilities, yet their success is often shadowed by failures on tasks that demand robust, algorithmic reasoning. A growing body of evidence shows that, instead of learning generalizable algorithms, these models frequently rely on brittle shortcuts and spurious correlations that exploit statistical cues in the training data (Niven & Kao, 2019; Geirhos et al., 2020; Tang et al., 2023; Yuan et al., 2024; Zhou et al., 2024d; Ye et al., 2025). This shortcut reliance contributes to poor out-of-distribution (OOD) generalization, brittleness under superficial input changes, and unreliability on multi-step reasoning tasks (Zou et al., 2023; Deng et al., 2024; Li et al., 2024; Mirzadeh et al., 2025). On deterministic tasks like shift ciphers, LLMs favor high-probability outputs over correct solutions (McCoy et al., 2023); in mathematical problem solving, strong in-distribution scores often fail to transfer as problem structure or size shifts (Saxton et al., 2019; Kao et al., 2024; Zhou et al., 2024b). This motivates a foundational question:

***When and why do Transformers learn heuristics over verifiably correct algorithms, even when the task admits an algorithmic solution?***

To study this question, we adopt *graph connectivity* as a controlled testbed. Connectivity is a fundamental problem in computational complexity (Wigderson, 1992) that combines three properties suited to our analysis. First, it admits an unambiguous algorithmic solution: reachability equals the transitive closure and is computable via matrix powering (Warshall, 1962; Floyd, 1962). Second, recent theory shows that Transformers with depth $L = \Theta(\log n)$ can express this algorithm through matrix powering constructions (Merrill & Sabharwal, 2025), so the architecture is provably expressive enough. Third, connectivity admits simple heuristics based on node degrees that correlate with the correct answer on typical random graphs but fail on adversarial instances. This makes it possible to test whether training recovers the algorithm or settles for the shortcut.

Despite these expressivity guarantees, whether gradient descent actually finds the algorithmic solution remains open. Our experiments reveal that standard Transformers achieve perfect in-distribution accuracy on random graphs yet fail catastrophically on simple OOD instances such as two disjoint chains (see §3.3 and Figure 1). To understand both the failure mode and how to correct it, we analyze the *Disentangled Transformer* (Friedman et al., 2023; Nichani et al., 2024), a simplified architecture that is more amenable to theoretical analysis while preserving the essential computations. We summarize our contributions below.

**An $L$-layer model solves connectivity up to diameter $3^L$, but no further.** We prove tight bounds that characterize model capacity in terms of graph diameter rather than the number of nodes. Let $\mathrm{diam}(G)$ denote the maximum shortest-path distance between any two connected nodes. On the expressivity side, we show that an $L$-layer Disen-

[*]Equal contribution [♠]Thomas Lord Department of Computer Science, University of Southern California. [♣]Department of Computer Science, Duke University. Correspondence to: Qilin Ye <qilin.ye@duke.edu>, Deqing Fu <deqingfu@usc.edu>.

*Preprint. February 19, 2026.*

tangled Transformer can solve connectivity on all graphs with $\mathrm{diam}(G) \leq 3^L$ by implementing a matrix powering algorithm (Theorem 4.3). On the limitations side, we prove a matching upper bound: for any choice of nonnegative weights, there exists a graph with diameter $3^L + 1$ on which the model fails (Theorem 4.5). Together, these results establish that $3^L$ is the maximum diameter an $L$-layer model can handle perfectly. We call this quantity the model's *capacity*. We empirically validate this diameter-depth scaling on both disentangled and standard Transformers.

**Learned weights decompose into algorithmic and heuristic channels.** We prove that under natural symmetry assumptions, specifically invariance to relabeling of graph vertices, the weights of a trained Disentangled Transformer decompose into two functionally distinct components (Theorem 4.7). The *algorithmic channel* preserves locality and implements multi-hop composition by computing powers of the adjacency matrix. The *heuristic channel* ignores graph structure and instead computes global statistics based on node degrees, predicting connectivity from whether two nodes both have high degree. We verify empirically that trained models satisfy the required symmetry, making this decomposition applicable to practical settings (§4.3).

**Training dynamics select between channels based on the data distribution.** Our analysis of the training dynamics reveals a sharp dichotomy driven by the data distribution. For graphs within the model's capacity (diameter $\leq 3^L$), population gradients suppress the heuristic channel and favor the algorithmic channel that implements matrix powering (Theorem C.5). Conversely, when the distribution contains a significant share of beyond-capacity graphs (diameter $> 3^L$), the gradients instead strengthen the heuristic channel, promoting the simple degree-counting shortcut (Theorem C.9). This precise characterization hinges on our exact $3^L$ capacity bound; an asymptotic one, such as the $\mathcal{O}(\exp(L))$ result from Merrill & Sabharwal (2025), would not yield such clear predictive implications.

**Transformers learns algorithmic solution *with the right data*.** These theoretical insights point to a direct mitigation strategy we call the *data lever*: restricting the training data exclusively to within-capacity graphs. Our experiments in §5 confirm the effectiveness of this approach, showing that it boosts the algorithmic component and improves OOD robustness (Figure 4), and that these benefits transfer successfully to standard Transformer models (Figure 7).

Our results, validated on both Disentangled and standard Transformers (Figures 2, 7 and 10), provide a precise account of how Transformers learn algorithms given the right data.

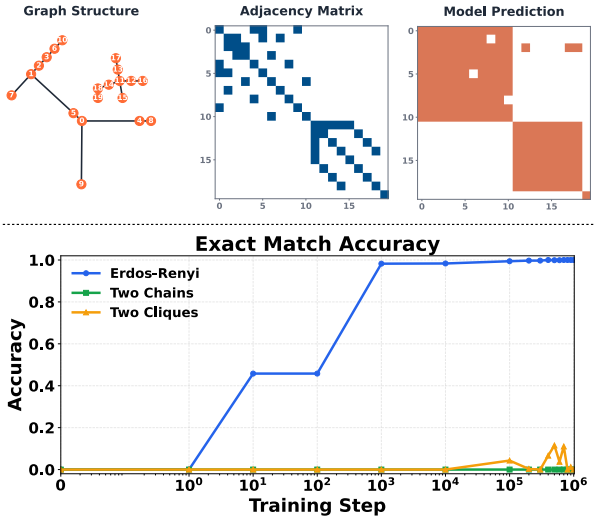



*Figure 1.* We train 2-layer Transformer models on Erdős-Rényi graphs. (**Top**) Visualization of a graph, its input adjacency matrix, and the model's (mis-)prediction of its connectivity. (**Bottom**) Although trained models are able to predict perfectly on every edge within distribution, they failed to generalize to out-of-distribution graphs such as graphs with two isolated chains or cliques.

## 2 Related Work

Theoretical analyses aim to define what Transformers can and cannot compute. Although Transformers are universal approximators for continuous sequence-to-sequence functions (Yun et al., 2020), they also face sharp complexity-theoretic limits. Fixed-depth attention struggles with periodic or hierarchical patterns (Hahn, 2020), and standard Transformers are restricted to the complexity class $\mathsf{TC}^0$ (Merrill & Sabharwal, 2023), with hard-attention variants also confined to low-level circuit classes (Hao et al., 2022; Barcelo et al., 2024). Allowing model depth to scale logarithmically with input length enables solving graph connectivity (Merrill & Sabharwal, 2025; Sanford et al., 2024). Chain-of-thought also increases expressivity (Merrill & Sabharwal, 2024; Feng et al., 2023) but we exclude it here to focus only on what the base architecture learns through gradient descent. Programmatic abstractions like RASP offer another lens, identifying which algorithms can be implemented in a length-generalizing way (Weiss et al., 2021; Zhou et al., 2024a). Empirically for the graph connectivity problem, Fu et al. (2024b) shows frontier LLMs can reach almost perfect performance on small graphs and Saparov et al. (2025) shows Transformers have greater difficulty in learning the task when graph size increases. Our matrix powering view is analogous to message passing in GNNs, where $L$ layers reach $L$-hop neighborhoods (Hamilton, 2020); attention achieves $3^L$ hops by tripling the range at each layer. We include more related work in the appendix §E.

# 3 Problem Setup and Preliminary Study

## 3.1 Graph Connectivity Task

**Definition 3.1** (Self-loop-augmented adjacency matrix)**.** Let $G = (V, E)$ be a graph with $n$ vertices. We define the **self-loop-augmented adjacency matrix** $A \in \{0,1\}^{n\times n}$ as $A_{i,j} = 1$ if $\{v_i, v_j\} \in E$ or $i = j$, and 0 otherwise.

This definition is equivalent to taking the standard adjacency matrix and adding the identity matrix. A key consequence is that the $(i, j)$-th entry of the matrix power $A^k$ counts the number of walks of length $k$ from $v_i$ to $v_j$. With self-loops, these walks may stay at the same vertex from one step to the next. Henceforth, "adjacency matrix" will refer to this self-loop-augmented version.

**Definition 3.2** (Connectivity)**.** For any graph $G = (V, E)$ with $n$ nodes, we define the connectivity matrix $R \in \{0,1\}^{n\times n}$ as follows: $R_{i,j} = 1$ if there is a path between $v_i$ and $v_j$ and 0 otherwise. In particular, $R_{i,j} = 1$ if and only if $[A^n]_{i,j} > 0$.

Our learning objective is to learn models $\mathcal{M} : \{0,1\}^{n\times n} \to \mathbb{R}^{n\times n}$. For a graph $G$ with adjacency matrix $A$ and connectivity matrix $R$, if $\mathcal{M}$ satisfies $[\mathcal{M}(A)]_{i,j} > 0 \Leftrightarrow R_{i,j} = 1$, then we say $\mathcal{M}$ is *perfect* on graph $G$. We train on Erdős-Rényi $\mathsf{ER}(n, p)$ graphs, i.e., graphs on $n$ vertices where each edge is present independently with probability $p$.

## 3.2 Transformer Architectures

We first introduce our setups on standard transformer models without causal attention masking.

**Definition 3.3** (Transformers for Graph Connectivity)**.** Let $A$ be the self-loop-augmented adjacency matrix of $G$. Fix depth $L$ and hidden width $d > n$. Define the linear read-in and read-out maps

$$\mathsf{ReadIn}(X) := XW_{\text{in}}, \quad W_{\text{in}} \in \mathbb{R}^{n\times d},$$
$$\mathsf{ReadOut}(H) := HW_{\text{out}}^{\top}, \quad W_{\text{out}} \in \mathbb{R}^{n\times d}.$$

An $L$-layer single-head transformer model acts as

$$\mathsf{TF}_{\Theta}^{L}(A) \;:=\; \mathsf{ReadOut}\Big(\mathsf{Transformers}^{L}\big(\mathsf{ReadIn}(A)\big)\Big)$$

where $\mathsf{Transformers}^{L}$ is a standard pre-norm Transformer with self-attention and with *no* causal attention masks. There is no additional positional encoding since $I_n$ is already added to the input $A$ as the absolute positional encoding. A full definition can be found in Definition A.1.

## 3.3 Preliminary Study

We train 2-layer Transformer models on $\mathsf{ER}(n = 20, p = 0.08)$ graphs and test them on two out-of-distribution datasets: (1) $\mathsf{2Chain}(n = 20, k = 10)$ graphs with $n$ nodes consisting of two isolated chains each with $k$ nodes, and (2) $\mathsf{2Clique}(n = 20, k = 10)$ graphs with $n$ nodes consisting of two isolated $k$-Cliques. We measure the performance of model $\mathcal{M}$ via an exact match accuracy on our graph distribution $\mathcal{G}$, i.e., the fraction of graphs on which $\mathcal{M}$ is perfect, or $\mathsf{ExactMatchAcc}(\mathcal{M}, \mathcal{G}) = \mathbb{E}_{G=(V,E)\in\mathcal{G}}\left[\prod_{v_i,v_j\in V} \mathbf{1}\left\{[\mathcal{M}(A_G)]_{i,j} = [R_G]_{i,j}\right\}\right]$.

**Transformers Fail to Generalize.** As shown in Figure 1, the 2-layer Transformer model is able to achieve almost perfect exact match accuracy on the held-out set of the training distribution. However, it fails to learn an algorithmic solution that transfers to other distributions. When the model is tested on the 2Chain and 2Clique distributions, its exact match accuracy falls to nearly zero, indicating over-fitted heuristics have dominated the model prediction. We repeat the experiments via extensive hyperparameter search and scaling up the number of layers, but all models fail to generalize. This motivates us to investigate why transformers prefer to learn brittle heuristics and how we can encourage them to learn algorithmic solutions instead.

# 4 Theory

## 4.1 Disentangled Transformer

To understand the generalization failure in §3.3 theoretically, we pivot to a simplified ***Disentangled Transformer***; this helps us not only with expressivity/capacity analysis in §4.2 but also with training dynamics analysis in §4.3. In the Disentangled Transformer, each attention block appends its output as a new coordinate slice of the residual stream rather than summing, so the representation dimension grows with depth and the read/write pathways become traceable (Friedman et al., 2023; Nichani et al., 2024). This model serves as a reasonable proxy for its standard counterpart: Nichani et al. (2024) show that any standard attention-only Transformer can be re-expressed as a disentangled model by specializing attention to implement feature concatenation, and Chen et al. (2024) adopt this architecture precisely because it preserves the computations of interest while being markedly more amenable to theoretical analysis. We now formalize the model.

**Definition 4.1** (Disentangled Transformer for Graphs)**.** Let $n$ be the number of nodes for any graph $G$ with adjacency matrix $A \in \mathbb{R}^{n\times n}$. Let $L$ be the depth of the Disentangled Transformer, and $\{d_0, d_1, \cdots, d_L\}$ be the set of dimensions of its hidden states with $d_\ell = 2^{\ell+1}n$. Let $\{W_\ell\}_{\ell=1}^{L}$ be the attention matrices with $W_\ell \in \mathbb{R}^{d_{\ell-1}\times d_{\ell-1}}$. Let $W_O \in \mathbb{R}^{n\times d_L} = [I_n, \cdots, I_n]$ be the output matrix. Let $\Theta = \{W_\ell\}_{\ell=1}^{L}$. An $L$-layer Disentangled Transformer $\mathsf{TF}_{\Theta}^{L}$ acts

on any graph's self-loop augmented adjacency matrix $A$ by

**Input hidden state** $h_0 := [I_n, A] \in \mathbb{R}^{n \times d_0}$

**$\ell$-th Hidden state** $h_\ell := [h_{\ell-1}, \mathrm{Attn}\,(h_{\ell-1}; W_\ell)] \in \mathbb{R}^{n \times d_\ell}$

**Output layer** $\mathsf{TF}^L_\Theta(A) := h_L W_O^\top$

where $\mathrm{Attn}\,(h_{\ell-1}; W_\ell) := \frac{1}{n}\mathsf{ReLU}\left(h_{\ell-1} W_\ell h_{\ell-1}^\top\right) h_{\ell-1}$. We remark that $h_\ell \in \mathbb{R}^{n \times d_\ell}$ where $d_\ell = 2^{\ell+1} n$ grows exponentially with respect to $\ell$.

### 4.2 Expressivity and Capacity

If a 2-layer Transformer fails to generalize in §3.3, should we attribute this to the architecture's expressivity? We argue not. Theorem 4.3 shows that an $L$-layer Disentangled Transformer can implement the matrix powering algorithm and is perfect on graphs of diameter at most $3^L$. Moreover, Theorem 4.5 shows this $3^L$ threshold is tight and exact[1]. To make this precise, we first formalize graph distance and diameter in Definition 4.2.

**Definition 4.2** (Graph distances and diameter)**.** Let $G = (V, E)$ be a finite, simple, undirected graph. Following standard definitions, for $u, v \in V$, we let $d_G(u, v)$ be the *shortest-path distance* between $u, v$, which is finite if they are connected and infinite otherwise. For a connected component, we define its *diameter* to be the longest path length within the component.

Throughout, we define the diameter of a graph, denoted $\mathrm{diam}(G)$, to be the maximum diameter among its connected components. Note this differs from the common convention on disconnected graphs, where the latter sets $\max_{u,v} d_G(u, v) = \infty$; ours is always finite.

We begin by establishing the expressive power of the Disentangled Transformer, showing that with sufficient depth, it can implement the correct matrix powering algorithm to solve connectivity.

**Theorem 4.3** (Expressivity of $\mathsf{TF}^L_\Theta$)**.** *There exists an $L$-layer Disentangled Transformer that makes perfect predictions for every graph $G$ satisfying $\mathrm{diam}(G) \le 3^L$.*

*Sketch of proof.* For all $\ell$, setting $W_\ell = I_{d_{\ell-1}}$ suffices. These choices of weights implement the matrix powering algorithm $\sum_{j=0}^{\mathrm{diam}(G)} \alpha_j A^j$ with positive coefficients $\alpha_j$. □

We next show a capacity bound that reveals the model's inherent limitations. We prove a tight, non-asymptotic upper bound on the graph diameter an $L$-layer model can handle, linking model depth directly to instance difficulty. To state this precisely, we define capacity as follows.

[1]Because the $\mathsf{2Chain}(n = 20, k = 10)$ graphs have maximum path length 9, they should be theoretically learnable by 2-layer models, unlike in §3.3.

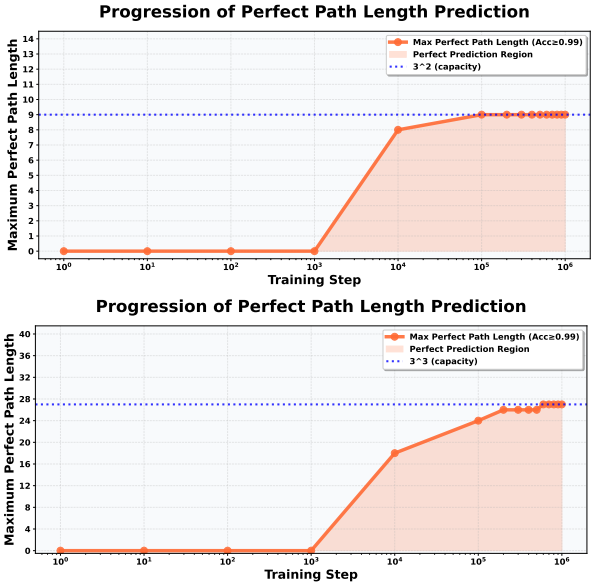



*Figure 2.* **Capacity of Disentangled Transformers.** We train 2-layer (**top**) and 3-layer (**bottom**) Disentangled Transformers on $\mathsf{ER}(n = 24)$ and $\mathsf{ER}(n = 64)$ graphs respectively. When evaluated on hold-out sets, both models can only make reliable predictions ($\ge 99\%$ accuracy) on node pairs $u, v$ if and only if $d_G(u, v) \le 3^L$. These findings resonate with our theoretical observations in Theorem 4.5.

**Definition 4.4** (Model Capacity)**.** The *capacity* of an $L$-layer Disentangled Transformer $\mathsf{TF}^L_\Theta$ is the largest integer $d$ such that there exist weights achieving perfect predictions on every graph $G$ with $\mathrm{diam}(G) \le d$.

Informally speaking, the capacity of an $L$-layer Disentangled Transformer $\mathsf{TF}^L_\Theta$ with nonnegative weights is $3^L$. We formalize this claim as follows.

**Theorem 4.5** (Capacity of $\mathsf{TF}^L_\Theta$)**.** *Let $\mathsf{TF}^L_\Theta$ be an $L$-layer Disentangled Transformer on $n = \Omega(3^L)$ nodes[2]. Further assume that the weights $W_\ell \ge 0$ for each $\ell$. Then there exists a graph $G$ on which $\mathsf{TF}^L_\Theta(A)$ does not equal the connectivity matrix $R$. Further, $G$ has diameter $3^L + 1$. In other words, $\mathsf{TF}^L_\Theta$ cannot master graph connectivity beyond path length $3^L$, and they cannot fully solve the task on $\Omega(3^L)$-node graphs.*

*Sketch of proof.* We split by whether a false positive across different connected components occurs at some intermediate layer; the full proof can be found in Section B.2.

*Case 1 (False positive occurs Lemma B.1).* Suppose for a graph $H$, an intermediate layer contains a false positive in its hidden states. That is, for two disconnected nodes $u, v$, the corresponding $(u, v)$ entry in some $\ell^{\text{th}}$ layer hidden states is positive (and this will propagate to the final output, whereas the $(u, v)$ entry in the ground-truth connectivity entry isn't). We isolate two sets of nodes that contribute to

[2]In particular, taking $n \ge (7/3) \cdot 3^L + 2$ suffices.

this false positive by backtracking the computation DAG. By making appropriate changes, we argue for the existence of a graph $G$ that (i) preserves this false positive on $(u, v)$ and (ii) contains a manually created path of length $3^L + 1$.

*Case 2 (No false positives Lemma B.2).* Suppose now that no intermediate layer has false positives for any $n$-node graph. We show that "information" spreads no faster than $3^L$ so that it never predicts "Yes" on node pairs with distance beyond $3^L$. We first apply the no-false-positives assumption on the empty (self-loops-only) graph. Inductively, each column of each hidden states is supported on exactly one row, which ranges from 1 to $n$. This naturally gives a "label" for each column in each hidden states. The crux of the proof is to inductively show that at layer $\ell$, two columns can "share" information, thereby creating a positive score on $(u, v)$, only if their labels, interpreted as graph nodes, are within distance $3^\ell$. Consequently, a no-false-positive model cannot recognize a connected pair with distance $3^L + 1$.

In both cases one can construct graphs with diameters $3^L+1$ on which $\mathsf{TF}^L_\Theta$ is not perfect. □

Given the tight $3^L$ capacity bounds for Transformers, it is natural and crucial to introduce a dichotomy around the $3^L$ capacity. For any connected node pair $(u, v)$, they are said to be within capacity if $d_G(u, v) \le 3^L$ and beyond capacity otherwise. Formally, we define the dichotomy as follow:

**Definition 4.6** (Within-capacity and beyond-capacity pairs at depth $L$)**.** Fix a graph $G$ and a depth $L$. We say a pair of nodes $(i, j)$ is **within capacity** if $[A^{3^L}]_{i,j} > 0$ and **beyond capacity** otherwise. In other words, a pair $(i, j)$ is within capacity iff their shortest-path distance is $\le 3^L$.

Overloading the notation, we say $G$ is a within-capacity graph if $\mathrm{diam}(G) \le 3^L$ and beyond-capacity otherwise.

### 4.3 Training Dynamics

If a capable 2-layer Transformer is able to perfectly predict connectivity up to path length $3^2 = 9$, and the $\mathsf{2Chain}(n = 20, k = 10)$ dataset does not contain longer paths, why didn't the Transformer model in §3.3 learn the algorithm? In this section, we show that this is because the training distribution contains too many graphs beyond the $3^L$ capacity, and those samples reward a learning a global heuristic over an algorithm. Equipped with Theorem 4.7, we can analyze the gradient dynamics in the two-channel parameterization (a superposition of heuristic and algorithmic channels). Theorem C.5 and Theorem C.9 give a simple criterion made possible by the exact $3^L$ characterization: if within-capacity pairs dominate, the algorithmic channel wins; if beyond-capacity pairs prevail, the heuristic wins.

**Parameterizing Model Weights.** To analyze the gradient dynamics, we first identify the relevant parameter space.

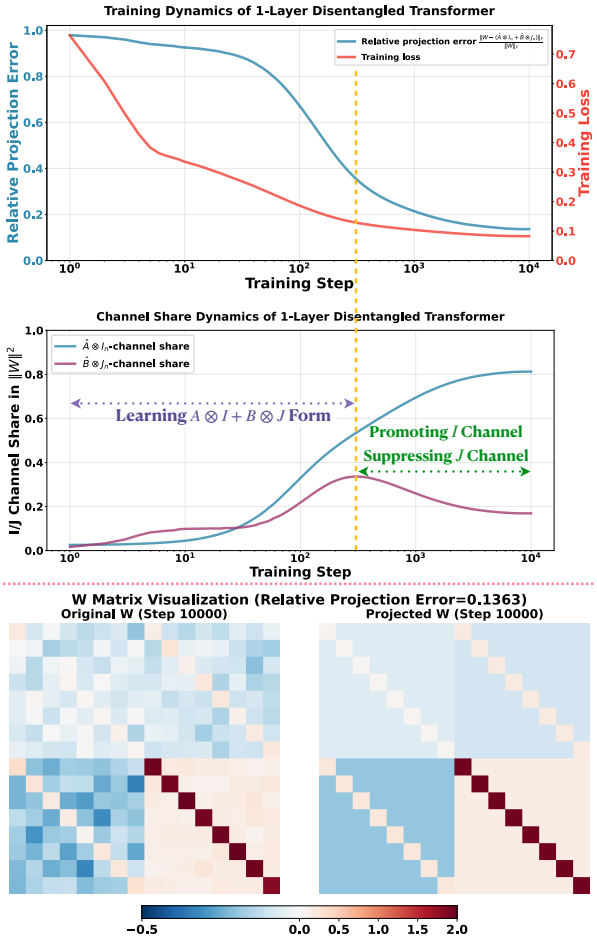



*Figure 3.* **Training Dynamics of Disentangled Transformers.** We train a 1-layer Disentangled Transformer on graphs from $\mathsf{ER}(n = 8, p = 0.2)$ distribution. Weight $W$ will approximately approach to $A \otimes I_n + B \otimes J_n$ form. (**Top**) There are two major phases during training, where during Phase 1, model focuses on learning the equivariant parameterizations so both $I$ and $J$ channel's share of energy (see §5.1) in $W$ increases, and during Phase 2, the algorithmic $I$-channel is promoted and the heuristic $J$-channel is suppressed. (**Bottom**) Visualization of the learned weights and its projection to the closest $\hat{W} = \hat{A} \otimes I_n + \hat{B} \otimes J_n$ form.

Our data and targets are symmetric under node relabeling: the ground truth mapping maps $A \mapsto R$ and $PAP^\top \mapsto PRP^\top$ for any permutation $P$. We also empirically observe that models trained from scratch on these graphs rapidly converge to a layerwise equivariant state (Figure 8). Based on these observations, we analyze the dynamics within the subspace of layerwise permutation-equivariant weights. The following theorem formally defines layerwise equivariance and characterizes exactly what weights look like.

**Theorem 4.7** (Layerwise Permutation-Equivariant Parameterization)**.** *Suppose an $L$-layer Disentangled Transformer $\mathsf{TF}^L_\Theta$ has non-negative weights. Let $K_{\ell-1} = 2^\ell$. Then $\mathsf{TF}^L_\Theta$ is layer-wise permutation equivariant, i.e., for each $\ell$, any*

*hidden states $h \in \mathbb{R}^{n \times d_{\ell-1}}$, and any permutation $P \in S_n$,*

$$\mathrm{Attn}(Ph(I_{K_{\ell-1}} \otimes P^\top); W_\ell) = P\ \mathrm{Attn}(h; W_\ell)\ (I_{K_{\ell-1}} \otimes P^\top),$$

*if and only if each layer weight $W_\ell$ decomposes as $W_\ell = A_\ell \otimes I_n + B_\ell \otimes J_n$ for some $A_\ell, B_\ell \in \mathbb{R}^{2^\ell \times 2^\ell}$ for all $\ell$, where $\otimes$ denotes the Kronecker product and $J_n = \mathbf{1}\mathbf{1}^T$ the all-ones $(n \times n)$ matrix.*

*Sketch of proof.* Sufficiency is immediate: If $W_\ell$ admits this form, then conjugating by any node permutation $P$ leaves both factors invariant, as $PI_nP^\top = I_n$ and $PJ_nP^\top = J_n$.

For necessity, the key observation is that with ReLU inactive due to non-negativity assumption, the layer map becomes bilinear in $h$. With algebra, the equivariance assumption can be shown to imply a conjugation-alike identity on *weights*: Writing $\sigma(P) = I_{K_{\ell-1}} \otimes P$ and $\Delta = \sigma(P)W_\ell\sigma(P)^\top - W_\ell$, the following must hold:

$$h\Delta h^\top h\sigma(P) = 0 \qquad \text{for all } h \geq 0.$$

We then argue that this forces $\Delta = 0$, i.e., $W_\ell$ is conjugation-invariant, by testing the above equation with special matrices $P$. Next, we inspect small blocks $W_\ell[u, v]$ of size $n \times n$ in $W_\ell \in \mathbb{R}^{(2^\ell n)\times(2^\ell n)}$, and argue that $W_\ell[u, v]$ must commute with all permutations $P$. This forces each $W_\ell[u, v]$ to lie in $\mathrm{span}(I_n, J_n)$. Aggregating all subblocks, $W_\ell$ can therefore be decomposed as $A_\ell \otimes I_n + B_\ell \otimes J_n$. □

It immediately follows that this subspace contains a canonical algorithmic solution (e.g. the identity construction $W = I$ used in Theorem 4.3). Furthermore, Theorem C.2 (in Appendix) shows that for *any* capacity-reaching model, the symmetric component of the weights, which drives the attention mechanism, must lie purely in the $I_n$-channel. Finally, algebraically this parameterization is closed under gradients (Theorem C.4), and this allows us to decompose the learning process into the competition between two functionally distinct channels, discussed next.

Under the conditions of Theorem 4.7, each layer weight decomposes as $W_\ell = A_\ell \otimes I_n + B_\ell \otimes J_n$. This decomposition separates the model's computation into two functionally distinct channels with different roles.

**The $I_n$-channel ($A_\ell \otimes I_n$) compute algorithms.** The term $A_\ell \otimes I_n$ preserves locality: when applied within the attention mechanism, it only combines features between nodes that share graph neighbors. Across layers, this channel composes multi-hop information by effectively computing powers of the adjacency matrix $A$. After $L$ layers, the readout aggregates a weighted sum $\sum_{j=0}^{3^L} \alpha_j A^j$ with non-negative coefficients, which is precisely the matrix powering algorithm that solves connectivity (Theorem 4.3).

**The $J_n$-channel ($B_\ell \otimes J_n$) collect heuristics.** The term $B_\ell \otimes J_n$ broadcasts information globally, ignoring graph structure. To see why, observe that $J_n$ is rank-one: for any vector $x \in \mathbb{R}^n$, we have $J_n x = (\mathbf{1}^\top x)\mathbf{1}$, which computes the sum of entries and broadcasts it uniformly to all nodes. When composed with the adjacency matrix, $AJ_n = \mathbf{d}\mathbf{1}^\top$ where $\mathbf{d} = A\mathbf{1}$ is the degree vector. Thus, this channel computes global statistics, specifically products of node degrees and their higher-order generalizations. Such statistics correlate with connectivity on dense random graphs but fail on adversarial instances. For example, two disjoint cliques both have high-degree nodes, yet no edge connects them.

**Training Dynamics.** Under this algorithmic-herustic dual-channel view, we can track the evolution of the two parameters, $A_\ell$ and $B_\ell$. To rigorously analyze the gradient descents, we adopt the following assumptions.

**Assumption 4.8.** We make assumptions on data distribution, model parameterization and the training loss.

1. **Data Distribution.** Let $\mathsf{ER}(n, p)$ be the Erdős-Rényi distribution with edge-probability $p \in (0, 1)$. Then $\mathbb{P}_{G\sim\mathsf{ER}(n,p)}\{G \text{ is disconnected}\} > 0$.
2. **Nonnegativity & Equivariant Parameterization.** For each layer $\ell$, assume $W_\ell \geq 0$ can be decomposed as $W_\ell = A_\ell \otimes I_n + B_\ell \otimes J_n$ for some $A_\ell, B_\ell \in \mathbb{R}^{2^\ell \times 2^\ell}$.
3. **Surrogate Loss.** Given scores $Z := \mathsf{TF}^L_\Theta(\cdot) \in \mathbb{R}^{n\times n}_{\geq 0}$, define the link $\phi(z) := 1 - e^{-\alpha z}$ with $\alpha > 0$;[3] the entrywise cross-entropy with respect to the connectivity matrix $R$ is $\mathcal{L}(Z; R) := -\sum_{i,j}\big(R_{i,j}\log\phi(Z_{i,j}) + (1 - R_{i,j})\log(1 - \phi(Z_{i,j}))\big)$. Its gradient with respect to $Z$ is $\frac{\partial\mathcal{L}}{\partial Z} = \alpha\,(1 - R/\phi(Z)) \in \mathbb{R}^{n\times n}$.

Under these assumptions, we can characterize the convergence and limiting behavior of gradient descent.

**Theorem 4.9** (Convergence to KKT Points)**.** *Let $\mathcal{R}(\Theta) := \mathbb{E}_{G\sim\mathsf{ER}(n,p)}[\mathcal{L}(\mathsf{TF}^L_\Theta(A_G); R_G)]$ denote the population risk. For $\lambda > 0$, define the regularized objective $\mathcal{R}_\lambda(\Theta) := \mathcal{R}(\Theta) + \frac{\lambda}{2}\|\Theta\|_F^2$. Let $\mathcal{C} := \{(A_\ell, B_\ell)_\ell : A_\ell \geq 0, B_\ell \geq 0, \forall\ell\}$ denote the constraint set, and consider the sequence $\{\Theta^{(k)}\}_{k\geq 0}$ generated by projected gradient descent on $\mathcal{R}_\lambda$:*

$$\Theta^{(k+1)} = \Pi_{\mathcal{C}}\left(\Theta^{(k)} - \eta\nabla\mathcal{R}_\lambda(\Theta^{(k)})\right), \tag{1}$$

*with step size $\eta > 0$ sufficiently small and initialization $\Theta^{(0)} \in \mathcal{C}$ of the form $W_\ell = A_\ell \otimes I + B_\ell \otimes J$. Then every limit point $\Theta^*_\lambda \in \mathcal{C}$ satisfies the KKT conditions:*

$$\begin{aligned} \nabla_{B_\ell}\mathcal{R}(\Theta^*_\lambda) + \lambda B^*_\ell \geq 0, \quad B^*_\ell \geq 0, \\ (\nabla_{B_\ell}\mathcal{R}(\Theta^*_\lambda) + \lambda B^*_\ell) \odot B^*_\ell = 0, \end{aligned} \tag{2}$$

[3] It is possible that $R_{i,j} = 1$ while $Z_{i,j} = 0$, resulting in undefined gradient $\partial\mathcal{L}/\partial Z$. To circumvent this, we approximate via $\phi_\epsilon = 1 - (1 - \epsilon)e^{-\alpha z}$. All subsequent analyses hold verbatim by replacing $\phi$ with $\phi_\epsilon$.

*and analogously for $A_\ell^*$. Moreover, the iterates converge to a KKT point at the standard $\mathcal{O}(1/\epsilon)$ rate for projected gradient descent.*

The proof adapts standard convergence analysis for projected gradient descent on smooth nonconvex functions (Bertsekas, 1997; Beck, 2017). See Section C.5 for details. In the next result, we analyze the structure of the KKT points of the objective.

**Theorem 4.10** (Learning the Algorithm, informal version of Theorem C.5 and Corollary C.20)**.** *Under suitable conditions where the gradient penalty from disconnected graphs outweighs the gradient reward from connected graphs, the only KKT-compliant value for the heuristic channel is $B_\ell^* = 0$, i.e., the model converges to learning the fully algorithmic matrix-powering algorithm.*

While Theorem 4.9 guarantees that training converges to limit points $\Theta_\lambda^*$ of the objective $\mathcal{R}_\lambda(\Theta)$, Theorem 4.10 further characterizes a sufficient condition for algorithmic alignment. The formal versions for the un-regularized and regularized objectives can be found at Theorem C.5 and Corollary C.20, respectively. With convergence guaranteed, our analysis of gradient descents reveals that the training process consists of two distinct phases:

**Phase 1: Both channels pick up easy examples**. In early updates, both channels quickly ramp up mass because there are plenty of within-component, within-capacity pairs. Concretely, the local $I$-channel composes neighborhood information, while the global $J$-channel can also boost under-predicted positives without facing much penalty (Remark C.7). Phase 1 is transient and ends once those easily connected pairs are mostly saturated. In Figure 3 (top), it only occupies around $2 \cdot 10^2$ steps out of $10^4$ total.

**Phase 2: Data determines which channel wins**. Once in this regime, the growth of $B_\ell$ is determined by the population-level balance (see Theorem C.5 for full definitions and details). Informally speaking, the derivative of $B_\ell$ is driven by a competition between a suppression force and a promotion force. The suppression force arises from disconnected graphs, where the heuristic generates false positives, and they push $B_\ell$ to zero. The promotion force arises from connected graphs, where the heuristic helps minimize loss by correctly identifying connected pairs, generating a reward gradient.

There are two outcomes, best understood through the distinction between within-capacity and beyond-capacity graphs (Theorem C.9). If batches are dominated by within-capacity graphs ($\mathrm{diam}(G) \le 3^L$), the algorithmic $I$-channel is consistently rewarded. Disconnected graphs in this regime penalize the heuristic $J$-channel for predicting false connections, driving $B_\ell \to 0$ (Theorem C.5(ii)) and leaving only the algorithmic solution. Conversely, if the distribution contains a significant share of connected beyond-capacity graphs ($\mathrm{diam}(G) > 3^L$), the algorithmic channel cannot bridge the distance. These samples force the model to rely on the global $J$-channel to minimize loss, promoting the degree-counting shortcut. Thus, the final learned solution depends directly on the proportion of beyond-capacity connected graphs in the training distribution (Remark C.13), as visualized in Figure 6.

## 5 Experiments

We test the theory in two parts. First, in §5.1 we verify the $3^L$ capacity threshold (Theorems 4.3 and 4.5) by measuring the maximum reliable path length one model can handle perfectly. Then, we trace training dynamics by projecting learned weights onto the algorithmic $A \otimes I$ channel and the heuristic $B \otimes J$ channel (Theorem 4.7). Next, in §5.2 we introduce a simple data lever that up-weights within-capacity graphs, and show that this simple method suppresses the heuristic and promotes the algorithmic channel, as predicted by Theorems C.5 and C.9. Finally, we show this data lever prescribed by our theoretical analysis on Disentangled Transformers can transfer back to standard Transformers and boost their generalization capabilities.

### 5.1 Capacity and Training Dynamics

**$L$-layer Transformers Hit Their Capacity at Exactly $3^L$.** We train Disentangled Transformers with 2 layers or 3 layers on Erdős-Rényi graphs with 24 or 64 nodes respectively. As shown in Figure 2, neither of the two models could make reliable predictions on node pairs $(u, v)$ with $d_G(u, v) > 3^L$ but their predictions on node pair with $d_G(u, v) \le 3^L$ are almost perfect with an $> 99\%$ accuracy. As shown in Figure 10, a 2-layer standard Transformer model also has the same empirical capacity. These results resonate with our exact capacity bound of Disentangled Transformers in Theorem 4.5 and justifies our dichotomy in Definition 4.6. Overall, for any graph $G = (V, E)$, the decisive factor for Transformer model depth is not simply the asymptotic $\Theta(\log |V|)$ relation to the number of nodes $n = |V|$ but more importantly the non-asymptotically exact relation to $\log_3 \mathrm{diam}(G)$. We also note that we do not enforce non-negativity of model weights during training, but our theoretical analysis remains predictive of model capacity.

**Transformers Learn an Algorithm-Heuristic Mixture.**To empirically understand training dynamics, we first train a 1-layer Disentangled Transformer model on $\mathsf{ER}(n = 8, p = 0.2)$ graphs, without enforcing any parameterization assumptions. As shown in Figure 3, a randomly initialized $W$ converges to a matrix approximately of the form $A \otimes I_n + B \otimes J_n$ for some matrices $A, B \in \mathbb{R}^{2\times 2}$. Deeper models also converge to such solutions, as shown in Figure 9. These results show the applicability of our decomposition Theorem 4.7. Then, we project the final weight $W$ onto

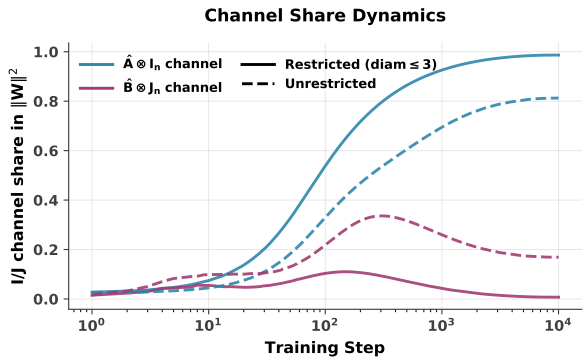



*Figure 4.* Following insights from Theorems C.5 and C.9, we repeat the same experiment setup as in Figure 3 but only training on within-capacity graphs (see Definition 4.6). As shown in the **solid** lines, restricting training samples by capacity pushes the energy share of the **algorithmic** mechanism (the $A \otimes I_n$ channel) further to nearly 100% in the weight $W$. It simultaneously prevents the growth of the **heuristic** portion (the $B \otimes J_n$ channel).

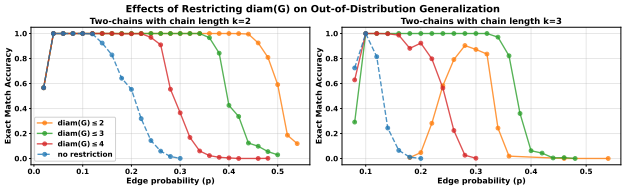



*Figure 5.* With 1-layer Disentangled Transformers with capacity $\mathsf{Cap} = 3$ following Theorem 4.5, we vary $d$ such that we restrict our training graphs to have $\mathrm{diam}(G) \leq d$. We also vary the edge probability of our training distribution $\mathsf{ER}(n = 8, p = \cdot)$ for generality. We test on $\mathsf{2Chain}(n = 8, k = \cdot)$ graphs with $k = 2$ or 3 and show the exact match accuracy on configurations where the accuracy is non-zero for readability. We find if the training $d \leq \mathsf{Cap}$, models still learns the algorithmic solution up to problem size $d$ (see $d = 2, k = 2$ case on the left in orange) but *fails to length generalize* (see $d = 2, k = 3$ in orange on the right). On the other hand, if the training $d > \mathsf{Cap}$, model struggles to learn the algorithmic solution (see $d = 4$ cases in red on both $k = 2$ or 3). The best case overall is when setting $d = \mathsf{Cap}$, i.e., preventing the model from seeing beyond-capacity samples but still preserving at-capacity samples for better generalization. As shown in the green lines, with $d = 3$, model achieves balanced testing accuracy on both $k = 2$ and 3.

this algebra as $\hat{W} = \hat{A} \otimes I_n + \hat{B} \otimes J_n$ by minimizing $\|W - \hat{W}\|_F$. We observe that the share (see §D) of $\hat{A} \otimes I_n$ in $\|W\|_F^2$ increases as training progresses but the share of $\hat{B} \otimes J_n$ first increases and then decreases. These provide empirical evidence supporting our two-phase story in §4.3.

### 5.2 Encouraging Transformers to Learn Algorithms

Now that we understand *why* Transformers and Disentangled Transformer models learn heuristics that hurt their algorithmic computations (as shown in Figures 1 and 3), a natural question is whether we can mitigate this unwanted behavior and encourage the models to up-weight the algorithmic channel.

**Mitigation via the Data Lever.**We propose a data-centric method: instead of training on all graphs from the ER distribution, we up-weigh graphs whose $\mathrm{diam}(G)$ is within capacity following the dichotomy in Definition 4.6. We dissect the training distribution $\mathcal{G}$ into two sub-distributions: $\mathcal{G}_{\leq} = \{G \in \mathcal{G} : \mathrm{diam}(G) \leq 3^L\}$ only includes graphs containing no beyond capacity node pairs, and $\mathcal{G}_{>}$ includes the rest. In Figure 4, we only train the 1-layer Disentangled Transformer on $\mathsf{ER}_{\leq}$, and then find the algorithmic $\hat{A} \otimes I_n$ channel is significantly promoted so that the learned weight only contains the algorithm channel. Furthermore, we find at-capacity graphs are crucial. In the case of Figures 5 and 12, where no graphs are to have $\mathrm{diam}(G) > 2$, the model also fails to learn generalizable solutions due to Transformers' poor length generalization abilities. These results imply that simply scaling up the model depth does not naturally equip it with algorithmic capabilities .

**Robustness to Noise.**To test the predictiveness of our theory, we evaluate if a small amount of beyond-capacity node pairs is enough to encourage the model learning a heuristic-dominated method. We define $\rho(\mathcal{G}) = \mathbb{E}_{G \in \mathcal{G}} |\{(u, v) \in V, d_G(u, v) > 3^L\}|/n^2$ to be the fraction of beyond-capacity node pairs in a graph distribution $\mathcal{G}$. In practice, $\rho$ can be controlled via stratified sampling from the mixture distribution $\mathcal{G}_q = q\mathcal{G}_{\leq} + (1 - q)\mathcal{G}_{>}$. In Figure 6, we performed stratified sampling between $\mathsf{ER}_{\leq}$ and $\mathsf{ER}_{>}$ and found that with a small $\rho(\mathcal{G})$, the model is still able to maintain high energy in the algorithm channel and make perfect predictions on out-of-distribution 2Chain graphs. It suggests that there exists a small $\rho^\star > 0$ such that the model can still rely on the algorithm-channel to make predictions if the training distribution satisfies $\rho(\mathcal{G}) \leq \rho^\star$.

**Transferability to Standard Transformers models.**Our theory from §4.3 makes a prescriptive suggestion to remove beyond capacity graphs to reduce dependence on heuristics, and Figure 4 demonstrated the effectiveness of this approach on Disentangled Transformers. We now evaluate this on standard Transformers. We train the same 2-layer Transformers model as in our preliminary study in §3.3 but this time, we train on the restricted distribution $\mathsf{ER}_{\leq}$ instead where all graphs $G \in \mathsf{ER}_{\leq}$ have $\mathrm{diam}(G) \leq 3^2 = 9$. As shown in Figure 7, when tested on the OOD 2Chain dataset with maximum chain length 10, the one trained on $\mathsf{ER}_{\leq}$ can successfully generalize but the one trained on unconstrained distribution ER cannot. It also helps model generalize to OOD 2Clique graphs as well (see Figure 11).

## 6 Discussion and Conclusion

In this paper, we separate expressivity from capacity and training dynamics for Transformers on graph connectivity. We prove that an $L$-layer model can implement matrix

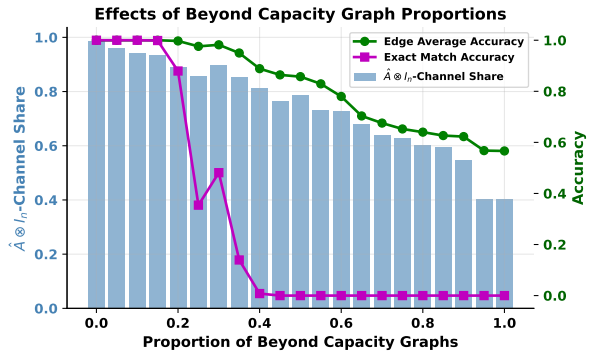



*Figure 6.* We vary the proportion of beyond-capacity graphs, and train the same Disentangled Transformer on stratified ER distribution and test on the same OOD 2Chain distribution. We find that Transformers are robust towards a small amount of noises (beyond-capacity graphs). Although the $W$ is not exactly in the $A \otimes I_n$ form, the model still performs perfectly when the energy share of $I$-channel dominates (beyond roughly 90%).

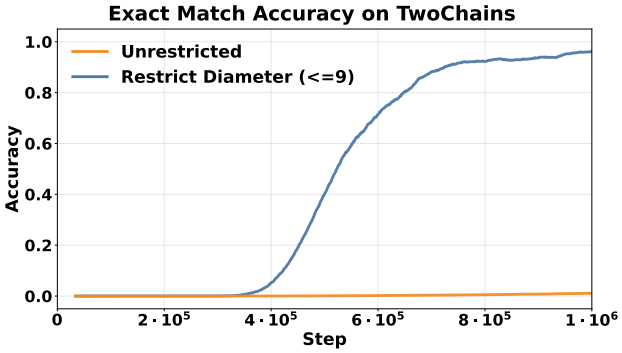



*Figure 7.* **Standard Transformer models learn generalizable solutions from within capacity data.** We train a 2-layer standard Transformer model on $\mathsf{ER}(n = 20)$ graphs, with and without restricting graph diameters. When tested on OOD 2Chain graphs, the one trained with *the right data* is able to generalize.

powering and is perfect on graphs with $\mathrm{diam}(G) \leq 3^L$, and we show this $3^L$ threshold is tight. The failures in §3.3 are explained by a capacity mismatch: training mass beyond $3^L$ steers learning toward a global shortcut rather than the intended multi-hop algorithmic computation. Our two-channel view makes this explicit and turns generalization into a property of the data distribution: when within-capacity pairs dominate, the algorithmic channel is selected. Experiments confirm the threshold and show that a simple capacity-aware data lever that up-weights within-capacity graphs suppresses the shortcut, promotes out-of-distribution generalization, and transfers to standard Transformers. By pinpointing when a model reaches for a shortcut and showing how simple data choices can steer it towards the true algorithmic solution, we outline a path to systematically control training data and model capacity to enable Transformers to learn solutions that generalize better.

## Impact Statement

This paper presents work whose goal is to advance the field of machine learning, specifically our understanding of when neural networks learn generalizable algorithms versus brittle heuristics. Our findings have potential implications for improving the reliability and interpretability of machine learning systems deployed in safety-critical applications, where algorithmic correctness is paramount. By identifying data distribution properties that encourage algorithmic learning, this work may inform better training practices. We do not foresee direct negative societal consequences from this foundational research.

## Acknowledgments

The authors acknowledge the Center for Advanced Research Computing (CARC) at the University of Southern California for providing computing resources that have contributed to the research results reported within this publication. We also acknowledge the use of the USC NLP cluster provided by USC NLP Group. This work used the Delta system at the National Center for Supercomputing Applications through allocation CIS250737 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. DF and RJ were also supported by gifts from the USC-Capital One Center for Responsible AI and Decision Making in Finance (CREDIF) and the USC-Amazon Center on Secure and Trusted Machine Learning. RJ was also supported by the National Science Foundation under Grant No. IIS-2403436. VS was supported by an NSF CAREER Award CCF-2239265, an Amazon Research Award, a Google Research Scholar Award and an Okawa Foundation Research Grant. The work was done in part while DF and VS were visiting the Simons Institute for the Theory of Computing. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the funding agencies.

## References

Attouch, H., Bolte, J., and Svaiter, B. F. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 137:91–129, 2013. URL https://api.semanticscholar.org/CorpusID:6597608.

Barcelo, P., Kozachinskiy, A., Lin, A. W., and Podolskii, V. Logical languages accepted by transformer encoders with hard attention. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gbrHZq07mq.

Beck, A. *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2017. doi: 10.1137/1.9781611974997. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611974997.

Bertsekas, D. P. Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334, 1997. doi: 10.1057/palgrave.jors.2600425. URL https://doi.org/10.1057/palgrave.jors.2600425.

Brinkmann, J., Sheshadri, A., Levoso, V., Swoboda, P., and Bartelt, C. A mechanistic analysis of a transformer trained on a symbolic multi-step reasoning task. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4082–4102, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.242. URL https://aclanthology.org/2024.findings-acl.242/.

Chan, L., Garriga-Alonso, A., Goldwosky-Dill, N., Greenblatt, R., Nitishinskaya, J., Radhakrishnan, A., Shlegeris, B., and Thomas, N. Causal scrubbing, a method for rigorously testing interpretability hypotheses. *AI Alignment Forum*, 2022. https://www.alignmentforum.org/posts/JvZhhzycHu2Yd57RN/causal-scrubbing-a-method-for-rigorously-testing.

Chang, T.-Y., Thomason, J., and Jia, R. Do localization methods actually localize memorized data in LLMs? a tale of two benchmarks. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3190–3211, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.176. URL https://aclanthology.org/2024.naacl-long.176/.

Chen, S., Sheen, H., Wang, T., and Yang, Z. Unveiling induction heads: Provable training dynamics and feature learning in transformers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=4fN2REs0Ma.

Deng, C., Zhao, Y., Tang, X., Gerstein, M., and Cohan, A. Investigating data contamination in modern benchmarks for large language models. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8706–8719, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.482. URL https://aclanthology.org/2024.naacl-long.482/.

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.

Feng, G., Zhang, B., Gu, Y., Ye, H., He, D., and Wang, L. Towards revealing the mystery behind chain of thought: A theoretical perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=qHrADgAdYu.

Floyd, R. W. Algorithm 97: Shortest path. *Commun. ACM*, 5(6):345, June 1962. ISSN 0001-0782. doi: 10.1145/367766.368168. URL https://doi.org/10.1145/367766.368168.

Friedman, D., Wettig, A., and Chen, D. Learning transformer programs. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Pe9WxkN8Ff.

Fu, D., Chen, T.-Q., Jia, R., and Sharan, V. Transformers learn to achieve second-order convergence rates for in-context linear regression. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a. URL https://openreview.net/forum?id=L8h6cozcbn.

Fu, D., Guo, R., Khalighinejad, G., Liu, O., Dhingra, B., Yogatama, D., Jia, R., and Neiswanger, W. Isobench: Benchmarking multimodal foundation models on isomorphic representations. In *First Conference on Language Modeling*, 2024b. URL https://openreview.net/forum?id=KZd1EErRJ1.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R. S., Brendel, W., Bethge, M., and Wichmann, F. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665 – 673, 2020. URL https://www.nature.com/articles/s42256-020-00257-z.

Hahn, M. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171, 2020. doi: 10.

1162/tacl_a_00306. URL https://aclanthology.org/2020.tacl-1.11/.

Hamilton, W. L. *Graph representation learning*. Morgan & Claypool Publishers, 2020.

Hao, Y., Angluin, D., and Frank, R. Formal language recognition by hard attention transformers: Perspectives from circuit complexity. *Transactions of the Association for Computational Linguistics*, 10:800–810, 2022. doi: 10.1162/tacl_a_00490. URL https://aclanthology.org/2022.tacl-1.46/.

Kao, K.-C., Wang, R., and Hsieh, C.-J. Solving for X and beyond: Can large language models solve complex math problems with more-than-two unknowns? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 16821–16843, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.980. URL https://aclanthology.org/2024.findings-emnlp.980/.

Li, Y., Guerin, F., and Lin, C. Latesteval: addressing data contamination in language model evaluation through dynamic and time-sensitive test construction. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i17.29822. URL https://doi.org/10.1609/aaai.v38i17.29822.

Lindner, D., Kramar, J., Farquhar, S., Rahtz, M., McGrath, T., and Mikulik, V. Tracr: Compiled transformers as a laboratory for interpretability. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=tbbId8u7nP.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach, 2019. URL https://arxiv.org/abs/1907.11692.

McCoy, R. T., Yao, S., Friedman, D., Hardy, M., and Griffiths, T. L. Embers of autoregression: Understanding large language models through the problem they are trained to solve, 2023. URL https://arxiv.org/abs/2309.13638.

Meng, K., Bau, D., Andonian, A. J., and Belinkov, Y. Locating and editing factual associations in GPT. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=-h6WAS6eE4.

Merrill, W. and Sabharwal, A. The parallelism tradeoff: Limitations of log-precision transformers. *Transactions of the Association for Computational Linguistics*, 11:531–545, 2023. doi: 10.1162/tacl_a_00562. URL https://aclanthology.org/2023.tacl-1.31/.

Merrill, W. and Sabharwal, A. The expressive power of transformers with chain of thought. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NjNGlPh8Wh.

Merrill, W. and Sabharwal, A. A little depth goes a long way: The expressive power of log-depth transformers. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=5pHfYe10iX.

Mirzadeh, S. I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., and Farajtabar, M. GSM-symbolic: Understanding the limitations of mathematical reasoning in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=AjXkRZIvjB.

Nanda, N., Chan, L., Lieberum, T., Smith, J., and Steinhardt, J. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9XFSbDPmdW.

Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87 of *Applied Optimization*. Springer, New York, NY, 2004. ISBN 978-1-4020-7553-7. doi: 10.1007/978-1-4419-8853-9. URL https://link.springer.com/book/10.1007/978-1-4419-8853-9.

Nichani, E., Damian, A., and Lee, J. D. How transformers learn causal structure with gradient descent. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=jNM4imlHZv.

Niven, T. and Kao, H.-Y. Probing neural network comprehension of natural language arguments. In Korhonen, A., Traum, D., and Màrquez, L. (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4658–4664, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1459. URL https://aclanthology.org/P19-1459/.

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, 2022. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Sanford, C., Fatemi, B., Hall, E., Tsitsulin, A., Kazemi, M., Halcrow, J., Perozzi, B., and Mirrokni, V. Understanding transformer reasoning capabilities via graph algorithms. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=AfzbDw6DSp.

Saparov, A., Pawar, S. A., Pimpalgaonkar, S., Joshi, N., Pang, R. Y., Padmakumar, V., Kazemi, M., Kim, N., and He, H. Transformers struggle to learn to search. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=9cQB1Hwrtw.

Saxton, D., Grefenstette, E., Hill, F., and Kohli, P. Analysing mathematical reasoning abilities of neural models. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1gR5iR5FX.

Tang, R., Kong, D., Huang, L., and Xue, H. Large language models can be lazy learners: Analyze shortcuts in in-context learning. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 4645–4657, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.284. URL https://aclanthology.org/2023.findings-acl.284/.

Vasudeva, B., Fu, D., Zhou, T., Kau, E., Huang, Y., and Sharan, V. Transformers learn low sensitivity functions: Investigations and implications. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4ikjWBs3tE.

Wang, K. R., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Warshall, S. A theorem on boolean matrices. *J. ACM*, 9 (1):11–12, January 1962. ISSN 0004-5411. doi: 10.1145/321105.321107. URL https://doi.org/10.1145/321105.321107.

Weiss, G., Goldberg, Y., and Yahav, E. Thinking like transformers. In *International Conference on Machine Learning*, pp. 11080–11090. PMLR, 2021.

Wigderson, A. The complexity of graph connectivity. In Havel, I. M. and Koubek, V. (eds.), *Mathematical Foundations of Computer Science 1992*, pp. 112–132, Berlin, Heidelberg, 1992. Springer Berlin Heidelberg. ISBN 978-3-540-47291-9.

Yao, Y., Zhang, N., Xi, Z., Wang, M., Xu, Z., Deng, S., and Chen, H. Knowledge circuits in pretrained transformers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=YVXzZNxcag.

Ye, W., Jiang, L., Xie, E., Zheng, G., Ma, Y., Cao, X., Guo, D., Qi, D., He, Z., Tian, Y., Coffee, M., Zeng, Z., Li, S., Ting-hao, Huang, Wang, Z., Rehg, J. M., Kautz, H., and Zhang, A. The clever hans mirage: A comprehensive survey on spurious correlations in machine learning, 2025. URL https://arxiv.org/abs/2402.12715.

Yuan, Y., Zhao, L., Zhang, K., Zheng, G., and Liu, Q. Do LLMs overcome shortcut learning? an evaluation of shortcut challenges in large language models. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12188–12200, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.679. URL https://aclanthology.org/2024.emnlp-main.679/.

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

Zhou, H., Bradley, A., Littwin, E., Razin, N., Saremi, O., Susskind, J. M., Bengio, S., and Nakkiran, P. What algorithms can transformers learn? a study in length generalization. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=AssIuHnmHX.

Zhou, R., Xu, M., Chen, S., Liu, J., Li, Y., Xinxin, L., Chen, Z., and He, J. Math for AI: On the generalization of learning mathematical problem solving. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*, 2024b. URL https://openreview.net/forum?id=xlnvZ85CSo.

Zhou, T., Fu, D., Sharan, V., and Jia, R. Pre-trained large language models use fourier features to compute addition. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024c. URL https://openreview.net/forum?id=i4MutM2TZb.

Zhou, Y., Xu, P., Liu, X., An, B., Ai, W., and Huang, F. Explore spurious correlations at the concept level in language models for text classification. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 478–492, Bangkok, Thailand, August 2024d. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.28. URL https://aclanthology.org/2024.acl-long.28/.

Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., and Fredrikson, M. Universal and transferable adversarial attacks on aligned language models, 2023. URL https://arxiv.org/abs/2307.15043.

# Appendix

## A Additional Details on Problem Setups and Preliminary Studies

**Definition A.1** (Transformer for Graph Connectivity: full specification)**.**

**Input and output**. Given a simple graph on $n$ nodes with adjacency matrix $A \in \{0,1\}^{n\times n}$, let $\bar{A} = A + I_n$ be its self-loop augmented adjacency matrix. We treat $\bar{A}$ as input embedding: row $i$ is the token for node $i$; column $j$ indexes a feature tied to node $j$. To ease notation, we will simply write $A$ in place of $\bar{A}$ and always assume the adjacency matrix is self-loop augmented. The model outputs an $n \times n$ score matrix $\mathrm{TF}^L_\Theta(A)$, the predicted connectivity matrix.

**Dimensions and parameters**. Fix depth $L$, hidden dimension $d > n$, number of heads $H$ with $d = Hd_h$, and feed-forward width $d_{\mathrm{ff}}$. The parameters we need include:

$$W_{\mathrm{in}}, W_{\mathrm{out}} \in \mathbb{R}^{n\times d}, \qquad W^Q_{\ell,h}, W^K_{\ell,h}, W^V_{\ell,h} \in \mathbb{R}^{d\times d_h}, \qquad W^O_\ell \in \mathbb{R}^{Hd_h\times d}$$

$$W^{(1)}_\ell \in \mathbb{R}^{d\times d_{\mathrm{ff}}},\ b^{(1)}_\ell \in \mathbb{R}^{d_{\mathrm{ff}}}, \qquad W^{(2)}_\ell \in \mathbb{R}^{d_{\mathrm{ff}}\times d},\ b^{(2)}_\ell \in \mathbb{R}^d$$

for $\ell = 1, \ldots, L$ and heads $h = 1, \ldots, H$. We use pre-norm residual blocks with LayerNorm (LN) and GeLU activations. We do not use attention masks or any extra positional encoding; the identity in $\bar{A}$ already pins each token to a node.

**The forward map**. The read-in is linear: $h^{(0)} = \bar{A}W_{\mathrm{in}} \in \mathbb{R}^{n\times d}$. From there, for each $\ell = 1, \ldots, L$, let $\tilde{h} = \mathrm{LN}(h^{(\ell-1)})$ as we use pre-norm. Within each block:

| | |
|---|---|
| **Multi-head self-attention** | $Q_{\ell,h} = \tilde{h}W^Q_{\ell,h}, \quad K_{\ell,h} = \tilde{h}W^K_{\ell,h}, \quad V_{\ell,h} = \tilde{h}W^V_{\ell,h}$ |
| **Attention scores** | $\alpha_{\ell,h} = \frac{1}{n}\mathrm{ReLU}(1/\sqrt{d_h} \cdot Q_{\ell,h}K^\top_{\ell,h}), \quad z_{\ell,h} = \alpha_{\ell,h}V_{\ell,h}$ |
| **Concatenation & residual** | $z_\ell = [z_{\ell,1} \mid \ldots \mid z_{\ell,H}]W^O_\ell \in \mathbb{R}^{n\times d}, \qquad u_\ell = h^{(\ell-1)} + z_\ell$ |
| **Feed-forward** | $\hat{u}_\ell = \mathrm{LN}(u_\ell), \qquad \mathrm{FFN}_\ell(\hat{u}_\ell) = \mathrm{GeLU}(\hat{u}_\ell W^{(1)}_\ell + b^{(1)}_\ell)W^{(2)}_\ell + b^{(2)}_\ell$ |

and finally $h^{(\ell)} = u_\ell + \mathrm{FFN}_\ell(\hat{u}_\ell) \in \mathbb{R}^{n\times d}$. The read-out is linear: $\mathrm{TF}^L_\Theta(A) = h^{(L)}W^\top_{\mathrm{out}} \in \mathbb{R}^{n\times n}$.

**Metrics for Permutation Equivariance.** Let $P \in S_n$ be the corresponding permutation matrix for any $\sigma \in \mathcal{S}_n$. For a given graph adjacency matrix $A$, we compute the model's prediction in respect to $P_\sigma$ as $\mathcal{M}(P_\sigma A P^\top_\sigma)$. Now we define a equivariance consistency metric, **Equivariance Consistency via Frobenius Cosine Similarity**:

$$\mathsf{ConsFrob}(\mathcal{M}) = \mathbb{E}_{\sigma\in\mathcal{S}_n, A\in\mathcal{G}}\left[\frac{\langle \mathcal{M}(P_\sigma A P^\top_\sigma), P_\sigma \mathcal{M}(A) P^\top_\sigma\rangle_F}{\|\mathcal{M}(P_\sigma A P^\top_\sigma)\|_F \|P_\sigma \mathcal{M}(A) P^\top_\sigma\|_F}\right] \tag{3}$$

When measuring intermediate model computations, this metric is modified depending on the model type. For standard Transformer models, $\mathcal{M}^\ell$ computes the final Readout to the hidden states at layer $\ell$. For Disentangled Transformers we are computing $P_\sigma \mathcal{M}^\ell(A)(P_\sigma \otimes I_n)^\top$.

# B Details for Capacity

## B.1 Expressivity

**Theorem 4.3.** *There exists an $L$-layer Disentangled Transformer that makes perfect predictions for every graph $G$ satisfying $\mathrm{diam}(G) \le 3^L$.*

*Proof.* Set $W_\ell = I_{d_{\ell-1}}$ for all layers and note that all matrices are entrywise nonnegative, so ReLU and the factor $1/n$ never changes supports. With $h_0 = [I \mid A] = [A^0 \mid A^1]$ and update $h_\ell = [h_{\ell-1} \mid (h_{\ell-1}h_{\ell-1}^\top)h_{\ell-1}/n]$, we can show by induction that every $n \times n$ block of $h_\ell$ lies in $\mathrm{span}\{A^0, \dots, A^{3^\ell}\}$, and that some block contains $A^{3^\ell}$ with a positive coefficient. Indeed, the base case holds trivially; for the inductive step, if a block within $h_{\ell-1}$ contains $A^m$, then $(h_{\ell-1}h_{\ell-1}^\top)h_{\ell-1}$ contains $A^{2m}A^m = A^{3m}$. Finally, the readout simply sums over all these blocks, so $\mathrm{supp}(\mathrm{TF}_\Theta^L(A)) = \mathrm{supp}(A^{3^L})$.

Finally, because $A$ has self-loops, supports are monotone in power and stabilizes at $t \ge \mathrm{diam}(G)$. Thus, if $\mathrm{diam}(G) \le 3^L$ we get $\mathrm{supp}(\mathrm{TF}_\Theta^L(A)) = \mathrm{supp}(A^{\mathrm{diam}(G)})$. □

## B.2 Capacity

**Theorem 4.5.** *Fix $L \ge 1$ and let $\mathsf{TF}_\Theta^L$ be an $L$-layer Disentangled Transformer on $n = \Omega(3^L)$ nodes. Further assume that the weights $W_\ell \ge 0$ for each $\ell$. Then there exists a graph $G$ with diameter $> 3^L$ on which $\mathsf{TF}_\Theta^L(A)$ is not perfect. In other words, diameter $3^L$ upper bounds the capacity of any $L$-layer Disentangled Transformer. In particular, taking $n \ge (7/3) \cdot 3^L + 2$ suffices.*

For each layer $\ell$ we define the post-ReLU score $R_\ell = \mathrm{ReLU}(h_{\ell-1}W_\ell h_{\ell-1}^\top)$. The proof of the theorem will be partitioned into two branches: whether some intermediate $R_\ell$ gives a false positive on some graph, or all $R_\ell$'s are free of false positives on all graphs. We say a pair of nodes $(u, v)$ from $G$ is a *witness* to false positives if they belong to different connected components while $R_\ell(G)_{u,v} > 0$. Throughout this section, we set $n \ge (7/3) \cdot 3^L + 2$.

**Lemma B.1.** *Assume the setup in Theorem 4.5. Suppose there exist some $n$-node graph, a layer index $\ell^* \in \{1, \dots, L\}$, and vertices $u, v$ belonging to different connected components of the graph such that $(R_{\ell^*})_{u,v} > 0$. Further assume that $\ell^*$ is globally minimal, in the sense that for all $n$-node graphs and all $\ell < \ell^*$, the corresponding $R_\ell$ has no false positive entries across components. Then there exists a graph $G$ such that $\mathrm{diam}(G) = 3^L + 1$, $\mathsf{TF}_\Theta^L(A(G))_{u,v} > 0$, where $u, v$ lie in different connected components of $G$.*

*Proof.* The proof roughly partitions into two parts. In the first half, we backtrack the computation DAG, tracing the "sources" that contribute to the false positiveness of $(u, v)$. This gives us subgraphs which we call *certificates* that, if kept untouched, suffice to guarantee a false positiveness of $(u, v)$. In the second half, we construct a graph $G$ that preserves these certificates while also containing a path of length $> 3^L$ disjoint from both certificates, and we show that $\mathsf{TF}_\Theta^L$ preserves false positiveness of $(u, v)$ on $G$, thereby proving the claim.

STEP 1. CONSTRUCTING THE CERTIFICATES. Intuitively, since $R_{\ell^*}(H)_{u,v} > 0$, there exist column indices $p, q$ with $(W_{\ell^*})_{p,q} > 0, h_{\ell^*-1}(H)_{u,p} > 0$, and $h_{\ell^*-1}(H)_{v,q} > 0$. We will backtrack the entries that contribute to the positiveness of the hidden states entries $h_{\ell^*-1}(H)_{u,p}$ and $h_{\ell^*-1}(H)_{v,q} > 0$ in the computation DAG, iteratively visiting previous layers. Formally, we define a *certificate* for an entry $h_t(H)_{i,c} > 0$ to be a small tree whose nodes are triples [of form (layer, row, column)] recording earlier entries that must be positive to guarantee that the current one is positive. The root is $(t, i, c)$ and we build it top-down by repeating one of the two rules until we hit the first layer, which we know looks like $[I_n \mid A]$. We now describe how to backtrack. Since $h_t = [h_{t-1} \mid \mathrm{Attn}(h_{t-1}; W_t)]$, we split the recursion on layer $t$ into two cases: whether the entry lies in the first half ($h_{t-1}$) or the second half ($\mathrm{Attn}(h_{t-1}; W_t)$).

- (First half) If column $c$ is in the inherited block of $h_t$, add a single child $(t-1, i, c)$ to $(t, i, c)$, as the value is simply copied from the previous layer $h_{t-1}$.
- (Second half) If column $c$ is in the newly appended block, then by definition

$$h_t(H)_{i,c} = \frac{1}{n} \sum_k R_t(H)_{i,k} h_{t-1}(H)_{k,c'} \qquad \text{for some } c',$$

and since this is a sum of nonnegative terms, there exists at least one $k$ with $R_t(H)_{i,k} > 0$ and $h_{t-1}(H)_{k,c'} > 0$. In turn,

$$R_t(H)_{i,k} = \sum_{r,s} h_{t-1}(H)_{i,r}(W_t)_{r,s}h_{t-1}(H)_{k,s} > 0,$$

which implies that there exist indices $r, s$ with $(W_t)_{r,s} > 0$, $h_{t-1}(H)_{i,r} > 0$, and $h_{t-1}(H)_{k,s} > 0$. Thus, for such $(t, i, c)$, we create three children:

$$(t-1, k, c'), \qquad (t-1, i, r), \qquad (t-1, k, s).$$

Now let $s(t)$ denote the maximal number of vertices needed to realize a single certificate for some entry $h_t(\cdot) > 0$ by the recursive procedure above. At $t = 0$ we may assume $s(0) \le 2$. The recursion gives $s(t) \le 3s(t-1)$, so $s(t) \le 2 \cdot 3^t$. Since $u, v$ lie in different connected components of $H$, and $\ell^*$ is minimal, every index $k$ selected by a certificate at any layer $t \le \ell^* - 1$ stays within the same component as its $i$, so the two certificates induce trees $T_u, T_v$ that occupy disjoint vertex sets $S_u, S_v$, with $|S_u \cup S_v| \le 4 \cdot 3^{\ell^*-1} \le 4 \cdot 3^{L-1}$ vertices.

STEP 2. BUILDING A NEW GRAPH. Initialize $G$ to the edgeless graph, keeping node isolated. We then embed $T_u, T_v$ onto $G$ by adding edges according to the trees. Finally, we connect $3^L + 2$ vertices outside $S_u \cup S_v$ arbitrarily into a long chain of path length $3^L + 1$. This is always possible because there are $n - |S_u \cup S_v|$ vertices outside the union, which is $\ge \big((7/3) \cdot 3^L + 2\big) - 4 \cdot 3^{L-1} = 3^L + 2$, and this is why require $n \ge (7/3) \cdot 3^L + 2$ in this Lemma.

We claim $G$ is the graph we seek. On one hand, every sum used by the certificates is a sum of nonnegative terms, and we have preserved a strictly positive summand at each step that appears in the tree. Hence $R_{\ell^*}(G)_{u,v} > 0$ with $u, v$ also disconnected in $G$. On the other hand, under the choice of $n$ specified by Theorem 4.5, there exist at least $3^L + 2$ vertices outside $S_u \cup S_v$, so connecting them into a long path guarantees $\mathrm{diam}(G) = 3^L + 1$. The claim then follows. □

**Lemma B.2.** *Assume the setup in Theorem 4.5. Further assume that for every $n$-node graph $G$ and every layer $\ell \in \{1, \dots, L\}$, the post-ReLU scores $R_\ell(G)$ has no positive entry between distinct connected components of $G$. Then, for every graph $G$ and every $u, v \in V(G)$, if $\mathsf{TF}^L_\Theta(A(G))_{u,v} > 0$, we must have $\mathrm{dist}_G(u, v) \le 3^L$. Consequently, if $G$ contains a connected component of diameter $3^L + 1$ then $\mathsf{TF}^L_\Theta$ is not perfect on $G$.*

*Proof.* Under the no-false-positives assumption, the idea is to show that "information" spreads no faster than power base 3 so $\mathsf{TF}^L_\Theta$ never predicts "Yes" on node pairs with distance beyond $3^L$. Concretely, columns exchange information as attention scores are calculated. We first define the "distances" between columns by giving each column a label $\in \{1, \dots, n\}$, and then show that by layer $\ell$, two columns can "share" information if and only if their labels, *interpreted as graph nodes,* are within distance $3^\ell$.

STEP 1. GIVING EACH COLUMN A LABEL. We first consider trivial graph $G_0$ with $n$ isolated nodes: immediately $h_0(G_0) = [I_n \mid I_n]$ and, by hypothesis, every $R_\ell(G_0)$ have no off-diagonal positives. Inductively this shows that every column of $h_\ell(G_0)$ has support in exactly one row. We define the label of this column to be the row index $\in \{1, \dots, n\}$ where the unique support is. With labels defined, the remaining proof is based on establishing the following locality claim.

CLAIM. Fix graph $G$, layer $\ell$, and $i, j \in \{1, \dots, n\}$. If column $c$ of $h_\ell(G)$ has label $j$ and if $h_\ell(G)_{i,c} > 0$, then $\mathrm{dist}_G(i, j) \le 3^\ell$. In other words, *every column spreads at most $3^\ell$ hops away from its label by depth $\ell$.*

STEP 2. ESTABLISHING THE CLAIM. We prove this claim via induction. The base case $\ell = 0$ directly follows from the fact that $h_0(G) = [I_n \mid A(G)]$. For the inductive step, we assume that the claim holds at depth $\ell - 1$ with radius $3^{\ell-1}$. As in Lemma B.1, there are two column types in $h_\ell$: inherited or newly appended columns. The former case is easy; if $c$ is inherited from $h_{\ell-1}$, then $h_\ell(G)_{i,c} = h_{\ell-1}(G)_{i,c}$, so the bound follows from the inductive hypothesis. We now assume $c$ is newly appended.

Suppose $(R_\ell(G)h_{\ell-1}(G))_{i,c} > 0$ for a column $c$ with label $j$. Then there exists a row $k$ with $R_\ell(G)_{i,k} > 0$ and $h_{\ell-1}(G)_{k,c} > 0$. By the IH, $\mathrm{dist}_G(k, j) \le 3^{\ell-1}$. Then we expand $R_\ell(G)_{i,k} > 0$ to obtain column witnesses $p, q$, with $h_{\ell-1}(G)_{i,p} > 0$, $h_{\ell-1}(G)_{k,q} > 0$, and $(W_\ell)_{p,q} > 0$, as in Lemma B.1. Let $a, b$ be the labels of $p, q$, respectively. By IH again, $\mathrm{dist}_G(i, a) \le 3^{\ell-1}$ and $\mathrm{dist}_G(k, b) \le 3^{\ell-1}$. We now split the analysis into two cases.

- If $a \ne b$, we derive a contradiction to the no-false-positives assumption by reusing the certificate procedure from Lemma B.1. Because $W_\ell \ge 0$ entrywise, every positive entry in $h_t(\cdot)$ admits a certificate supported on at most $s(t) \le 2 \cdot 3^t$ vertices. In particular, there exist certificates witnessing $h_{\ell-1}(G)_{i,p} > 0$ (labeled $a$) and $h_{\ell-1}(G)_{k,q} > 0$

(labeled $b$). Let $S_a, S_b$ be the corresponding certificate vertex sets. Form a new graph $G'$ on the same $n$ vertices whose connected components are two disjoint induced copies $S'_a, S'_b$ of the subgraphs on $S_a, S_b$ (leaving all other vertices outside $S_a \cup S_b$ isolated). Note this is feasible because $|S_a| + |S_b| \le 4 \cdot 3^{L-1} \le n$ assumed by Theorem 4.5. By construction, there exist $i' \in S'_a$ and $k' \in S'_b$ with $h_{\ell-1}(G')_{i',p} > 0$ and $h_{\ell-1}(G')_{k',q} > 0$. Thus,

$$(h_{\ell-1}(G')W_\ell h_{\ell-1}(G')^\top)_{i',k'} \ge h_{\ell-1}(G')_{i',p}(W_\ell)_{p,q}h_{\ell-1}(G')_{k',q} > 0,$$

meaning $R_\ell(G')_{i',k'} > 0$. But $i', k'$ belong to different connected components in $G'$, contradiction! Therefore,

- $a = b$. Triangle inequality gives $\mathrm{dist}_G(i,j) \le \mathrm{dist}_G(i,a) + \mathrm{dist}_G(a,k) + \mathrm{dist}_G(k,j) \le 3 \cdot 3^{\ell-1} = 3^\ell$, completing the induction. END PROOF OF CLAIM / STEP 2.

The model's output $\mathsf{TF}^L_\Theta(A(G)) = h_L(G)W_O^\top$ is an entrywise nonnegative sum over the $n \times n$ blocks of $h_L(G)$. Since each block respects the $3^L$ locality bound, we have $\mathsf{TF}^L_\Theta(A(G))_{u,v} = 0$ whenever $\mathrm{dist}_G(u,v) \ge 3^L + 1$. Hence, on any graph whose largest component has diameter $3^L + 1$, the model will inevitably miss a pair $(u,v)$ of nodes realizing this diameter. □

*Proof of Theorem 4.5.* Combine Lemmas B.1 and B.2. □

# C Details for Training Dynamics

## C.1 Characterizing Block Weights $W_\ell$

As discussed in Section 4.3, due to the symmetric nature of the graph connectivity problem, it is natural to demand that a "good" model should map not only adjacency matrices $A$ to connectivity matrices $R$, but also $PAP^\top$ to $PRP^\top$ for any permutation $P$. We further generalize equivariance. Observe that given a permutation matrix $P$ and any hidden states $h \in \mathbb{R}^{n\times(kn)}$ consisting of $k$ consecutive $n \times n$, the mapping $h \mapsto Ph(I_K \otimes P^\top)$ relabels both rows and columns within each $n \times n$ block in a way that is consistent with the effects of $P$. Hence, the notion of equivariance can be generalized to any (nonnegative) hidden states, beyond just the ones induced by adjacency matrices.

Similarly, we are now also able to define an $L$-layer Disentangled Transformer on arbitrary inputs of appropriate dimensions. For any nonnegative initial state $h_0 \in \mathbb{R}^{n\times 2n}$, recursively define $h_\ell = [h_{\ell-1} \mid \mathrm{Attn}(h_{\ell-1}; W_\ell)]$ for $\ell = 1, \ldots, L$. Let $\mathrm{Sum}(h)$ denote the sum of the consecutive left-aligned $n \times n$ blocks of $h$. Then the generalized output is $\mathsf{TF}^L_\Theta(h_0) = \mathrm{Sum}(h_L)$. We define two equivariance-related conditions. The first one is a direct generalization of $P\mathsf{TF}^L_\Theta(A)P^\top = \mathsf{TF}^L_\Theta(PAP^\top)$; the second one, as discussed in Section 4.3, makes theoretical analysis significantly more tractable while also being supported by empirical evidence.

**Definition C.1** (Output Equivariance and Layerwise Attention Equivariance)**.** Let $\mathsf{TF}^L_\Theta$ be an $L$-layer Disentangled Transformer with nonnegative weights. Let $K_\ell = 2^{\ell+1}$.

(i) For $h_0 \in \mathbb{R}^{n\times 2n}_{\geq 0}$ and for any $P$, define $h_0^P = Ph_0(I_{K_0} \otimes P^\top)$. We say $\mathsf{TF}^L_\Theta$ is **output-level value equivariant** iff $P\mathsf{TF}^L_\Theta(h_0)P^\top = \mathsf{TF}^L_\Theta(h_0^P)$ holds for all $P$ and all $h_0 \in \mathbb{R}^{n\times 2n}_{\geq 0}$.

(ii) We say $\mathsf{TF}^L_\Theta$ is **layer-wise attention equivariant** iff for each $\ell$ and any hidden states $h \in \mathbb{R}^{n\times d_{\ell-1}}$ (i.e., any hidden states of dimension feasible for layer $\ell$),

$$\mathrm{Attn}(Ph(I_{K_{\ell-1}} \otimes P^\top); W_\ell) = P\ \mathrm{Attn}(h; W_\ell)\ (I_{K_{\ell-1}} \otimes P^\top),$$

**Theorem C.2** (Parameterization of "Good" Models)**.** *Let $n = \Omega(3^L)$ as in Theorem 4.5. Fix an $L$-layer Disentangled Transformer $\mathsf{TF}^L_\Theta$ with nonnegative weights. Suppose that*

*(i) $\mathsf{TF}^L_\Theta$ is output-level value-equivariant, and*

*(ii) $\mathsf{TF}^L_\Theta$ reaches its capacity bound of $3^L$, i.e., for every graph, we have $\mathrm{supp}(\mathsf{TF}^L_\Theta(A)) = \mathrm{supp}(A^{3^L})$.*

*Then, either $\mathsf{TF}^L_\Theta$ or a functionally equivalent version of it satisfies the following: for each layer $\ell$, there exists a nonnegative matrix $\Lambda_\ell \in \mathbb{R}^{K_{\ell-1}\times K_{\ell-1}}$ such that $W_\ell + W_\ell^\top = \Lambda_\ell \otimes I_n$. In other words, $W_\ell$ can be decomposed into this form up to an antisymmetric part.*

*(Note that the theorem is a direct generalization of equivariance under all graph permutations; replacing $h_0$ by $[I_n \mid A]$ gives the desired result for a fixed graph with adjacency matrix $A$.)*

*Proof.* To prove the claim, it suffices to show that if we partition $W_\ell$ into $K_{\ell-1} \times K_{\ell-1}$ contiguous sub-blocks of size $n \times n$, then each block must be diagonal, with symmetry conditions meeting $W_\ell + W_\ell^\top = \Lambda_\ell \otimes I_n$.

To do so, the proof is split into two parts: we prove that each $n \times n$ block must be diagonal using (ii) and Lemma B.2, and that the diagonal entries must realize the said forms by examining the forward maps under a curated, parameterized class of initial hidden states.

STEP 1: EACH BLOCK MUST BE DIAGONAL. In this step, we argue that if a block admits a positive off-diagonal entry, then the certificate trick from Lemma B.1 will create a false positive entry on some output, contradicting (ii).

Formally, let $R_\ell = \mathsf{ReLU}(h_{\ell-1}W_\ell h_{\ell-1}^\top)$. If for some graph and some $\ell$, there exists a false positive entry $(R_\ell)_{i,k} > 0$ for some $i, k$ across different connected components, then the false positiveness would persist to the output, contradicting (ii). Hence $\mathsf{TF}^L_\Theta$ must have no false positives.

Consider feeding the graph $G_0$ of $n$ isolated vertices into $\mathsf{TF}^L_\Theta$, so that $h_0(G_0) = [I_n \mid I_n]$. The premises of Lemma B.2 hold, so every column of every $h_\ell(G_0)$ is supported in exactly one row, which we called its label in $\{1, \ldots, n\}$. Hence, if we write $h_\ell(G_0) = [X_1^{(\ell)} \mid \ldots \mid X_{K_\ell}^{(\ell)}]$ of contiguous $n \times n$ blocks, then each such block $X_r^{(\ell)}$ must be nonnegative and

diagonal. Now expand

$$R_\ell = \mathsf{ReLU}(h_{\ell-1}W_\ell h_{\ell-1}^\top) = h_{\ell-1}W_\ell h_{\ell-1}^\top = \sum_{r,s} X_r^{(\ell-1)} W_\ell[r,s](X_s^{(\ell-1)})^\top.$$

We first claim that every $n \times n$ sub-block $W_\ell[r,s]$ is diagonal. Suppose not, that there exist indices $r, s$ and distinct nodes $i \neq k$ such that $(W_\ell[r,s])_{i,k} > 0$. For a node $i$ and a block $r$, we say $(i,r)$ is *activatable* at depth $\ell-1$ if there exists *some* graph $G$ such that $X_r^{(\ell-1)}(G)[i,i] = h_{\ell-1}[i,(r-1)n+i] > 0$. Two cases:

- If at least one of $(i,r)$ or $(k,s)$ is not activatable, then for every graph $G$, at least one factor $X_r^{(\ell-1)}(G)[i,i]$ or $X_s^{(\ell-1)}(G)[k,k]$ is zero, and thus $(W_\ell[r,s])_{i,k}$ is functionally inert and never contributes to any $R_\ell$ entry. Hence we may simply set it to 0 without altering the model's output on any graph.
- If both $(i,r)$ and $(k,s)$ are activatable, take graphs $G_i, G_k$ that make $X_r^{(\ell-1)}(G_i)[i,i] > 0$ and $X_x^{(\ell-1)}(G_k)[k,k] > 0$. Using the certificate mechanism in Lemma B.1, each positiveness admits a finite certificate subgraph with at most $2 \cdot 3^{\ell-1}$ vertices. We then create a new graph $G'$ and disjointly embed both certificates into it, leaving all other vertex isolated. The two labels $i, k$, viewed as nodes, now lie in different components. But then the product

  $$X_r(G')[i,i] \cdot (W_\ell[r,s])_{i,k} \cdot X_k(G')[k,k] > 0,$$

  making $(R_\ell(G'))_{i,k} > 0$, contradiction.

Therefore $W_\ell[r,s]$ is diagonal for all block indices $(r,s)$. This concludes STEP 1.

STEP 2. $W_\ell$ IS NODE-SYMMETRIC. Given a triplet $(\ell, r, s)$, we can now write $W_\ell[r,s]$ as $\mathrm{diag}(w_{\ell,r,s}(1), \ldots, w_{\ell,r,s}(n))$. Our goal is to show that for each $(\ell,r,s)$, $w_{\ell,r,s}(j) + w_{\ell,s,r}(j) = w_{\ell,r,s}(k) + w_{\ell,s,r}(k)$ for all $j, k \in [n]$. We formalize this in matrix form: For each node $i \in [n]$ and each layer $\ell$, let $\Lambda_\ell^{(i)} = [w_{\ell,r,s}(i)]_{r,s} \in \mathbb{R}^{K_{\ell-1} \times K_{\ell-1}}$ and define the symmetric part $\mathrm{Sym}(\Lambda_\ell^{(i)}) = (\Lambda_\ell^{(i)} + \Lambda_\ell^{(i)T})/2$; the goal is to show that given $\ell$, all $\Lambda_\ell^{(i)}$ are the same, so that $\mathrm{Sym}(W_\ell) = \Lambda_\ell \otimes I_n$ or equivalently, $W_\ell + W_\ell^\top = \Lambda_\ell \otimes I_n$, as claimed.

Throughout out this step, we will use a family of special hidden states parameterized by a scalar $\lambda > 0$ and a vector $u = (u_1, u_2) \in \mathbb{R}^2_{\geq 0}$. Fix distinct nodes $j \neq k$. For $\lambda, u$, define the initial state $h_0(\lambda, u) \in \mathbb{R}^{n \times 2n}$ by setting exactly four entries nonzero:

$$\begin{cases} h_0(\lambda,u)[j,j] = \lambda u_1 & h_0(\lambda,u)[j,n+j] = \lambda u_2 \\ h_0(\lambda,u)[k,k] = \lambda u_1 & h_0(\lambda,u)[k,n+k] = \lambda u_2. \end{cases}$$

Note that $h_0(\lambda,u)$ is invariant under the transposition $P = (j,k)$, i.e., $Ph_0(\lambda,u)(I_{K_0} \otimes P^\top) = h_0(\lambda,u)$. Therefore, by assumption (i), we must have $\mathsf{TF}_\Theta^L(h_0)_{j,j} = \mathsf{TF}_\Theta^L(h_0)_{k,k}$. Let $h_\ell(\lambda,u)$ be the network state at depth $\ell$. Because of STEP 1, there is no cross-row interaction for this input at any depth. Writing the row-$i$ vector as $v_\ell^{(i)}(\lambda,u) \in \mathbb{R}^{K_\ell}$, recursion gives, for $i \in \{j,k\}$,

$$v_0^{(i)}(\lambda,u) = \lambda u, \qquad v_\ell^{(i)}(\lambda,u) = [v_{\ell-1}^{(i)}(\lambda,u) \mid q_\ell^{(i)}(\lambda,u) v_{\ell-1}^{(i)}(\lambda,u)]$$

where

$$q_\ell^{(i)}(\lambda,u) = \frac{1}{n} \cdot v_{\ell-1}^{(i)}(\lambda,u)^\top \, \mathrm{Sym}(\Lambda_\ell^{(i)}) \, v_{\ell-1}^{(i)}(\lambda,u).$$

Taking $\ell_1$-norms gives

$$\|v_\ell^{(i)}(\lambda,u)\| = (1 + q_\ell^{(i)}(\lambda,u))\|v_{\ell-1}^{(i)}(\lambda,u)\| \quad \text{and} \quad \|v_L^{(i)}(\lambda,u)\| = \|v_0^{(i)}(\lambda,u)\| \prod_{\ell=1}^{L} (1 + q_\ell^{(i)}(\lambda,u)). \tag{4}$$

Because the readout weight $W_O$ is a concatenation of $I_n$'s, and under our specific input $h_0(\lambda,u)$, every nonzero row $i$ lies in columns with indices $i$ modulo $n$, the $(i,i)$ output numerically equals $\|v_L^{(i)}(\lambda,u)\|$. Hence, assumption (i) requires $\|v_L^{(j)}(\lambda,u)\| = \|v_L^{(k)}(\lambda,u)\|$.

Let $\ell^*$ be the minimal layer such that $\mathrm{Sym}(\Lambda_{\ell^*}^{(j)}) \neq \mathrm{Sym}(\Lambda_{\ell^*}^{(k)})$. If no such $\ell^*$ exists for all $j \neq k$, then all $\Lambda_\ell^{(i)}$'s are the same given any fixed $\ell$, and STEP 2 holds. Otherwise, for every $\ell < \ell^*$, the symmetric parts coincide, and $v_{\ell-1}^{(j)}(\lambda,u) = v_{\ell-1}^{(k)}(\lambda,u)$

and $q_\ell^{(j)}(\lambda,u) = q_\ell^{(k)}(\lambda,u)$ for all $\lambda, u$. We may use $v_{\ell^*-1}(\lambda,u)$ to denote both $v_{\ell^*-1}^{(j)}(\lambda,u)$ and $v_{\ell^*-1}^{(k)}(\lambda,u)$ for they are now equal.

Because of the structure of $h_0(\lambda,u)$, by induction, the row vectors of each hidden state admits an odd power expansion

$$v_{\ell-1}^{(i)}(\lambda,u) = \lambda u + \lambda^3 \xi_{1,\ell-1}(u) + \lambda^5 \xi_{2,\ell-1}(u) + \dots$$

from which we conclude $q_\ell^{(i)}(\lambda,u) = O(\lambda^2)$ for every $\ell$. In particular, at $\ell = \ell^*$,

$$q_{\ell^*}^{(j)}(\lambda,u) - q_{\ell^*}^{(k)}(\lambda,u) = \frac{1}{n} \cdot v_{\ell^*-1}(\lambda,u)^\top (\mathrm{Sym}(\Lambda_{\ell^*}^{(j)}) - \mathrm{Sym}(\Lambda_{\ell^*}^{(k)})) v_{\ell^*-1}(\lambda,u) = \lambda^{2m} c(u) + o(\lambda^{2m})$$

for some $m \geq 1$ and some nondegenerate polynomial $c(u)$ as $\lambda \searrow 0$. In particular,

$$q_{\ell^*}^{(j)}(\lambda,u) - q_{\ell^*}^{(k)}(\lambda,u) = \Theta(\lambda^{2m}).$$

We now put this back into the comparison between the output's $(j,j)$ and $(k,k)$ entry. Recall that $v_0^{(j)}(\lambda,u) = v_0^{(k)}(\lambda,u) = \lambda u$. Further, since $v_{\ell-1} = \lambda u + O(\lambda^3)$, we know $q_\ell(\lambda,u) = O(\lambda^2)$ for every $\ell$ and every $i$. We drop $\lambda, u$ for notational simplicity. It follows from equation 4 that

$$\begin{aligned}
\|v_L^{(j)}\| - \|v_L^{(k)}\| &= \lambda \|u\| \cdot \left[\prod_{l<\ell^*} (1+q_\ell)\right] \cdot \left[\left[1+q_{\ell^*}^{(j)}\right] \prod_{\ell>\ell^*} \left[1+q_\ell^{(j)}\right] - \left[1+q_{\ell^*}^{(k)}\right] \prod_{\ell>\ell^*} \left[1+q_\ell^{(k)}\right]\right] \\
&= \lambda \|u\| \cdot \left[\prod_{\ell<\ell^*} (1+O(\lambda^2))\right] \cdot \left[q_{\ell^*}^{(j)} - q_{\ell^*}^{(k)}\right] \cdot \left[\prod_{\ell>\ell^*} (1+O(\lambda^2))\right] \\
&= \lambda \|u\| \Theta(\lambda^{2m})(1+o(1)) = \Theta(\lambda^{2m+1})
\end{aligned}$$

which is nonzero for small $\lambda$. Hence the $(j,j)$ and $(k,k)$ entries can be made different, contradicting assumption (i), and the proof is complete! □

**Theorem 4.7.** *Suppose an $L$-layer Disentangled Transformer $\mathsf{TF}_\Theta^L$ has nonnegative parameters. Suppose $\mathsf{TF}_\Theta^L$ is layerwise permutation equivariant, i.e., for each $\ell$, any hidden states $h \in \mathbb{R}^{n \times d_{\ell-1}}$, and any permutation $P \in S_n$,*

$$\mathrm{Attn}(Ph(I_{K_{\ell-1}} \otimes P^\top); W_\ell) = P\ \mathrm{Attn}(h; W_\ell)\ (I_{K_{\ell-1}} \otimes P^\top),$$

*then each block $W_\ell = A_\ell \otimes I_n + B_\ell \otimes J_n$ for some $A_\ell, B_\ell \in \mathbb{R}^{K_{\ell-1}, K_{\ell-1}}$. In other words, each block-aligned $n \times n$ submatrix of $W_\ell$ necessarily lies in* $\mathrm{span}\{I_n, J_n\}$.

*Remark* C.3. The equivariance condition presented in the theorem is strictly harder than what we need for graph-level, layerwise equivariance:

$$\mathrm{Attn}(h_{\ell-1}(PAP^\top); W_\ell) = P\ \mathrm{Attn}(h_{\ell-1}(A); W_\ell)\ (I_{K_\ell - 1} \otimes P^\top).$$

For graphs, it suffices to assume that the hidden states are induced by some $n$-node graph.

*Proof.* STEP 1. RELATING TO WEIGHT CONJUGATION. Fix a layer $\ell$. Write $K = K_{\ell-1}$, $h = h_\ell$, $W = W_\ell$, and let $\sigma(P) = I_K \otimes P$. The first step is to relate the conjugation of hidden states, $T_P(h) : h \mapsto Ph(I_K \otimes P^\top)$, to a conjugation of layer weights, $W_\ell \mapsto \sigma(P) W_\ell \sigma(P)^\top$.

Concretely, since $W_\ell \geq 0$, $\mathsf{ReLU}$. Hence

$$\begin{aligned}
\mathrm{Attn}(T_P(h); W) &= \frac{1}{n} \mathsf{ReLU}[(Ph\sigma(P))\ W\ (\sigma(P)^\top h^\top P^\top)](Ph\sigma(P)) \\
&= \frac{1}{n} P[h\sigma(P)\ W\ (\sigma(P)^\top h^\top)](h\sigma(P))
\end{aligned}$$

and

$$T_P(\mathrm{Attn}(h; W)) = \frac{1}{n} P(hWh^\top) h\sigma(P).$$

Layer-wise attention equivariance requires the two quantities above to equal for all $h$, and left multiplication by $P^{-1}$ gives

$$h\Delta h^\top h\,\sigma(P) = 0 \qquad \text{for all } h \geq 0 \qquad \text{where} \qquad \Delta := \sigma(P)W\sigma(P)^\top - W. \tag{*}$$

STEP 2. PROVING $\Delta = 0$. To do so, we consider special hidden states, with only two nonzero entries $h_{i,p} = 1$ and $h_{j,q} = t$. Equivalently, pick columns $p \neq q$ and rows/nodes $i \neq k$ and set $h_{i,\cdot} = e_p^\top$, $h_{j,\cdot} = te_q^\top$, and $h = 0$ everywhere else, where $e_p$ is standard basis vector pivoted at $p$.

Because $h$ only uses columns $p$ and $q$, the matrix $h\Delta h^\top$ can be embedded on rows/columns $\{i, j\}$ with values

$$h\Delta h^\top = \begin{pmatrix} \Delta_{p,p} & t\Delta_{p,q} \\ t\Delta_{q,p} & t^2\Delta_{q,q} \end{pmatrix}.$$

Recall $\sigma(P)$ is a permutation on columns; let $\pi$ be the permutation induced by it. Since $h\sigma(P)$ has the same two nonzero rows with $(h\sigma(P))_{i,\cdot} = e_{\pi(p)}^\top$ and $(h\sigma(P))_{j,\cdot} = te_{\pi(q)}^\top$, we get that $(h\Delta h^\top)(h\sigma(P))$ only has rows $i$ and $j$ potentially nonzero:

$$\begin{cases} \text{row } i : \Delta_{p,p}e_{\pi(p)}^\top + t^2\Delta_{p,q}e_{\pi(q)}^\top \\ \text{row } j : t\Delta_{q,p}e_{\pi(p)}^\top + t^2\Delta_{q,q}e_{\pi(q)}^\top. \end{cases}$$

But recall (*): $(h\Delta h^\top)(h\sigma(P)) = 0$ for all $t > 0$. The two standard basis vectors $e_{\pi(p)}, e_{\pi(q)}$ are linearly independent, so the coefficients must be uniformly zero! Hence $\Delta_{p,p} = \Delta_{p,q} = \Delta_{q,p} = \Delta_{q,q} = 0$. Finally, because $p \neq q$ were arbitrary, this forces $\Delta = 0$ entrywise. and that $\sigma(P)W_\ell\sigma(P)^\top = W_\ell$ for this $P$. And because $P$ is arbitrary, we conclude that $\sigma(P)W\sigma(P)^\top = W$ for every permutation $P$.

STEP 3. RELATING TO $n \times n$ BLOCKS. Consider any $n \times n$ block $W[u, v]$ of $W$ where $1 \leq u, v \leq K_\ell$. Using $\sigma(P) = I_{K_\ell} \otimes P^\top$ and taking the $(u, v)$ block on both sides,

$$(\sigma(P)W\sigma(P)^\top)[u, v] = \sum_{a,b}(I_{K_\ell})_{u,a}P^\top W[a, b]P(I_{K_\ell})_{b,v} = P^\top W[u, v]P.$$

The LHS equals $W[u, v]$, so we conclude that

$$P^\top W[u, v]P = W[u, v] \qquad \text{for all } P \in S_n.$$

In other words, layerwise equivariance implies each block must be invariant under $P^\top(\cdot)P$. Taking any transposition forces all diagonal entries of a block to equal, while for any $i \neq j, k \neq \ell$, any arbitrary permutation mapping $\pi(i) = k, \pi(j) = \ell$ forces entries $(i, j)$ and $(k, \ell)$ to be equal. This implies that each block lies in $\mathrm{span}\{I_n, J_n\}$ as claimed. □

### C.2 Population Gradient Lives in the Equivariant Algebra

**Theorem C.4** (Population gradient lives in the equivariant algebra)**.** *Under Assumption 4.8, in particular using layerwise parameterization $W_\ell = A_\ell \otimes I_n + B_\ell \otimes J_n$, fix a layer $\ell$ and let $K = K_{\ell-1}$. Then the population gradient with respect to $W_\ell$ lies in $M_K(\mathbb{R}) \otimes \mathrm{span}\{I_n, J_n\}$: there exist matrices $G_\ell^{(I)}, G_\ell^{(J)} \in \mathbb{R}^{K\times K}$ such that*

$$\mathbb{E}\Big[\frac{\partial \mathcal{L}}{\partial W_\ell}\Big] = G_\ell^{(I)} \otimes I_n + G_\ell^{(J)} \otimes J_n. \tag{5}$$

*Proof.* We let $S_n$ act on node indices. Since $W_\ell$ can be parametrized as $W_\ell = A_\ell \otimes I_n + B_\ell \otimes J_n$, the attention map is equivariant under left-right action:

$$\mathrm{Attn}(Ph(I_K \otimes P^\top); W_\ell) = P\,\mathrm{Attn}(h; W_\ell)(I_K \otimes P^\top),$$

and so is the full map $A \mapsto Z$. For any fixed permutation $P$, the data $\mathsf{ER}(n, p)$ is permutation-invariant, i.e., $A$ and $PAP^\top$ are identically distributed. Because the model map and the loss are equivariant under $A \mapsto PAP^\top$ with $R \mapsto PRP^\top$, the sample gradient covaries as

$$\nabla_{W_\ell}\mathcal{L}(PAP^\top) = (I_k \otimes P)\nabla_{W_\ell}\mathcal{L}(A)(I_K \otimes P^\top).$$

Taking expectation over $A$ gives

$$\mathbb{E}_A[\nabla_{W_\ell}\mathcal{L}(PAP^\top)] = (I_k \otimes P)\mathbb{E}_A[\nabla_{W_\ell}\mathcal{L}(A)](I_K \otimes P^\top)$$

for every $P$. Hence the population gradient lies in the commutant of $\{I_K \otimes P : P \in S_n\}$. It remains to identify this commutant. View $G_\ell$ as a $K \times K$ block matrix with $n \times n$ sub-blocks. The relation $(I_K \otimes P)^\top G_\ell (I_K \otimes P) = G_\ell$ says each $n \times n$ block $B$ satisfies $P^\top BP$ for all permutations $P$, so the block must have one value on the diagonal and one on the off-diagonals. It is well known that the fixed-point algebra of conjugation on $n \times n$ matrices is $\mathrm{span}(I_n, J_n)$. Hence every block lies in this span, i.e., $G_\ell \in M_K(\mathbb{R}) \otimes \mathrm{span}\{I_n, J_n\}$. □

### C.3 Which Conditions Encourage $W_\ell \approx A_\ell \otimes I_n$?

To facilitate the following analyses, it will be beneficial to first (re)introduce some notations.

Throughout the analysis of training dynamics, we inherit the notations used in Assumption 4.8: we use $Z$ to denote the model output, $R$ the reachability matrix, $A$ the adjacency matrix, $\mathcal{L} = \mathcal{L}(Z; R)$ the loss, and $\mathcal{R}(\Theta)$ the population risk $\mathcal{R}(\Theta) := \mathbb{E}_{G\sim\mathrm{ER}(n,p)}[\mathcal{L}(\mathsf{TF}^L_\Theta(A_G); R_G)]$.

Fix a layer $\ell$ and a nonnegative direction $\Delta \geq 0$ in the $J$-channel. Write $D = \frac{\partial Z}{\partial B_\ell}[\Delta]$ (more details in Theorem C.5). We say a node pair $(i, j)$ is **active** for $\Delta$ if $D_{i,j} > 0$. In particular, we say $\Delta$ is active on cross-component pairs if $D_{i,j} > 0$ for some $(i, j)$ belonging to different connected components (note $\Delta$ could also be active on within-component pairs).

Because we constrain $W_\ell \geq 0$, under the parameterization $W_\ell = A_\ell \otimes I_n + B_\ell \otimes J_n$, we must also have $B_\ell \geq 0$. Then, the appropriate notion of stationarity is KKT: in our setting, this reduces to

$$\nabla_{B_\ell}\mathcal{R}(\Theta) \geq 0, \qquad B_\ell \geq 0, \qquad \text{and} \qquad \nabla_{B_\ell}\mathcal{R}(\Theta) \odot B_\ell = 0$$

which we use in the Theorem below.

**Theorem C.5** (Population Training Conditionally Suppresses the $J$-Channel)**.** *Assume Assumption 4.8. Fix any layer $\ell$ and decompose $W_\ell = A_\ell \otimes I_n + B_\ell \otimes J_n$. Let $Z$ be the output, $R$ the reachability matrix (ground truth), $\mathcal{L} = \mathcal{L}(Z; R)$ the loss, and $\mathcal{R}(\Theta)$ the population risk.*

1. *(**Directional derivative on nonnegative $J$-channel directions.**) Let $\Delta \in \mathbb{R}^{K_{\ell-1}\times K_{\ell-1}}$ be entrywise nonnegative and define the one-sided Fréchet derivative $D := \frac{\partial Z}{\partial B_\ell}[\Delta] \in \mathbb{R}^{n\times n}$. Then $D \geq 0$ entrywise, and the population directional derivative satisfies*

$$D_{B_\ell}\mathcal{R}(\Theta)[\Delta] = \mathbb{E}\left\langle \left[\frac{\partial \mathcal{L}}{\partial Z}, D\right]\right\rangle_F = \alpha \cdot \mathbb{E}\Bigg[\underbrace{\sum_{R_{i,j}=0} D_{i,j}}_{\substack{\text{cross component}\\ \text{penalty}}} - \underbrace{\sum_{R_{i,j}=1} \frac{1-\phi_\epsilon(Z_{i,j})}{\phi_\epsilon(Z_{i,j})} D_{i,j}}_{\substack{\text{within-component}\\ \text{reward}}}\Bigg]. \tag{6}$$

   *In particular, $D_{B_\ell}\mathcal{R}(\Theta)[\Delta] \geq 0$ iff the Population-Level Dominance Condition holds (i.e. Equation (6) is positive). Throughout this Appendix, we will use "cross component penalty" and "within-component reward" to denote these two competing terms.*

2. *(**Consequences for KKT stationary points.**) Assume $\Theta$ is KKT-stationary for $B_\ell \geq 0$:*

$$\nabla_{B_\ell}\mathcal{R}(\Theta) \geq 0, \qquad B_\ell \geq 0, \qquad \textit{and} \qquad \nabla_{B_\ell}\mathcal{R}(\Theta) \odot B_\ell = 0 \tag{7}$$

   *Let $\Delta = |B_\ell|$ (entrywise absolute value) and let $D = \frac{\partial Z}{\partial B_\ell}[|B_\ell|]$. If, with positive probability under $\mathsf{ER}(n, p)$, $\Delta$ activates at least one cross-component pair, and if the Population-Level Dominance Condition holds, then $B_\ell = 0$. Equivalently, under activation at $\Delta = |B_\ell|$ and strict dominance by cross-component penalty, the only KKT stationary point in the $J_n$-channel is $B_\ell = 0$.*

**Lemma C.6** (Monotonicity in the $J$-channel)**.** *Fix $\ell$ and hold all parameters except $B_\ell$. Write $h_{\ell-1} = [X_1 \mid \ldots \mid X_{K_{\ell-1}}]$ and $u_p = X_p \mathbf{1} \in \mathbb{R}^n_{\geq 0}$. Then*

$$h_{\ell-1} W_\ell h_{\ell-1}^\top = \sum_{p,q} (A_\ell)_{p,q} X_p X_q^\top + \sum_{p,q} (B_\ell)_{p,q} u_p u_q^\top. \tag{8}$$

*Consequently, for every nonnegative direction $\Delta \geq 0$ in the $J$-channel, the one-sided Fréchet derivative at $0^+$ exists and is entrywise nonnegative. Hence, along the ray $\{B_\ell + \delta\Delta \mid \delta \geq 0\}$, the output is entrywise nondecreasing:*

$$\frac{\partial Z}{\partial B_\ell}[\Delta] \in \mathbb{R}^{n\times n}_{\geq 0}, \qquad Z(B_\ell + \delta\Delta) - Z(B_\ell) \geq 0 \text{ for all } \delta \geq 0.$$

*Moreover, if $G$ is disconnected, and either (i) $\Delta_{p,p} > 0$ for a block $p$ such that $u_p$ has support in at least two components, or (ii) there exist blocks $p, q$ with $\Delta_{p,q} > 0$ and $u_p, u_q$ supported in different components, then there exist cross component pairs $(i, j)$ with $(\frac{\partial Z}{\partial B_\ell}[\Delta])_{i,j} > 0$.*

*Proof.* Since $J_n x = (\mathbf{1}^\top x)\mathbf{1}$ for $x \in \mathbb{R}^n$, we have $X_p J_n X_q^\top = (X_p\mathbf{1})(X_q\mathbf{1})^\top = u_p u_q^\top$, yielding the displayed decomposition. For $B_\ell \mapsto B_\ell + \delta\Delta$ with $\Delta \geq 0$, the layer scores

$$R_\ell(B_\ell + \delta\Delta) - R_\ell(B_\ell) = \delta \sum_{p,q} \Delta_{p,q} u_p u_q^\top \geq 0,$$

so the one-sided derivative exists and is entrywise nonnegative. Because all subsequent maps are entrywise monotone, this implies $Z(B_\ell + \delta\Delta) - Z(B_\ell) \geq 0$ as stated.

For the "moreover" part, in casse (i), $u_p u_p^\top$ places positive mass on index pairs spnaning the componentns where $u_p > 0$, and in case (ii), $u_p u_q^\top$ (or its transpose) places positive mass across two components supporting $u_p$ and $u_q$. Monotonicity propagates these positives to $D = \frac{\partial Z}{\partial B_\ell}[\Delta]$. □

*Proof of Theorem C.5.* For the population risk $\mathcal{R}(\Theta) = \mathbb{E}[\mathcal{L}(Z; R)]$, applying definitions gives the directional derivative along $\Delta$ gives

$$D_{B_\ell}\mathcal{R}(\Theta)[\Delta] = \left\langle \mathbb{E}\Big[\frac{\partial \mathcal{L}}{\partial Z}\Big], D \right\rangle_F = \alpha \cdot \mathbb{E}\Big[\sum_{i,j}\left(1 - \frac{R_{i,j}}{\phi_\epsilon(Z_{i,j})}\right) D_{i,j}\Big].$$

Separating indices by $R_{i,j} \in \{0, 1\}$ proves equation 6.

For the second claim, evaluate equation 6 at $\Delta = |B_\ell|$. Under the activation premise (Lemma C.6) and strict dominance by cross-component penalty, we obtain $D_{B_\ell}\mathcal{R}(\Theta)[|B_\ell|] = \langle \nabla_{B_\ell}\mathcal{R}(\Theta), \Delta\rangle_F > 0$. Since $\nabla_{B_\ell}\mathcal{R}(\Theta) \geq 0$ and $|B_\ell| \geq 0$, a strictly positive inner product violates the KKT complementary condition $\nabla_{B_\ell}\mathcal{R}(\Theta) \odot B_\ell = 0$ unless $B_\ell = 0$. □

*Remark* C.7. While Theorem C.5 mostly discusses the suppression of $B_\ell$, its (i) in fact reveals a quite interesting, opposite phenomenon: **early training promotes** $B_\ell$. Before the model learns to pick up easy connected pairs, the corresponding values $\phi_\epsilon(Z_{i,j}) \ll 1$. Consequently, the fractions $(1 - \phi_\epsilon(Z_{i,j}))/\phi_\epsilon(Z_{i,j})$ are large, making equation 6 negative. Gradient descent then pushes $B_\ell$ up "without feeling pressure." As training proceeds, these easy connected pairs saturate ($\phi_\epsilon(Z_{i,j}) \to 1$), while simultaneously $\Delta$ begins to active cross pairs (the "moreover" part of Lemma C.6), increasing the $R = 0$ term in equation 6 and potentially flipping the sign. This is when the $J$-channel starts to incur penalty. This explains the transient "Phase 1" in §4.3.

*Remark* C.8. The $B_\ell \otimes J_n$-channel injects rank-one dense terms $u_p u_q^\top$ into the attention core. On disconnected graphs, these terms produce cross-component positives, which the reachability target $R$ labels as negatives. Because disconnected graphs appear with positive probability in the data, the population gradient penalizes every nonnegative direction in the $J$-channel active on cross-component pairs whenever the cross-component penalty dominates within-component reward. Under the same activation and cross-component penalty dominance assumptions, any KKT stationary point must have $B_\ell = 0$. In short: under these conditions, population drives the node-side factor towards locality, i.e., $W_\ell \approx A_\ell \otimes I_n$.

### C.4 Which Samples Push Which Channel? (Local $I_n$ vs. Global $J_n$)

Recall $W_\ell = A_\ell \otimes I_n + B_\ell \otimes J_n$ and Lemma C.6. The $I$-channel controls local propagation within components; the $J$-channel couples to the global / mean direction and injects dense rank-one terms. In this section, we first shift to a micro-level perspective, focusing on the effects of individual samples (graphs), and then draw connection to how the training distribution determines the model's eventual behavior (algorithmic vs. heuristic, §4.3).

We decompose the single-sample loss $\mathcal{L}_G(\Theta) := \mathcal{L}(\mathsf{TF}^L_\Theta(A); R)$ and examine directional derivatives at a fixed $\Theta$, with the link gradient $\partial\mathcal{L}/\partial Z = \alpha(1 - R/\phi_\epsilon(Z))$. Throughout, we say a pair $(i, j)$ is **saturated** if its per-pair loss gradient

vanishes; for within-component pairs ($R_{i,j} = 1$) this is equivalent to $\phi_\epsilon(Z_{i,j}) = R_{i,j}$. We say a direction $\Delta$ is **active** over $(i, j)$ if the correpsonding channel directive $D_{i,j} > 0$, where $D$ denotes $\frac{\partial Z}{\partial A_\ell}[\Delta]$ or $\frac{\partial Z}{\partial B_\ell}[\Delta]$ as appropriate.

Our first main result is the following Theorem, which intuitively claims two things:

- *(Within capacity) Small-diameter graphs "reward" the local $I$-channel and, if disconnected, penalizes the global $J$-channel if activated.*
- *(Beyond capacity) Large-diameter connected graphs demand a global shortcut: the $J$-channel is promoted, while the $I$-channel remains confined to short-range corrections.*

**Theorem C.9** (Per-sample pushes by diameter)**.** *Fix a layer $\ell$ and nonnegative directions $\Delta_A, \Delta_B \geq 0$ for $A_\ell, B_\ell$, respectively. Assume $B_1 = \ldots = B_L = 0$.*

*(i) (Within capacity) If $\operatorname{diam}(G) \leq 3^L$, then $D_{A_\ell}\mathcal{L}_G(\Theta)[\Delta_A] \leq 0$, with strict $< 0$ whenever $\Delta_A$ is active on at least one unsaturated within-component pair. If, in addition, $G$ is disconnected, then $D_{B_\ell}\mathcal{L}_G(\Theta)[\Delta_B] > 0$ if both of the following hold: $\Delta_B$ is active at at least one cross-component pair, and Population-Level Dominance Condition holds.*

*(ii) (Beyond capacity) If $\operatorname{diam}(G) > 3^L$ and $G$ is connected, then we have $D_{A_\ell}\mathcal{L}_G(\Theta)[\Delta_A] \leq 0$ where only within-capacity pairs can contribute, and $D_{B_\ell}\mathcal{L}_G(\Theta)[\Delta_B] < 0$ for $\Delta_B$ that is active on at least one unsaturated pair.*

To prove this Theorem, we split the argument into the following four lemmas, each isolating one ingredient of the dynamics. Firstly, Lemma C.10 shows that the local $I$-channel is monotone: any nonnegative $A_\ell$ cannot increase the loss and is strictly helpful on unsaturated within-component pairs. This lets us treat local corrections as "harmless," while Lemma C.11 analyze the sign of the global $J$-channel (connected vs. disconnected), and Lemma C.12 determines which pairs are ever affected when $B = 0$. Together, they yield the two cases in Theorem C.9.

**Lemma C.10** (Local channel always helps)**.** *Assume $B_1 = \ldots = B_L = 0$. For any graph $G$, any layer $\ell$, and any direction $\Delta \geq 0$ in the $I$-channel,*

$$D_{A_\ell}\mathcal{L}_G(\Theta)[\Delta] = \left\langle \frac{\partial \mathcal{L}}{\partial Z}, \frac{\partial Z}{\partial A_\ell}[\Delta] \right\rangle_F \leq 0, \tag{9}$$

*with strict inequality whenever there exists a within-component, unsaturated pair, on which $\Delta$ is active.*

*Proof.* From the block decomposition from equation 8, the $I$-channel contributes $\sum_{p,q} \Delta_{p,q} X_p X_q^\top$., which is block-diagonal with respect to the component partition. Hence $\frac{\partial Z}{\partial A_\ell}[\Delta]$ has support only on pairs $(i, j)$ in the same component. On those pairs, $R_{i,j} = 1$, and thus

$$\left(\frac{\partial \mathcal{L}}{\partial Z}\right)_{i,j} = \alpha \cdot \left(1 - \frac{1}{\phi_\epsilon(Z_{i,j})}\right) = -\alpha \cdot \frac{1 - \phi_\epsilon(Z_{i,j})}{\phi_\epsilon(Z_{i,j})} \leq 0,$$

with strict negativity whenever $\phi_\epsilon(Z_{i,j}) < 1$. Entrywise, nonnegativity of the forward map (Lemma C.6) gives $\frac{\partial Z}{\partial A_\ell}[\Delta] \geq 0$. Therefore the Frobenius inner product $\leq 0$, and $< 0$ under the stated conditions. □

We now switch from the local $I$-channel to the global $J$-channel and will use that the forward sensitivity in the $J$-channel is entrywise nonnegative, so the sign of the directional derivative is controlled entirely by the per-pair loss gradient.

**Lemma C.11** (Global channel helps connected graphs and conditionally hurts disconnected graphs)**.** *Fix a layer $\ell$ and a nonnegative direction $\Delta \geq 0$ in the $J$-channel.*

*(i) If $G$ is connected, then $D_{B_\ell}\mathcal{L}_G(\Theta) \leq 0$, with strict $< 0$ whenever there exists an unsaturated pair $(i, j)$ (i.e., $\phi_\epsilon(Z_{i,j}) < 1$) on which $\Delta$ is active ($D_{i,j} > 0$).*

*(ii) If $G$ is disconnected, then*

$$D_{B_\ell}\mathcal{L}_G(\Theta)[\Delta] = \alpha \cdot \left[ \sum_{R_{i,j}=0} D_{i,j} - \sum_{R_{i,j}=1} \frac{1 - \phi_\epsilon(Z_{i,j})}{\phi_\epsilon(Z_{i,j})} D_{i,j} \right], \tag{10}$$

*hence $D_{B_\ell}\mathcal{L}_G(\Theta)[\Delta] \geq 0$ whenever the Population-Level Dominance Condition holds. Strict $> 0$ holds if the inequality is strict, and $\Delta$ is active on at least one cross pair.*

*Proof.* By the chain rule,

$$D_{B_\ell}\mathcal{L}_G(\Theta)[\Delta] = \left\langle \frac{\partial \mathcal{L}}{\partial Z}, \frac{\partial Z}{\partial B_\ell}[\Delta] \right\rangle_F = \left\langle \alpha \cdot \left(1 - \frac{R}{\phi_\epsilon(Z)}\right), D \right\rangle_F . \tag{11}$$

By Lemma C.6 , $D \geq 0$ entrywise; moreover, $D_{i,j} > 0$ exactly on pairs where $\Delta$ is active.

(i) If $G$ is connected, then the reachability matrix $R$ is all-ones. Hence $\frac{\partial \mathcal{L}}{\partial Z} = -\alpha(1-\phi_\epsilon(Z))/\phi_\epsilon(Z) \leq 0$ entrywise, with strict negativity whenever $\phi_\epsilon(Z_{i,j}) < 1$. Pairing with $D \geq 0$ and $D_{i,j} > 0$ on active pairs gives $D_{B_\ell}\mathcal{L}_G(\Theta)[\Delta] \leq 0$ and strict $< 0$ under the stated saturation / activation conditions.

(ii) If $G$ is disconnected, split equation 11 over $R_{i,j} = 0$ and $R_{i,j} = 1$ to obtain the displayed identity. Since $D \geq 0$, the stated dominance condition yields $\geq 0$. Strictness requires a cross pair with $D_{i,j} > 0$, holds exactly when $\Delta$ is active on at least one cross-component pair. □

We now show that when the global $J$-channel is disabled, the model can only light up within-capacity pairs. Note this is somewhat a converse to Theorem C.2, where a "good" model that only lights up within-capacity pairs necessarily have each $W_\ell[r,s]$ diagonal. The following Lemma isolates the role of the $J$-channel as the only "nontrivial" shortcut.

Recall from Definition 4.6: for a depth $L$ and a graph $G$ with adjacency matrix $A$, we call a pair $(i,j)$ **within capacity** if $[A^{3^L}]_{i,j} > 0$ and **beyond capacity** otherwise.

**Lemma C.12** ($I$-channel reaches within-capacity pairs; $J$-channel is the only dense shortcut)**.** *At any $\Theta$ with $B_1 = \ldots = B_L = 0$, the output satisfies*

$$Z_{i,j} > 0 \implies [A^{3^L}]_{i,j} > 0.$$

*Equivalently, beyond-capacity pairs receive no positive mass from the $I$-channel alone. In contrast, for any $\ell$ and any $\Delta \geq 0$ in the $J$-channel, $\frac{\partial Z}{\partial B_\ell}[\Delta] \geq 0$ and is strictly positive on active pairs by definition.*

*Proof.* Since $B_\ell = 0$ implies $h_{\ell-1}W_\ell h_{\ell-1}^\top = \sum_{p,q}(A_\ell)_{p,q}X_pX_q^\top$ from Lemma C.6, it is easy to see that they are block-diagonal w.r.t. connected components. Hence Lemma B.2 applies and support expands by at most a factor of 3 per layer, and only within-capacity pairs receive mass. The density statement follows from Lemma C.6: For any $\Delta \geq 0$ in the $J$-channel, we have $\frac{\partial Z}{\partial B_\ell} = \sum_{p,q}\Delta_{p,q}u_pu_q^\top u \geq 0$. The strict positiveness characterization follows directly from Lemma C.6. □

With the previous lemmas established, we can now assemble the per-sample sign rules. Intuitively, the $I$-channel makes only local corrections, never hurting the loss and only touching within-capacity pairs when $B = 0$, while the $J$-channel is the sole dense shortcut, helpful on connected graphs but penalized by cross-component pairs when the graph is disconnected.

*Proof of Theorem C.9.* Let $\ell, \Delta_A, \Delta_B$ be given as described. Set $D_A = \frac{\partial Z}{\partial A_\ell}[\Delta_A]$ and $D_B = \frac{\partial Z}{\partial B_\ell}[\Delta_B]$. Recall from chain rule

$$D_{(\cdot)}\mathcal{L}_G(\Theta)[\cdot] = \left\langle \frac{\partial \mathcal{L}}{\partial Z}, \frac{\partial Z}{\partial (\cdot)}[\cdot] \right\rangle_F = \alpha \cdot \left\langle 1 - \frac{R}{\phi_\epsilon(Z)}, \frac{\partial Z}{\partial (\cdot)}[\cdot] \right\rangle_F .$$

(i) (Within capacity) By Lemma C.10, for any $\Delta_A \geq 0$ the $I$-channel directional derivative is $\leq 0$, with strict inequality under the stated conditions. The result on disconnected graphs $G$ follows from Lemma C.11.

(ii) By Lemma C.12, with $B = 0$, only within-capacity pairs can be affected by the $I$-channel, so Lemma C.10 gives $D_{A_\ell}\mathcal{L}_G(\Theta)[\Delta_A] \leq 0$. Since $G$ is connected and $\mathrm{diam}(G) > 3^L$, there will be unsaturated pairs; then Lemma C.11(i) yields $D_{B_\ell}\mathcal{L}_G(\Theta)[\Delta_B] < 0$, as claimed. □

*Remark* C.13 (Population-level consequence under $\mathrm{ER}(n,p)$). Fix a layer $\ell$ and nonnegative directions $\Delta_A, \Delta_B \geq 0$. Partition the graphs into $\mathcal{G}_0 = \{G : \mathrm{diam}(G) \leq 3^L\}$ and $\mathcal{G}_1 = \{G : \mathrm{diam}(G) > 3^L\}$. Writing the population directional derivatives as mixtures,

$$D_{B_\ell}\mathcal{R}(\Theta)[\Delta_B] = \mathbb{P}(\mathcal{G}_0)\mathbb{E}[D_{B_\ell}\mathcal{L}_G(\Theta)[\Delta_B] \mid G \in \mathcal{G}_0] + \mathbb{P}(\mathcal{G}_1)\mathbb{E}[D_{B_\ell}\mathcal{L}_G(\Theta)[\Delta_B] \mid G \in \mathcal{G}_1]. \tag{12}$$

We claim the following on the population gradient.

(i) (Local) From Lemma C.10, once the global $J$-channel has been suppressed, the local $I$-channel is consistently promoted until saturation.

(ii) (Global) The population gradient along the global $J$-channel is an explicit mixture of two regimes: large, connected graphs beyond capacity that promote the $J$-channel, and small, disconnected graphs within capacity that suppress it whenever cross-component errors persist. Formally:

(ii.a) If $G$ is connected and $\mathrm{diam}(G) > 3^L$, then by Lemma C.12, every beyond-capacity pair has $Z_{ij} = 0$ while $R_{ij} = 1$. For those pairs, we have $\partial\mathcal{L}/\partial Z = -\alpha(1-\phi_\epsilon(Z))/\phi_\epsilon(Z) < 0$. By Lemma C.6, the inner product $\langle \partial\mathcal{L}/\partial Z, \partial Z/\partial B_\ell[\Delta_B]\rangle_F < 0$ too. Integrating over all beyond-capacity, connected graphs yields

$$\mathbb{E}[D_{B_\ell}\mathcal{L}_G(\Theta)[\Delta_B] \mid G \in \mathcal{G}_1 \;\&\; G \text{ connected}] < 0. \tag{13}$$

(ii.b) If $G$ is disconnected and $\mathrm{diam}(G) \le 3^L$, then by Lemma C.11, $D_{B_\ell}\mathcal{L}_G(\Theta)[\Delta_B] \ge 0$ with strict $> 0$ if cross-component errors persist (the $\sum_{R=0} D$ term strictly dominates the $\sum_{R=1}(1-\phi_\epsilon(Z))/\phi_\epsilon(Z)\cdot D$ term), and if $\Delta_B$ is active on cross pairs (i.e. $D^{(B)}_{ij} > 0$ for some $R_{ij} = 0$). The latter holds by Lemma C.6 if $\Delta_B$ is active on at least one cross pair. Integrating thus yields

$$\mathbb{E}\big[D_{B_\ell}\mathcal{L}_G(\Theta)[\Delta_B] \,\big|\, G \in \mathcal{G}_0 \;\&\; G \text{ disconnected}\big] \ge 0, \tag{14}$$

and strictly $>0$ provided the two additional assumptions above.

### C.5 Convergence of Projected Gradient Descent to KKT Points

In this subsection, we establish that projected gradient descent on the regularized population risk converges to points satisfying the Karush-Kuhn-Tucker (KKT) conditions. This justifies the stationarity assumption underlying Theorem C.5. We begin by stating the main result.

**Theorem 4.9.** *Let $\mathcal{R}(\Theta) := \mathbb{E}_{G\sim\mathsf{ER}(n,p)}[\mathcal{L}(\mathsf{TF}^L_\Theta(A_G); R_G)]$ denote the population risk. For $\lambda > 0$, define the regularized objective $\mathcal{R}_\lambda(\Theta) := \mathcal{R}(\Theta) + \frac{\lambda}{2}\|\Theta\|_F^2$. Let $\mathcal{C} := \{(A_\ell, B_\ell)_\ell : A_\ell \ge 0, B_\ell \ge 0, \forall \ell\}$ denote the constraint set, and consider the sequence $\{\Theta^{(k)}\}_{k\ge 0}$ generated by projected gradient descent on $\mathcal{R}_\lambda$:*

$$\Theta^{(k+1)} = \Pi_{\mathcal{C}}\left(\Theta^{(k)} - \eta\nabla\mathcal{R}_\lambda(\Theta^{(k)})\right), \tag{15}$$

*with step size $\eta > 0$ sufficiently small and initialization $\Theta^{(0)} \in \mathcal{C}$ of the form $W_\ell = A_\ell \otimes I + B_\ell \otimes J$. Then every limit point $\Theta^*_\lambda \in \mathcal{C}$ satisfies the KKT conditions:*

$$\nabla_{B_\ell}\mathcal{R}(\Theta^*_\lambda) + \lambda B^*_\ell \ge 0, \quad B^*_\ell \ge 0, \quad (\nabla_{B_\ell}\mathcal{R}(\Theta^*_\lambda) + \lambda B^*_\ell)\odot B^*_\ell = 0, \tag{16}$$

*and analogously for $A^*_\ell$. Moreover, after $K$ iterations, there exists $k < K$ such that $\Theta^{(k)}$ is an $\epsilon$-approximate KKT point with $\epsilon = O(1/K)$.*

The proof adapts the standard convergence analysis for gradient descent on smooth nonconvex functions (Nesterov, 2004) to the projected setting. The argument proceeds in three stages: we first introduce the gradient mapping as the appropriate measure of stationarity for constrained problems, then establish a sufficient decrease property for each iteration, and finally combine these ingredients via a telescoping argument to obtain the convergence rate.

#### C.5.1 Preliminaries

We work in the decomposed parameter space $\Theta = (A_\ell, B_\ell)_\ell$, where $W_\ell = A_\ell \otimes I + B_\ell \otimes J$ as in Assumption 4.8. By Theorem C.4, if the initialization lies in this subalgebra, then the gradient $\nabla\mathcal{R}_\lambda(\Theta)$ also decomposes as a direct sum over the $(A_\ell, B_\ell)$ components, and projected gradient descent preserves this structure. The constraint set $\mathcal{C} = \{(A_\ell, B_\ell)_\ell : A_\ell \ge 0, B_\ell \ge 0\}$ is the non-negative orthant in this decomposition, and the projection $\Pi_{\mathcal{C}}$ acts component-wise as $\Pi_{\mathcal{C}}(\Theta) = \max\{0, \Theta\}$.

Recall that a point $\Theta^* \in \mathcal{C}$ satisfies the KKT conditions for minimizing $\mathcal{R}_\lambda$ over $\mathcal{C}$ if

$$\nabla\mathcal{R}_\lambda(\Theta^*) \ge 0, \quad \Theta^* \ge 0, \quad \text{and} \quad \nabla\mathcal{R}_\lambda(\Theta^*)\odot\Theta^* = 0, \tag{17}$$

where $\odot$ denotes the Hadamard product and inequalities hold component-wise. These conditions assert that interior components (where $\Theta^* > 0$) have vanishing gradient, while boundary components (where $\Theta^* = 0$) have non-negative gradient pointing outward from the feasible region.

In unconstrained optimization, the gradient norm $\|\nabla f(\Theta)\|$ measures proximity to stationarity. For constrained problems, the appropriate generalization is the gradient mapping.

**Definition C.14** (Gradient Mapping)**.** For step size $\eta > 0$, the gradient mapping at $\Theta \in \mathcal{C}$ is

$$G_\eta(\Theta) := \frac{1}{\eta}\left(\Theta - \Pi_{\mathcal{C}}\left(\Theta - \eta \nabla \mathcal{R}_\lambda(\Theta)\right)\right). \tag{18}$$

The gradient mapping quantifies the displacement induced by a projected gradient step: the update rule can be written as $\Theta^{(k+1)} = \Theta^{(k)} - \eta G_\eta(\Theta^{(k)})$. When $\mathcal{C}$ is unconstrained, the projection is the identity and $G_\eta(\Theta) = \nabla \mathcal{R}_\lambda(\Theta)$. The following lemma confirms that the gradient mapping vanishes precisely at KKT points.

**Lemma C.15.** *For any $\eta > 0$, a point $\Theta^* \in \mathcal{C}$ satisfies $G_\eta(\Theta^*) = 0$ if and only if $\Theta^*$ is a KKT point.*

*Proof.* The condition $G_\eta(\Theta^*) = 0$ is equivalent to $\Theta^* = \Pi_{\mathcal{C}}(\Theta^* - \eta \nabla \mathcal{R}_\lambda(\Theta^*))$. For the non-negative orthant, this becomes

$$\Theta^* = \max\left\{0, \Theta^* - \eta \nabla \mathcal{R}_\lambda(\Theta^*)\right\}. \tag{19}$$

On the interior $\{\Theta^* > 0\}$, equality requires $\nabla \mathcal{R}_\lambda(\Theta^*) = 0$. On the boundary $\{\Theta^* = 0\}$, the condition reduces to $0 = \max\{0, -\eta \nabla \mathcal{R}_\lambda(\Theta^*)\}$, which holds if and only if $\nabla \mathcal{R}_\lambda(\Theta^*) \geq 0$. These are precisely the KKT conditions. □

### C.5.2 Regularity of the Objective

The convergence analysis requires two properties of the regularized objective: coercivity, which ensures that iterates remain bounded, and smoothness, which enables a descent inequality.

**Lemma C.16** (Coercivity)**.** *For any $\lambda > 0$, the sublevel sets of $\mathcal{R}_\lambda$ are bounded: if $\mathcal{R}_\lambda(\Theta) \leq c$, then $\|\Theta\|_F \leq \sqrt{2c/\lambda}$.*

*Proof.* Since $\mathcal{R}(\Theta) \geq 0$, we have $\mathcal{R}_\lambda(\Theta) \geq \frac{\lambda}{2}\|\Theta\|_F^2$. The bound follows by rearrangement. □

Coercivity is essential: without regularization, the scaling symmetry of ReLU networks ($W_\ell \mapsto \alpha W_\ell$, $W_{\ell+1} \mapsto \alpha^{-1} W_{\ell+1}$) renders the sublevel sets unbounded, and iterates could escape to infinity.

**Lemma C.17** (Lipschitz Gradient)**.** *Under Assumption 4.8, for any bounded set $\mathcal{B} \subseteq \mathcal{C}$, there exists $L_{\mathcal{B}} > 0$ such that*

$$\|\nabla \mathcal{R}_\lambda(\Theta_1) - \nabla \mathcal{R}_\lambda(\Theta_2)\| \leq L_{\mathcal{B}} \|\Theta_1 - \Theta_2\| \quad \text{for all } \Theta_1, \Theta_2 \in \mathcal{B}. \tag{20}$$

*In particular, $\nabla \mathcal{R}_\lambda$ is Lipschitz continuous on the sublevel set $\{\Theta \in \mathcal{C} : \mathcal{R}_\lambda(\Theta) \leq \mathcal{R}_\lambda(\Theta^{(0)})\}$.*

*Proof.* Within the constraint set $\mathcal{C}$, ReLU activations act as the identity. Since the adjacency matrix $A_G \geq 0$ and $W_\ell = A_\ell \otimes I + B_\ell \otimes J \geq 0$ for $(A_\ell, B_\ell) \in \mathcal{C}$, induction shows $h_{\ell-1} W_\ell h_{\ell-1}^\top \geq 0$ at every layer, so ReLU is the identity throughout the forward pass. The transformer output is therefore a polynomial in $\Theta$, the loss function is smooth (with $\phi_\epsilon \geq \epsilon > 0$ keeping the logarithm away from zero), and the regularization is quadratic. The composition is smooth on $\mathcal{C}$, and smooth functions have Lipschitz gradients on bounded sets. The second statement follows from Lemma C.16, which ensures the sublevel set is bounded. □

### C.5.3 Sufficient Decrease

The core of the analysis is a progress bound showing that each projected gradient step decreases the objective by an amount proportional to the squared gradient mapping norm.

**Lemma C.18** (Sufficient Decrease)**.** *For step size $\eta \leq 1/L$, the iterates satisfy*

$$\mathcal{R}_\lambda(\Theta^{(k+1)}) \leq \mathcal{R}_\lambda(\Theta^{(k)}) - \frac{\eta}{2}\|G_\eta(\Theta^{(k)})\|^2. \tag{21}$$

*Proof.* Let $\Theta = \Theta^{(k)}$, $\Theta^+ = \Theta^{(k+1)}$, $g = \nabla \mathcal{R}_\lambda(\Theta)$, and $G = G_\eta(\Theta)$. The descent lemma for $L$-smooth functions gives

$$\mathcal{R}_\lambda(\Theta^+) \leq \mathcal{R}_\lambda(\Theta) + \langle g, \Theta^+ - \Theta \rangle + \frac{L}{2}\|\Theta^+ - \Theta\|^2. \tag{22}$$

Since $\Theta^+ - \Theta = -\eta G$, this becomes

$$\mathcal{R}_\lambda(\Theta^+) \leq \mathcal{R}_\lambda(\Theta) - \eta \langle g, G \rangle + \frac{L\eta^2}{2}\|G\|^2. \tag{23}$$

It remains to show that $\langle g, G\rangle \geq \|G\|^2$. The projection $\Theta^+ = \Pi_{\mathcal{C}}(\Theta - \eta g)$ satisfies the first-order optimality condition: for all $z \in \mathcal{C}$,

$$\langle (\Theta - \eta g) - \Theta^+, z - \Theta^+\rangle \leq 0. \tag{24}$$

Taking $z = \Theta \in \mathcal{C}$ and using $\Theta - \Theta^+ = \eta G$ yields $\langle \eta G - \eta g, \eta G\rangle \leq 0$, which simplifies to $\langle g, G\rangle \geq \|G\|^2$. Substituting into equation 23 and using $\eta \leq 1/L$ completes the proof. □

### C.5.4 Proof of **Theorem** 4.9

*Proof.* We establish each component of the theorem in turn.

**Bounded iterates.**The sufficient decrease property (Lemma C.18) implies that $\mathcal{R}_\lambda(\Theta^{(k)})$ is non-increasing, so all iterates lie in the initial sublevel set. By Lemma C.16, this set is bounded: $\|\Theta^{(k)}\|_F \leq \sqrt{2\mathcal{R}_\lambda(\Theta^{(0)})/\lambda}$ for all $k$.

**Non-asymptotic rate.**Rearranging Lemma C.18 and summing from $k = 0$ to $K - 1$:

$$\sum_{k=0}^{K-1} \|G_\eta(\Theta^{(k)})\|^2 \leq \frac{2}{\eta} \sum_{k=0}^{K-1} \left[\mathcal{R}_\lambda(\Theta^{(k)}) - \mathcal{R}_\lambda(\Theta^{(k+1)})\right] = \frac{2}{\eta}\left[\mathcal{R}_\lambda(\Theta^{(0)}) - \mathcal{R}_\lambda(\Theta^{(K)})\right]. \tag{25}$$

The right-hand side telescopes. Since $\mathcal{R}_\lambda \geq 0$, we obtain

$$\sum_{k=0}^{K-1} \|G_\eta(\Theta^{(k)})\|^2 \leq \frac{2\mathcal{R}_\lambda(\Theta^{(0)})}{\eta}. \tag{26}$$

The minimum of $K$ non-negative terms is at most their average:

$$\min_{k=0,\ldots,K-1} \|G_\eta(\Theta^{(k)})\|^2 \leq \frac{2\mathcal{R}_\lambda(\Theta^{(0)})}{\eta K}. \tag{27}$$

Thus, within $K$ iterations, at least one iterate achieves $\|G_\eta(\Theta^{(k)})\|^2 \leq \epsilon$ for $\epsilon = O(1/K)$.

**Asymptotic convergence.**Taking $K \to \infty$, the bound $\sum_{k=0}^{\infty} \|G_\eta(\Theta^{(k)})\|^2 < \infty$ implies $\|G_\eta(\Theta^{(k)})\| \to 0$.

**Limit points are KKT.**Boundedness of the iterates follows from Lemma C.16. That every limit point of projected gradient descent on a smooth function over a closed convex set satisfies the first-order stationarity condition is a standard result; see, e.g., Bertsekas (1997) (Proposition 2.3.2) or Beck (2017) (Theorem 10.15). The required smoothness holds by Lemma C.17. Restricting to the $B_\ell$ components and expanding $\nabla\mathcal{R}_\lambda = \nabla\mathcal{R} + \lambda\Theta$ yields equation 16. □

*Remark* C.19 (Sufficiency for Main Results). **Theorem** 4.9 establishes convergence to KKT points in the sense of limit points; it does not rule out the possibility that the sequence oscillates between multiple KKT points. Establishing convergence to a unique limit requires additional structure, such as the Kurdyka-Łojasiewicz property (Attouch et al., 2013). However, for our purposes, convergence to limit points suffices: Theorem C.5 shows that *any* KKT point satisfying our assumptions has $B_\ell^* = 0$, so the heuristic channel is suppressed regardless of which limit point is approached.

### C.5.5 Connecting Regularized Convergence to Heuristic Suppression

Theorem 4.9 establishes that projected gradient descent on the regularized objective $\mathcal{R}_\lambda(\Theta) = \mathcal{R}(\Theta) + \frac{\lambda}{2}\|\Theta\|_F^2$ converges to stationary points satisfying the KKT conditions. To complete the picture, we must link this convergence guarantee back to the gradient properties derived in Theorem C.5, which analyzed the unregularized population risk $\mathcal{R}(\Theta)$.

We now demonstrate that the regularization term $\frac{\lambda}{2}\|\Theta\|_F^2$ works in tandem with the data-driven suppression mechanism. Specifically, for any nonnegative direction $\Delta \geq 0$ in the $J_n$-channel, the directional derivative of the regularized objective satisfies:

$$D_{B_\ell}\mathcal{R}_\lambda(\Theta)[\Delta] = \underbrace{D_{B_\ell}\mathcal{R}(\Theta)[\Delta]}_{\text{Population Risk Gradient}} + \underbrace{\lambda\langle B_\ell, \Delta\rangle_F}_{\text{Regularization Penalty}}. \tag{28}$$

When the cross-component penalty dominates the within-component reward (as characterized in Theorem C.5), the population risk gradient is already nonnegative ($D_{B_\ell}\mathcal{R}(\Theta)[\Delta] \geq 0$). Since $\lambda > 0$ and $B_\ell \geq 0$, the regularization term is also nonnegative, and strictly positive whenever $B_\ell \neq 0$. Consequently, the regularization only strengthens the inequality, ensuring that the heuristic channel is suppressed at any stationary point.

**Corollary C.20** (Regularized KKT Points Suppress the Heuristic Channel)**.** *Assume the setting of Assumption 4.8. Let $\lambda > 0$ and let $\Theta^*_\lambda = (A^*_\ell, B^*_\ell)_{\ell=1}^L$ be a KKT point of the regularized objective $\mathcal{R}_\lambda(\Theta)$ over the constraint set $\mathcal{C}$. Fix a layer $\ell \in \{1, \dots, L\}$ and assume the Population-Level Dominance Condition (cf. Equation (6)) holds at $\Theta^*_\lambda$:*

$$\mathbb{E}\left[\sum_{R_{ij}=0} D_{ij}\right] > \mathbb{E}\left[\sum_{R_{ij}=1} \frac{1-\phi_\epsilon(Z_{ij})}{\phi_\epsilon(Z_{ij})} D_{ij}\right], \tag{29}$$

*where $D = \frac{\partial Z}{\partial B_\ell}[B^*_\ell]$ is the Jacobian along the direction of the learned weights $B^*_\ell$. Then the heuristic channel is fully suppressed: $B^*_\ell = 0$.*

*Proof.* We proceed by contradiction. Suppose $B^*_\ell \neq 0$.

Since $\Theta^*_\lambda$ is a KKT point of $\mathcal{R}_\lambda$ over $\mathcal{C}$, the first-order optimality conditions for the non-negative parameter $B^*_\ell$ require

$$(\nabla_{B_\ell} \mathcal{R}(\Theta^*_\lambda) + \lambda B^*_\ell) \odot B^*_\ell = 0. \tag{30}$$

Summing over all entries (taking the Frobenius inner product with $B^*_\ell$), we obtain:

$$\langle \nabla_{B_\ell} \mathcal{R}(\Theta^*_\lambda), B^*_\ell \rangle_F + \lambda \|B^*_\ell\|_F^2 = 0. \tag{31}$$

We now analyze the population gradient term $\langle \nabla_{B_\ell} \mathcal{R}(\Theta^*_\lambda), B^*_\ell \rangle_F$. By Theorem C.5, this term decomposes into the difference between the expected penalty on disconnected graphs and the expected reward on connected graphs. By the Population-Level Dominance Condition assumed in the Corollary statement, the expected penalty strictly exceeds the expected reward. Therefore, the unregularized gradient contribution is strictly positive:

$$\langle \nabla_{B_\ell} \mathcal{R}(\Theta^*_\lambda), B^*_\ell \rangle_F > 0. \tag{32}$$

Intuitively, this means the data distribution itself is pushing the weights $B^*_\ell$ toward zero.

Substituting this into equation 31, we arrive at a contradiction:

$$\underbrace{\langle \nabla_{B_\ell} \mathcal{R}(\Theta^*_\lambda), B^*_\ell \rangle_F}_{>0} + \underbrace{\lambda \|B^*_\ell\|_F^2}_{>0} = 0. \tag{33}$$

Both terms on the left-hand side are strictly positive (the second term because $\lambda > 0$ and we assumed $B^*_\ell \neq 0$). Their sum cannot be zero. Thus, we must have $B^*_\ell = 0$. □

*Remark* C.21 (Role of Regularization). The regularization parameter $\lambda > 0$ plays a dual role. First, it ensures coercivity (Lemma C.16), which is necessary to prove the existence of limit points for the training dynamics in Theorem 4.9. Second, as shown in Corollary C.20, it acts as a strict enforcer of suppression: even if the data-driven gradient were merely zero (a "tie" between penalty and reward), the regularization force $\lambda B^*_\ell$ would still drive the weights to zero via the stationarity condition.

**Corollary C.22** (Convergence to Algorithmic Solutions)**.** *Under Assumption 4.8, let $\{\Theta^{(k)}\}_{k\geq 0}$ be the sequence generated by projected gradient descent on $\mathcal{R}_\lambda$ with step size $\eta \leq 1/L$ and initialization $\Theta^{(0)} \in \mathcal{C}$. Suppose that at every limit point $\Theta^*$, the Population-Level Dominance Condition from Corollary C.20 holds for all layers $\ell$.*

*Then every limit point satisfies $B^*_\ell = 0$ for all $\ell$, and consequently $W^*_\ell = A^*_\ell \otimes I_n$. The model at any limit point implements the algorithmic $I_n$-channel exclusively and reaches its theoretical capacity of $3^L$.*

*Proof.* Theorem 4.9 guarantees that every limit point $\Theta^*_\lambda$ satisfies the KKT conditions for $\mathcal{R}_\lambda$ over $\mathcal{C}$. Applying Corollary C.20 to each layer yields $B^*_\ell = 0$. With $B^*_\ell = 0$, the heuristic channel is eliminated. The capacity statement then follows from Lemma C.12, which establishes that when $B_1 = \cdots = B_L = 0$, the model output satisfies $Z_{ij} > 0$ only if $[A^{3^L}]_{ij} > 0$, achieving the tight capacity bound of Theorem 4.5. □

# D Additional Experiment Details and Results

## D.1 Experiment Details

**Standard Transformers.**When training 2-layer standard Transformers, we adopt the implementation from RoBERTa (Liu et al., 2019) with single-head per-layer and using normalized ReLU activation function as defined in Definition A.1. We use a hidden dimension of $d = 512$ to make sure the hidden size is not the blocker for expressivity. We trained on 1 Billion ER graphs with a batch size of 1000 and $10^6$ steps. Each graph is only seen by the model once to resembling the training regime of modern LLMs. We note that although 1 billion graphs sounds a lot but with $n = 20$ nodes, this is far from enumerating all possible graphs: there can be $2^{\binom{n}{2}}$ graphs if we don't consider graph isomorphism. When $n = 20$, this is about more than $10^{57}$ graphs in total, and 1 billion ($10^9$) is only a very small number of training instances. We train with AdamW optimizer with a learning rate of `1e-4` and weight decay of `1e-4` and a cosine learning rate decay.

**Disentangled Transformers.** For 1-layer Disentangled Transformers in Section 5, we train on a fixed set 4096 i.i.d. samples of $\mathsf{ER}(n = 8)$ graphs and running standard *Gradient Descent* without any mini-batching. In this case, we have a learning rate of $0.1$ with cosine learning rate decay. For 2-layer Disentangled Transformers, we train on the same set of 1 billion number of $\mathsf{ER}(n = 20)$ graphs as with standard Transformers. For 3-layer models, we train on 1 billion number of $\mathsf{ER}(n = 64)$ graphs. Both 2- and 3-layer models are trained with AdamW with a learning rate of `1e-3`. We would like to note that the hidden dimensions $d_\ell$ of Disentangled Transformerss are fixed to be $d_\ell = 2^\ell n$ rather than a hyper-parameter (see Definition 4.1).

**Computing Energy Share of $I_n$/$J_n$ Channels.** In the experiments on 1-Layer Disentangled Transformers, we compute energy shares of the $A \otimes I_n$ and $B \otimes J_n$ within $\|W\|_F^2$. Here is the formalized versions. We consider the noisy decomposition $W = \hat{A} \otimes I_n + \hat{B} \otimes J_n + W_\epsilon$, where $W_\epsilon$ is the projection error term. We define Frobenius-norm energy share on the $I_n$ channel as

$$\mathsf{EnergyShare}(\hat{A} \otimes I_n, W) = \frac{\langle W, \hat{A} \otimes I_n \rangle}{\|W\|_F^2} = \frac{\|\hat{A} \otimes I_n\|_F^2 + \langle \hat{A} \otimes I_n, \hat{B} \otimes J_n \rangle + \langle \hat{A} \otimes I_n, W_\epsilon \rangle}{\|W\|_F^2},$$

and by symmetry, the $J_n$-channel share is

$$\mathsf{EnergyShare}(\hat{B} \otimes J_n, W) = \frac{\langle W, \hat{B} \otimes J_n \rangle}{\|W\|_F^2} = \frac{\|\hat{B} \otimes J_n\|_F^2 + \langle \hat{B} \otimes J_n, \hat{A} \otimes I_n \rangle + \langle \hat{B} \otimes J_n, W_\epsilon \rangle}{\|W\|_F^2}.$$

This is a well-designed quantity because if you expand $\|W\|_F^2$ you obtain $\langle W, \hat{A} \otimes I_n + \hat{B} \otimes J_n + W_\epsilon \rangle$, and the $I/J$-channels' energy shares will sum to one when the projection error $W_\epsilon$ converges to zero.

## D.2 Additional Experiments on Disentangled and Standard Transformers

In Figure 8, we show the training dynamics of a 3-Layer Disentangled Transformer. In Figure 9, we show the learned weights by Disentangled Transformers.

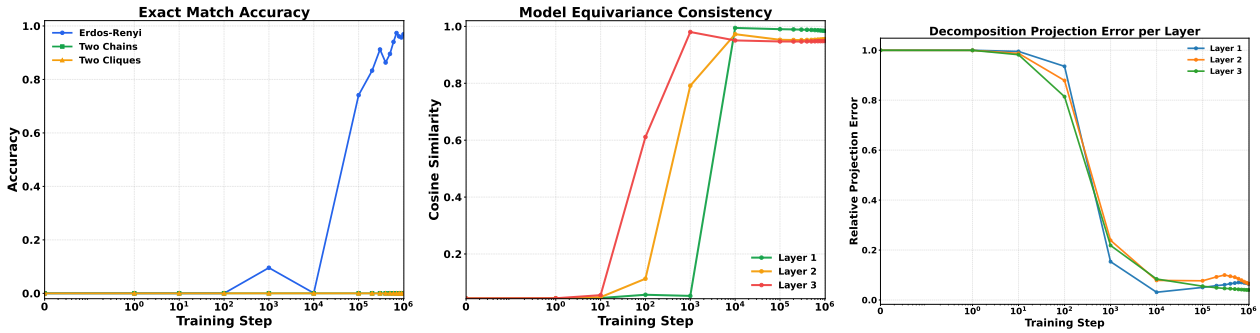



*Figure 8.* We plot the model behavior of a 3-Layer Disentangled Transformer model trained on $\mathsf{ER}(n = 64)$ graphs. They also quickly pick up almost *layer-wise equivariant* properties (measured by Eqn. 3). All layers show very small projection error onto the $A \otimes I_n + B \otimes J_n$ decomposition, resonating our theoretical claims in Theorem 4.7.

In Figure 9, we show that the trained 2-layer and 3-layer converge to weight spaces $W_\ell = A_\ell \otimes I_n + B_\ell \otimes J_n$ in the particular form echoing Theorem 4.7.

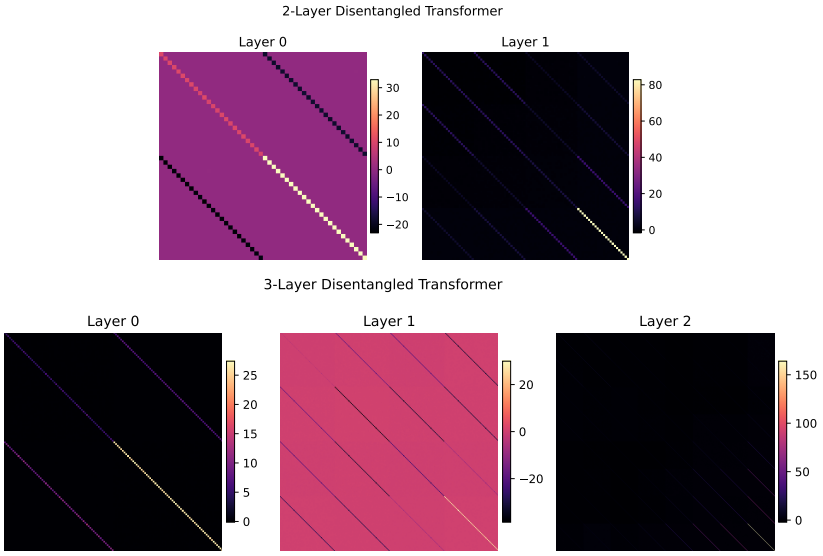



*Figure 9.* Here we visualize the weights $W_\ell$ learned by a 2-Layer and 3-Layer Disentangled Transformer respectively. All models are randomly initialized **without** any restriction on parameterization. Resonating Theorem 4.7, they all converge to a form of $W_\ell = A_\ell \otimes I_n + B_\ell \otimes J_n$.

In Figure 10, we show that the capacity theorems (Theorem 4.5) also transfer to standard 2-layer Transformer models.

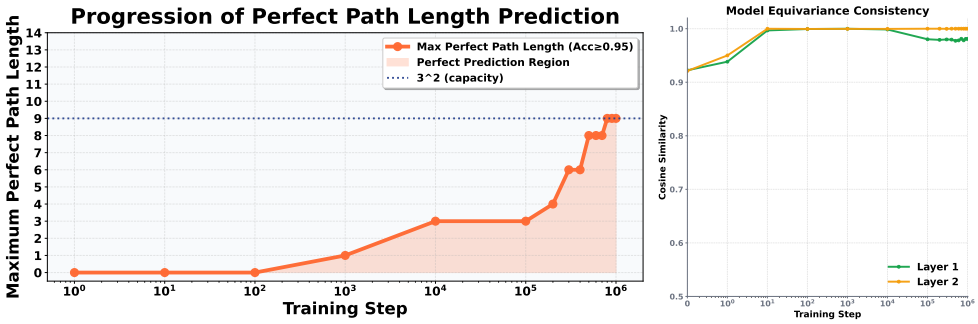



*Figure 10.* (**left**) Standard Transformers models studied in §3.3 also hit its capacity wall at $3^L$, showing that our theoretical results transfer beyond the theoretical simplification of Disentangled Transformers. (**right**) Standard Transformer models also learn an almost layer-wise equivariant solution measured by Eqn. 3.

In Figure 11, we show that when evaluated in 2Clique dataset, the one trained with the right data generalize better.

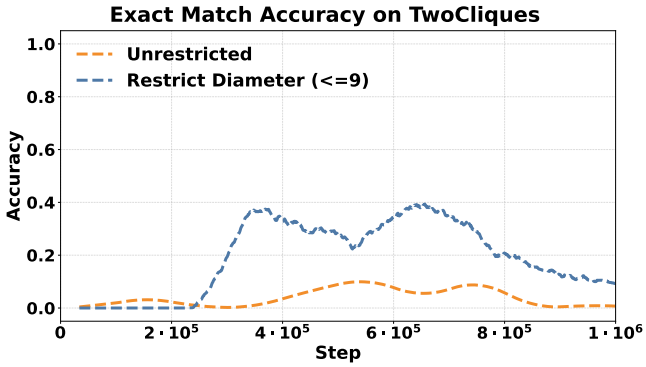
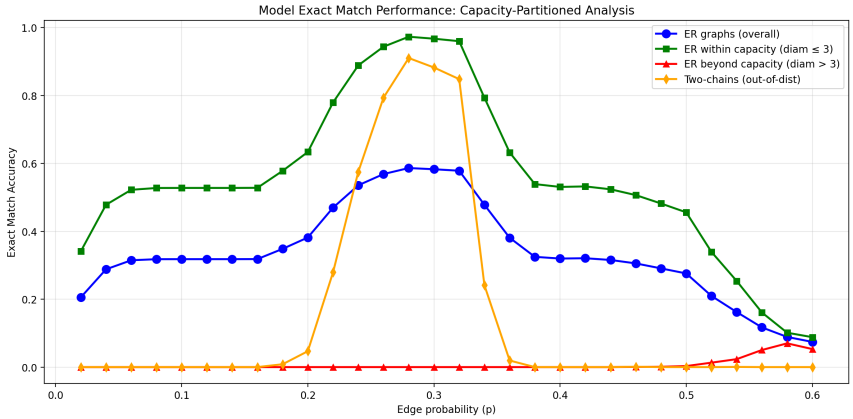



*Figure 11.* Under the same setup as Fig. 7, when tested on 2Clique graphs, the one trained with *the right data* is able to generalize better.

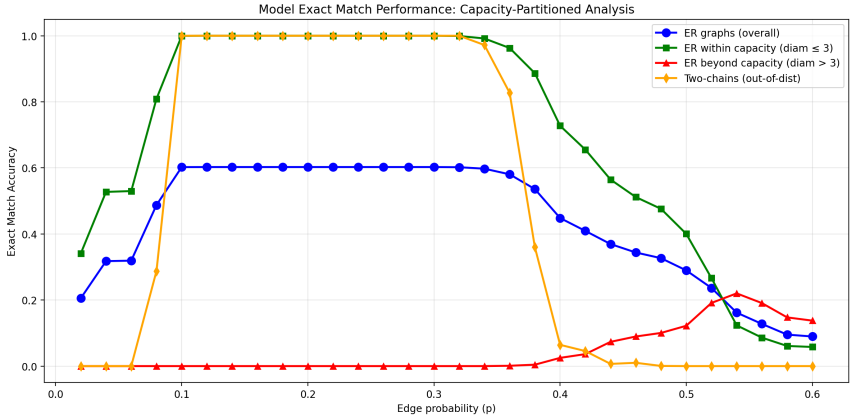
### D.3 Scaling Effects of Diameter and Capacity

## E Additional Related Work

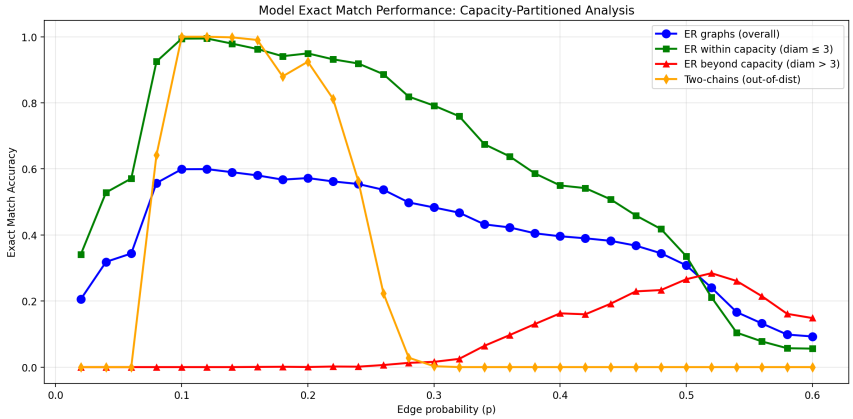
**Mechanistic Interpretability of Transformers.**A growing body of work reverse-engineers the *algorithmic circuits* that Transformers learn for tasks like copying, induction, and reasoning (Elhage et al., 2021; Olsson et al., 2022; Wang et al., 2023; Brinkmann et al., 2024). These can range from Fourier-style circuits for modular addition (Nanda et al., 2023; Zhou et al., 2024c) to Newton-like updates for in-context linear regression (Fu et al., 2024a). Researchers validate hypotheses by compiling programs into model weights (Lindner et al., 2023), decompiling models into code (Friedman et al., 2023), and using causal interventions to localize function (Chan et al., 2022; Meng et al., 2022; Yao et al., 2024; Chang et al., 2024). Theoretical work on inductive biases, like a preference for low-sensitivity functions, helps explain why models often favor robust heuristics over exact algorithms (Vasudeva et al., 2025).




*(a)* When training 1-layer Disentangled Transformers, instead of restricting training graphs to have diameter at most 3, we restrict $\mathrm{diam}(\mathbf{G}) \leq \mathbf{2}$ and varying the edge probability in $\mathsf{ER}(n = 8, p = p)$ training distribution. When measured by exact match accuracy, restricting $\mathrm{diam}(G) \leq 2$ make the models unable to generalize as well, indicating the importance of **at-capacity graphs** $(\mathrm{diam}(G) = 3)$




*(b)* When restricting $\mathrm{diam}(\mathbf{G}) \leq \mathbf{3}$, with reasonable $p \in [0.1, 0.32]$, 1-layer Disentangled Transformer can learn the algorithmic channel.




*(c)* When restricting $\mathrm{diam}(\mathbf{G}) \leq \mathbf{4}$, allowing some beyond-capacity graphs, 1-layer Disentangled Transformer struggle to learn the algorithmic channel, and starts to rely on the heuristic $J_n$-channel to make predictions on beyond-capacity graphs (red lines).

*Figure 12.* Effects of **at-capacity graphs** $(\mathrm{diam}(G) = 3^L)$ for $L = 1$. Without at-capacity graphs, models struggle to learn the algorithmic solution. With beyond-capacity graphs, models weight too much on heuristics. In short, models not only need most graphs within capacity and but also require at-capacity graphs to learn algorithms over heuristics.