ON PROVABLE LENGTH AND COMPOSITIONAL GENERALIZATION

Kartik Ahuja
FAIR, Meta
kartikahuja@meta.com

Amin Mansouri
Mila-Quebec AI Institute
sayed.mansouri-tehrani@mila.quebec

ABSTRACT

Length generalization – the ability to generalize to longer sequences than ones seen during training, and compositional generalization – the ability to generalize to token combinations not seen during training, are crucial forms of out-of-distribution generalization in sequence-to-sequence models. In this work, we take the first steps towards provable length and compositional generalization for a range of architectures, including deep sets, transformers, state space models, and simple recurrent neural nets. Depending on the architecture, we prove different degrees of representation identification, e.g., a linear or a permutation relation with ground truth representation, is necessary for length and compositional generalization.

1 INTRODUCTION

Large language models (LLMs), such as the GPT models (Achiam et al., 2023) and the Llama models (Touvron et al., 2023), have led to a paradigm shift in the development of future artificial intelligence (AI) systems. Despite being trained on a massive corpus of text data, these models struggle with planning and reasoning tasks, as pointed out by Bubeck et al. (2023), Stechly et al. (2023), and Valmeekam et al. (2023). The successes and failures of these models have sparked a debate about whether they actually learn general algorithms or if their success is primarily due to memorization.

A model's ability to perform well across different distribution shifts – for example, solving tasks involving longer problem instances than those seen during training – highlights its ability to learn general algorithms. A considerable amount of research has demonstrated that the right inductive biases, such as regularizers that enforce invariance or sparsity, lead to provable out-of-distribution (OOD) generalization guarantees for models operating with fixed-dimensional inputs (Peters et al., 2016; Arjovsky et al., 2019; Ahuja et al., 2021). Despite a large body of empirical works investigating OOD generalization in sequence-to-sequence models (Anil et al., 2022; Jelassi et al., 2023), we lack research efforts in developing provable OOD generalization guarantees for these models.

Length generalization – the ability to generalize to longer sequences than ones seen during training, and compositional generalization – the ability to generalize to token combinations not seen during training, are crucial forms of OOD generalization in sequence-to-sequence models. Zhou et al. (2023) provide a conjecture on the conditions under which transformers (Vaswani et al., 2017) length generalize, which has inspired the analysis we undertake here. In this work, we identify conditions under which different sequence-to-sequence models achieve length and compositional generalization. While transformers are the go-to choice for building LLMs, recently, alternative architectures based on state-space models, as noted by Gu et al. (2021), Orvieto et al. (2023b), and Gu and Dao (2023), have shown a lot of promise. This motivates us to study a range of architectures, including deep sets (Zaheer et al., 2017), transformers, state space models, and certain simple recurrent neural networks. A common theme in the results is that depending on the architecture, different degrees of representation identification (e.g., linear or permutation relationship to the ground truth representation) is necessary for length generalization. Our research confirms that these models can achieve provable length and compositional generalization in specific scenarios, offering valuable insights. However, frequent failures (Anil et al., 2022; Zhou et al., 2023) also reveal the urgency for a sharp

characterization of these failures. This will enhance our comprehension of these models and help us build better architectures.

2 Provable Length and Compositional Generalization

We are given a dataset comprising of a sequence of inputs $\{x_1, \cdots, x_t\}$ and a corresponding sequence of labels $\{y_1, \cdots, y_t\}$, where each $x_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}^m$. We denote a sequence $\{s_1, \cdots, s_t\}$ as $s_{\leq t}$. Each $x_{\leq t}$ is sampled from some distribution $\mathbb{P}_{X_{\leq t}}$ and label y_t is generated from labeling function $f(x_{\leq t})$. We define the set of input distribution and labeling function at all lengths $\mathcal{P} = \left\{ \{\mathbb{P}_{X_{\leq t}}\}_{t=1}^{\infty}, f \right\}$. Further, $\mathcal{P}(T)$ denotes the subset of $\mathcal{P}$ comprising of input distribution up to length T. We are tasked to learn a model from the dataset that takes an input sequence $x_{\leq t}$ as input and predicts the true label y_t as well as possible. If the model succeeds to predict well on sequences that are longer than any sequence seen under training distribution, then it is said to achieve length generalization (a more formal definition follows later). Further, if the model succeeds to predict well on sequences comprising of combination of tokens that are never seen under training distribution, then it is said to achieve compositional generalization (a more formal definition follows later.). We study both these forms of generalization from a first principles perspective.

Learning Objective Consider a map h that accepts sequences of n-dimensional inputs to generate a m-dimensional output. We measure the loss of predictions of h against true labels as $\ell\big(h(x_{\leq t}), y_t\big)$, where y_t is the true label for sequence $x_{\leq t}$. In what follows, we use ℓ_2 loss but appropriate generalizations for cross-entropy loss are possible. The expected risk across all time instances up to maximum length T is defined as $R(h;T) = \sum_{t=1}^{T} \mathbb{E}\big[\ell(h(x_{\leq t}), y_t)\big]$. The objective of the learner is to learn an h^* such that it minimizes

$$h^* \in \arg\min_{h \in \mathcal{H}} R(h;T), \tag{1}$$

where $\mathcal{H}$ is the hypothesis class of models. We now make the notions of length and compositional generalization formal.

Definition 1. *Consider the setting where a model is trained on sequences $(x_{\leq t}, y_{\leq t})$ of length up to T drawn from $\mathcal{P}(T)$. If the model achieves zero generalization error on sequences $(x_{\leq t}, y_{\leq t})$ of length up to $\tilde{T}$ drawn from $\mathcal{P}(\tilde{T}), \forall \tilde{T} \geq T$, then it achieves length generalization.*

Define the support of each token X_t under $\mathbb{P}_{X_{\leq t}}$, $\mathsf{supp}(X_t; \mathbb{P})$. Define the set of test input distribution and labeling function as $\mathcal{Q} = \left\{ \{\mathbb{Q}_{X_{\leq t}}\}_{t=1}^{\infty}, f \right\}$. We assume that the labeling function is the same between train and test. Further, we also assume that the support of the marginal distribution of each token is also equal, i.e., $\mathsf{supp}(X_t; \mathbb{P}) = \mathsf{supp}(X_t; \mathbb{Q}), \forall t \in \{1, \cdots, \infty\}$. We collect all such sets of test distributions $\mathcal{Q}$, which have the same labeling function as the train and have the same support over the marginal distribution of each token as train $\mathcal{P}$, into a set $\tilde{\mathcal{Q}}$.

Definition 2. *Consider the setting where a model is trained on sequences $(x_{\leq t}, y_{\leq t})$ of length up to T drawn from $\mathcal{P}(T)$. If the model achieves zero generalization error on sequences $(x_{\leq t}, y_{\leq t})$ of length up to T drawn from $\mathcal{Q}(T), \forall \mathcal{Q} \in \tilde{\mathcal{Q}}$, then it achieves compositional generalization.*

In the first definition, we ask if the model generalizes to longer sequences. In the second definition, we ask if the model generalizes to unseen combinations for seen lengths. This definition of compositionality follows from Wiedemer et al. (2023). Let us consider the following example. Suppose we train the model on sequences of length T where each x_i is a scalar. Consider the sphere $\mathcal{D} = \{x \in \mathbb{R}^T;\ \|x - \frac{1}{2}\mathbf{1}\| \leq \frac{1}{2}\}$, where $\mathbf{1}$ is a T dimensional vector of ones in each component. If our sequences come from $\mathcal{D}$, then the question we seek to understand is if the model generalizes to all combinations in $[0,1]^t$ for all t. This support $\mathcal{D}$ satisfies the assumptions that we make in the subsequent sections and hence is a good example to keep in mind.

Impossibility of Length and Compositional Generalization We argue that in the absence of any constraints on the hypothesis class $\mathcal{H}$ neither length generalization nor compositional generalization

are achievable. Let us consider the case of length generalization. If we minimize equation 1, then we learn an h that agrees with f on all sequences that are in the support (differences appear only on measure zero sets). However, there is no constraint on the behavior of h on sequences longer than T. If there are no constraints on $\mathcal{H}$, we can always construct a function that is equal to f on all sequences of length up to T but disagrees on sequences longer than T. Therefore, if this solution were selected, then length generalization will not be achieved. The same argument applies to the case of compositional generalization. In sections that follow, we study a constrained hypothesis class $\mathcal{H}$. We take inspiration from Zhou et al. (2023) and study the realizable case, i.e. true $f \in \mathcal{H}$ and constrain $\mathcal{H}$ to be somewhat simple.

2.1 Deep Sets

Deep sets were introduced in Zaheer et al. (2017). Informally stated, Zaheer et al. (2017) show that a function of a set $\mathcal{X}$, which by definition is permutation invariant, can be decomposed as $\rho\big(\sum_{x\in\mathcal{X}} \phi(x)\big)$, where x comes from a countable universe. Consider the example of addition operator, which takes $\{x_1, x_2\}$ and returns the sum $y = x_1 + x_2$. This operation is permutation invariant. Another example, can be sorting a list in increasing order, no matter the order of the inputs the output remains the same. This functional decomposition describes one of the simplest forms to define a function that accepts inputs of arbitrary lengths. Hence, we start with the $\mathcal{H}$ as deep sets.

Assumption 1. *Each function in the hypothesis class $\mathcal{H}$ used by the learner is given as $\omega\Big(\sum_{j\le i} \psi(x_j)\Big)$, where ω is a single layer perceptron with continuously differentiable bijective activation (e.g., sigmoid) and ψ is a map that is differentiable (almost everywhere).*

Assumption 2. *The joint support $\mathsf{supp}(X_{\le i}; \mathbb{P})$ is a regular closed set (in standard topology in $\mathbb{R}^{ni}$) for all $i \le T$.*

The above assumption is a mild regularity assumption (Lachapelle et al., 2023), which allows us to two equate the derivative of two functions if the two functions are equal as well.

Theorem 1. *If $\mathcal{H}$ follows Assumption 1 and the realizability condition holds, i.e., $f \in \mathcal{H}$, and $\mathsf{supp}(X_j; \mathbb{P}) = [0,1]^n,\ \forall j \in \{1, \cdots, \infty\}$ and the regular closedness condition in Assumption 2 holds, then the model trained to minimize the risk in equation 1 with ℓ_2 loss, achieves length and compositional generalization.*

The proof is provided in the Appendix. An important insight from the proof is if the output layer matrix has a left inverse, then the hidden representation learned by the model is a linear transform of the true hidden representation, i.e., $\psi = A\phi$. As a result, we obtain that such linear representation identification (Roeder et al., 2021) is necessary for length and compositional generalization.

2.2 Transformers

Ever since their introduction in Vaswani et al. (2017), transformers have revolutionized all domains of AI. In this section, we seek to understand length generalization for these models. In a simplified form, transformer architectures are represented as alternating layers of attention and position-wise non-linearity; we drop layer norms for tractability. Following similar notation as previous section, we denote position-wise non-linearity as ρ and attention layer as ϕ, we obtain the simplest form of causal transformer model as $\rho\Big(\sum_{j\le i} \phi(x_i, x_j) + x_i\Big)$. This decomposition captures linear attention, ReLU attention, sigmoid attention, ReLU squared attention, which were studied previously in Wortsman et al. (2023); Hua et al. (2022); Shen et al. (2023) and found to be quite effective in several settings (See further discussion on it in the Appendix). This decomposition does not capture softmax-based attention and developing provable length generalization guarantees for the same is an exciting future work. Other works Bai et al. (2023) also replaced softmax with other position-wise non-linearities for the ease of analysis.

Assumption 3. *Each function in the hypothesis class $\mathcal{H}$ used by the learner is given as $\omega\Big(\sum_{j\le i} \psi(x_i, x_j) + x_i\Big)$, where ω is a single layer perceptron with continuously differentiable bijective activation (e.g., sigmoid) and ψ is a map that is differentiable (almost everywhere).*

Theorem 2. *If $\mathcal{H}$ follows Assumption 3, the realizability condition holds, i.e., $f \in \mathcal{H}$, $\mathsf{supp}(X_i, X_j; \mathbb{P}) = [0,1]^{2n},\ \forall i \neq j \in \{1, \cdots, \infty\}$ and the regular closedness condition in Assumption 2 holds, then the model trained to minimize the risk in equation 1 (with $T \geq 3$) with ℓ_2 loss, achieves length and compositional generalization.*

The proof is provided in the Appendix. An important element that falls out of the proof is that if we use residual connections, then exact identification, i.e., $\omega = \rho$ and $\psi = \phi$, turns out to be necessary (and sufficient) for length and compositional generalization. If we do not use a residual connection, then we obtain a condition exactly similar to the previous theorem, i.e., linear identification of the attention head is necessary for length and compositional generalization.

We now extend Theorem 1,2 to general position-wise non-linearities.

Assumption 4. *Each function in $\mathcal{H}$ is expressed as $\omega(\sum_{j<i} \psi(x_i, x_j))$, where ω is a continuously differentiable bijection with a continuously differentiable inverse.*

The reader would notice that the summation carries a less than sign instead of less than equals. We believe the theorem that we present next extends to the more general case with less than equals to sign in the sum as well and the analysis for that case is underway.

Theorem 3. *If $\mathcal{H}$ follows Assumption 4, the realizability condition holds, i.e., $\mathsf{supp}(X_i, X_j; \mathbb{P}) = [0,1]^{2n},\ \forall i \neq j \in \{1, \cdots, \infty\}$ and mild assumption on the support (Assumption 7) holds, then the model trained to minimize the risk in equation 1 (with $T \geq 3$) with ℓ_2 loss, achieves length and compositional generalization.*

The proof is provided in the Appendix. We obtain a condition exactly similar to the previous theorem, i.e., linear representation identification is necessary for length and compositional generalization. The above result also extends to multiple attention heads (See the Appendix). We can use the same proof technique to extend Theorem 1 to position-wise non-linearities.

2.3 Linear recurrence and position-wise non-linearity

In recent years, state space models Gu et al. (2021); Orvieto et al. (2023b) have emerged as a promising competitor to transformers. In (Orvieto et al., 2023a;b), the authors give a first principles perspective on these models. We adopt their formulation and use linear recurrent layer with position-wise non-linearities as our main block towards understanding length generalization in these models. We illustrate the dynamics of these models to show the generation of sequence $y_{\leq T}$ from $x_{\leq T}$.

$$\begin{aligned} &h_1 = Bx_1; \quad h_2 = \Lambda h_1 + Bx_2; \ \cdots, h_T = \Lambda h_{T-1} + Bx_T \\ &y_1 = \rho(h_1);\ y_2 = \rho(h_2);\ \cdots\cdots, \qquad y_T = \rho(h_T), \end{aligned} \tag{2}$$

where h_t is hidden state at point t, $\Lambda \in \mathbb{R}^{n\times k}, B \in \mathbb{R}^{n\times k}$ and $\rho : \mathbb{R}^k \to \mathbb{R}^m$. We can succinctly write $h_T = \sum_{j=0}^{T-1} \Lambda^j B x_{T-j}$.

Assumption 5. *Each function in the hypothesis class $\mathcal{H}$ used by the learner is given as $\omega\Big(\sum_{j=0}^{T-1} \Lambda^j B x_{T-j}\Big)$, where ω is a single layer perceptron with continuously differentiable bijective activation σ (e.g., sigmoid), parametrized as $\sigma \circ A$. A, B, Λ are square, i.e., $k = m = n$, and invertible.*

Theorem 4. *If $\mathcal{H}$ follows Assumption 5, and the realizability condition holds, i.e., $f \in \mathcal{H}$, and a mild regularity condition on the support, i.e., Assumption 8, holds, then the model trained to minimize the risk in equation 1 with ℓ_2 loss ($T \geq 2$), achieves length and compositional generalization.*

The proof is in the Appendix. In the above result as well, linear representation identification, i.e., $\tilde{h}_t$ (predicted hidden state) and true hidden state h_t bear a linear relationship, turns out to be necessary for length and compositional generalization.

2.4 Vanilla RNNs

Standard RNNs have a non-linear recurrence unlike linear recurrence studied in previous section. We use the same notation as the previous section and only add an activation for non-linear recurrence.

$$\begin{aligned} &h_1 = \sigma(Bx_1); \quad h_2 = \sigma(\Lambda h_1 + Bx_2); \cdots, h_T = \sigma(\Lambda h_{T-1} + Bx_T) \\ &y_1 = \rho(h_1); \quad\quad y_2 = \rho(h_2); \cdots\cdots\cdots, \quad\quad y_T = \rho(h_T), \end{aligned} \tag{3}$$

Assumption 6. *Each function in the hypothesis class $\mathcal{H}$ used by the learner is a vanilla RNN, where the position-wise non-linearity is a single layer perceptron $\sigma \circ A$ and Λ, B govern the hidden state dynamics (as in equation 3). A, Λ, B are square invertible matrices, and σ is the sigmoid activation.*

Theorem 5. *If $\mathcal{H}$ follows Assumption 6, and the realizability condition holds, i.e., $f \in \mathcal{H}$ and regular closedness condition in Assumption 2 holds, then the model trained to minimize the risk in equation 1 with ℓ_2 loss (with $T \geq 2$), achieves length and compositional generalization.*

The proof is in the Appendix. Different from Theorem 4, here permutation identification, i.e., $\tilde{h}_t$ and h_t are related by a permutation matrix, is necessary for length and compositional generalization.

Discussion on RASP conjecture from Zhou et al. (2023) The RASP conjecture from Zhou et al. (2023) states – if the task of interest is realizable on a transformer and has simple description in the RASP programming language (Weiss et al., 2021), then the transformer achieves length generalization provided the data is sufficiently diverse. Our theoretical results capture the essential ingredients of this conjecture – we require the tasks to be realizable and simple. From the point of view of diversity, we require the marginal distribution of tokens or pairs of tokens to be sufficiently diverse, i.e., $\mathsf{supp}(X_j, \mathbb{P}) = [0, 1]^n$. And lastly, our results apply to architectures beyond transformers. That being said our characterization is far from complete. We believe a lot more work needs to be done to arrive at a more complete characterization of success and failure cases and use the insights gained to better tackle the empirical failures of length generization (Zhou et al., 2023; Anil et al., 2022).

3 Conclusion

We have initiated the formalization and identification of conditions for provable length generalization and compositional generalization in sequence-to-sequence models. This effort aims to give way to a foundation for the recently proposed RASP conjecture Zhou et al. (2023). Moving forward, it's crucial to extend our theory to cover even more complex tasks, reflecting the success and failure of these architectures across a broad spectrum of real-world tasks.

Acknowledgement

We thank Sébastien Lachapelle for insightful feedback and especially for pointing to Lemma 6 in Lachapelle et al. (2023), which helped us precisely use property of regular closed sets. We also thank various participants at the Bellairs workshop on causality 2024 that provided useful feedback on this work.

References

Emmanuel Abbe, Samy Bengio, Aryo Lotfi, and Kevin Rizk. Generalization on the unseen, logic reasoning and degree curriculum. *arXiv preprint arXiv:2301.13105*, 2023.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34:3438–3450, 2021.

Cem Anil, Yuhuai Wu, Anders Andreassen, Aitor Lewkowycz, Vedant Misra, Vinay Ramasesh, Ambrose Slone, Guy Gur-Ari, Ethan Dyer, and Behnam Neyshabur. Exploring length generalization

in large language models. *Advances in Neural Information Processing Systems*, 35:38546–38556, 2022.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Robert B Ash and Catherine A Doléans-Dade. *Probability and measure theory*. Academic press, 2000.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *arXiv preprint arXiv:2306.04637*, 2023.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

Weizhe Hua, Zihang Dai, Hanxiao Liu, and Quoc Le. Transformer quality in linear time. In *International Conference on Machine Learning*, pages 9099–9117. PMLR, 2022.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795, 2020.

Samy Jelassi, Stéphane d'Ascoli, Carles Domingo-Enrich, Yuhuai Wu, Yuanzhi Li, and François Charton. Length generalization in arithmetic transformers. *arXiv preprint arXiv:2306.15400*, 2023.

Sébastien Lachapelle, Divyat Mahajan, Ioannis Mitliagkas, and Simon Lacoste-Julien. Additive decoders for latent variables identification and cartesian-product extrapolation. *arXiv preprint arXiv:2307.02598*, 2023.

Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, pages 2873–2882. PMLR, 2018.

Joao Loula, Marco Baroni, and Brenden M Lake. Rearranging the familiar: Testing compositional generalization in recurrent networks. *arXiv preprint arXiv:1807.07545*, 2018.

Boris Mityagin. The zero set of a real analytic function. *arXiv preprint arXiv:1512.07276*, 2015.

Antonio Orvieto, Soham De, Caglar Gulcehre, Razvan Pascanu, and Samuel L Smith. On the universality of linear recurrences followed by nonlinear projections. *arXiv preprint arXiv:2307.11888*, 2023a.

Antonio Orvieto, Samuel L Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. Resurrecting recurrent neural networks for long sequences. *arXiv preprint arXiv:2303.06349*, 2023b.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016.

Geoffrey Roeder, Luke Metz, and Durk Kingma. On linear identifiability of learned representations. In *International Conference on Machine Learning*, pages 9030–9039. PMLR, 2021.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018.

Kai Shen, Junliang Guo, Xu Tan, Siliang Tang, Rui Wang, and Jiang Bian. A study on relu and softmax in transformer. *arXiv preprint arXiv:2302.06461*, 2023.

Kaya Stechly, Matthew Marquez, and Subbarao Kambhampati. Gpt-4 doesn't know it's wrong: An analysis of iterative prompting for reasoning problems. *arXiv preprint arXiv:2310.12397*, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Karthik Valmeekam, Matthew Marquez, and Subbarao Kambhampati. Can large language models really improve by self-critiquing their own plans? *arXiv preprint arXiv:2310.08118*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Gail Weiss, Yoav Goldberg, and Eran Yahav. Thinking like transformers. In *International Conference on Machine Learning*, pages 11080–11090. PMLR, 2021.

Thaddäus Wiedemer, Prasanna Mayilvahanan, Matthias Bethge, and Wieland Brendel. Compositional generalization from first principles. *arXiv preprint arXiv:2307.05596*, 2023.

Mitchell Wortsman, Jaehoon Lee, Justin Gilmer, and Simon Kornblith. Replacing softmax with relu in vision transformers. *arXiv preprint arXiv:2309.08586*, 2023.

Changnan Xiao and Bing Liu. Conditions for length generalization in learning reasoning skills. *arXiv preprint arXiv:2311.16173*, 2023.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

Hattie Zhou, Arwen Bradley, Etai Littwin, Noam Razin, Omid Saremi, Josh Susskind, Samy Bengio, and Preetum Nakkiran. What algorithms can transformers learn? a study in length generalization. *arXiv preprint arXiv:2310.16028*, 2023.

A APPENDIX

A.1 RELATED WORKS

Length generalization In the field of length generalization, significant strides have been made, as evidenced by various studies. Shaw et al. (2018) discovered the drawbacks of absolute positional embeddings and suggested relative positional embeddings as an alternative. Subsequent empirical analyses, notably by Anil et al. (2022) and Jelassi et al. (2023), explored length generalization in transformer-based models and arithmetic tasks, respectively. Key findings revealed that larger model sizes don't necessarily enhance generalization and that the utility of scratchpads varies, improving significantly when combined with in-context learning. Additionally, the effectiveness of relative positional embeddings appeared task-dependent, proving beneficial in simpler tasks like addition but faltering in more complex ones like multiplication. This led to the innovative approach of model priming with a few long sequence examples. Zhou et al. (2023) proposed the RASP conjecture, which we discussed already, to delineate the tasks where transformers excel or fall short in length generalization, emphasizing the necessity of task simplicity and data diversity.

On a more theoretical side, in Abbe et al. (2023), the authors showed an implicit bias of neural network training towards min-degree interpolators. This bias was used to explain the failures of length generalization on the parity task from Anil et al. (2022). In Xiao and Liu (2023), the authors use DAGs to formulate the computation in reasoning tasks and characterize conditions under which there exist functions that permit length generalization. Our results crucially differ in the sense we show a range of conditions under which both length and compositional generalization are actually achieved.

Compositional generalization The breadth of research on compositional generalization, encompassing studies like Lake and Baroni (2018); Hupkes et al. (2020); Loula et al. (2018), is too expansive to address comprehensively here. However, we reference several pertinent works that align with a specific theoretical viewpoint focused on representation identification. Recent studies, notably those by Wiedemer et al. (2023); Lachapelle et al. (2023), draw inspiration from object-centric architectures and approach compositional generalization from a first principles perspective. Our work adopts a definition of compositional generalization akin to these studies but diverges by centering on sequence-to-sequence models as the primary subject of interest.

A.2 PROOFS

In all the results that follow, we assume that the Radon-Nikodym derivative of $X_{\leq T}$ is absolutely continuous w.r.t Lebesgue measure. In all the results that follow, we work with standard topology in $\mathbb{R}^{nT}$, where n is dimension of each token and T is the sequence length. We remind the reader the definition of a regular closed set – if a set is equal to the closure of its interior, then it is said to be a regular closed set.

Lemma 1. *Let $\mathcal{X} \subseteq \mathbb{R}^n$. If $f : \mathcal{X} \to \mathbb{R}^m$ and $g : \mathcal{X} \to \mathbb{R}^m$ are continuously differentiable functions that satisfy $f(x) = g(x)$ almost everywhere in $\mathcal{X}$, where $\mathcal{X}$ is a regular non-empty closed set, then $f(x) = g(x), \forall x \in \mathcal{X}$ and $\nabla f(x) = \nabla g(x), \forall x \in \mathcal{X}$, where ∇ is the Jacobian w.r.t x.*

Proof. Let us consider the interior of $\mathcal{X}$ and denote it as $\mathcal{X}^{\mathsf{int}}$. We first argue that the two functions f and g are equal at all points in the interior. Suppose there exists a point $x \in \mathcal{X}^{\mathsf{int}}$ at which $f(x) \neq g(x)$.

Consider a ball centered at x of radius r denoted as $B(x, r) \subset \mathcal{X}^{\mathsf{int}}$ (such a ball exists as this point is in the interior of $\mathcal{X}$.). We argue that there exists at least one point x_1 in this ball at which $f(x_1) = g(x_1)$. If this were not the case, then the equality will not hold on the entire ball, which would contradict the condition that the equality $f(x) = g(x)$ can only be violated on a set of measure zero. Note this condition holds true for all $r > 0$. Suppose the distance of x_1 from x is $r_1 \leq r$. Consider another ball with radius $r_2 < r_1$ and let $x_2 \in B(x, r_2)$ where the equality holds. By repeating this argument, we can construct a sequence $\{x_k\}_{k\in\mathbb{N}}$ that converge to x, where $\mathbb{N}$ is the set of natural numbers. On this sequence, the following conditions hold.

$$f(x_k) = g(x_k), \forall k \in \mathbb{N} \tag{4}$$

Further, from the continuity of f and g it follows that

$$\lim_{k\to\infty} f(x_k) = f(x), \lim_{k\to\infty} g(x_k) = g(x) \tag{5}$$

Combining the above two conditions, we get that $f(x) = g(x)$. This leads to a contradiction since we assumed that $f(x) \neq g(x)$. Thus there can be no such x in the interior at which $f(x) \neq g(x)$. From this it follows that $f(x) = g(x)$ for all $x \in \mathcal{X}^{\mathsf{int}}$. Now let us consider the closure of $\mathcal{X}^{\mathsf{int}}$, which is $\mathcal{X}$ itself since it is a regular closed set. Every point $x \in \mathcal{X}$ in the closure can be expressed as limit of points in $\mathcal{X}^{\mathsf{int}}$. Consider an $x \in \mathcal{X}$ and from the definition of regular closed set it follows that $\lim_{k\to\infty} x_k = x$, where $x_k \in \mathcal{X}^{\mathsf{int}}$. We already know from the fact that f and g are equal in the interior

$$f(x_k) = g(x_k), \forall k \in \mathbb{N} \tag{6}$$

From the continuity of f and g it follows

$$\lim_{k\to\infty} f(x_k) = f(x), \lim_{k\to\infty} g(x_k) = g(x) \tag{7}$$

Combining the above two we get that $f(x) = g(x)$ for all $x \in \mathcal{X}$. After this we can use Lemma 6 from (Lachapelle et al., 2023) to conclude that $\nabla f(x) = \nabla g(x), \forall x \in \mathcal{X}$. We repeat their proof here for completeness. For all points in the interior of $\mathcal{X}$, it follows that $\nabla f(x) = \nabla g(x), \forall x \in \mathcal{X}^{\mathsf{int}}$.

Now consider any point $x \in \mathcal{X}$. Since $\mathcal{X}$ is a regular closed set, $\lim_{k\to\infty} x_k = x$. Since each x_k is in the interior of $\mathcal{X}$ it follows that

$$\nabla f(x_k) = \nabla g(x_k), \forall k \in \mathbb{N} \tag{8}$$

From the continuity of ∇f and ∇g it follows that

$$\lim_{k\to\infty} \nabla f(x_k) = \nabla f(x), \lim_{k\to\infty} \nabla g(x_k) = \nabla g(x) \tag{9}$$

Combining the above conditions, we get that $\nabla f(x) = \nabla g(x)$. This completes the proof.

□

Theorem 1. *If $\mathcal{H}$ follows Assumption 1 and the realizability condition holds, i.e., $f \in \mathcal{H}$, and $\mathsf{supp}(X_j; \mathbb{P}) = [0,1]^n,\ \forall j \in \{1, \cdots, \infty\}$ and the regular closedness condition in Assumption 2 holds, then the model trained to minimize the risk in equation 1 with ℓ_2 loss, achieves length and compositional generalization.*

Proof. Consider any h that solves equation 1. Since ℓ is ℓ_2 loss and realizability condition holds, f is one of the optimal solutions to equation 1. For all $x_{\leq T} \in \mathsf{supp}(X_{\leq T}; \mathbb{P})$ except over a set of measure zero the following condition holds

$$h(x_{\leq T}) = f(x_{\leq T}). \tag{10}$$

The above follows from the fact that h solves equation 1, i.e., $\mathbb{E}[\|h - f\|^2] = 0$ and from Theorem 1.6.6. (Ash and Doléans-Dade, 2000). Since $\mathsf{supp}(X_{\leq T})$ is regular closed, f, h are both continuously differentiable, we can use Lemma 1, it follows that the above equality holds for all $x_{\leq T} \in \mathsf{supp}(X_{\leq T})$. From realizability condition it follows that true $f(x_{\leq T}) = \rho\Big(\sum_{j\leq T} \phi(x_j)\Big)$. We substitute the functional decomposition from Assumption 1 to get

$$\omega\Big(\sum_{j\leq T} \psi(x_j)\Big) = \rho\Big(\sum_{j\leq T} \phi(x_j)\Big). \tag{11}$$

Since ω and ρ are single layer perceptron with bijective activation σ. We substitute the parametric form of ω and ρ to obtain

$$\begin{aligned} &\sigma\Big(A\sum_{j\leq T} \psi(x_j)\Big) = \sigma\Big(B\sum_{j\leq T} \phi(x_j)\Big) \implies \\ &A\sum_{j\leq T} \psi(x_j) = B\sum_{j\leq T} \phi(x_j). \end{aligned} \tag{12}$$

The second equality in the above simplification follows from the fact that the activation σ is bijective, the inputs to σ are equal. We take the derivative of the expressions above w.r.t x_r to get the following condition and equate them (follows from Lemma 1). For all $x_r \in \mathsf{supp}(X_r; \mathbb{P})$, i.e., $x_r \in [0,1]^n$,

$$\nabla_{x_r}\Big(A\sum_{j\leq T}\psi(x_j)\Big) = \nabla_{x_r}\Big(B\sum_{j\leq T}\phi(x_j)\Big). \tag{13}$$

We drop the subscript r to simplify the notation. Therefore, for all $x \in [0,1]^n$

$$A\nabla_x\psi(x) = B\nabla_x\phi(x), \tag{14}$$

where $\nabla_x\psi(x)$ is the Jacobian of $\psi(x)$ w.r.t x and $\nabla_x\phi(x)$ is the Jacobian of $\phi(x)$ w.r.t x.

Let us consider the derivative w.r.t some component x^k of x. Denote the remaining components as x^{-k}. From the above condition it follows that for all $x \in [0,1]^n$,

$$A\frac{\partial\psi(x)}{\partial x^k} = B\frac{\partial\phi(x)}{\partial x^k}. \tag{15}$$

Using fundamental theorem of calculus, we can integrate both sides for fixed x^{-k} and obtain the following for all $x^k \in [0,1]$,

$$\begin{aligned} &A\psi(x^k, x^{-k}) = B\phi(x^k, x^{-k}) + C_k(x^{-k}) \implies \\ &A\psi(x) - B\phi(x) = C_k(x^{-k}). \end{aligned} \tag{16}$$

The above condition is true of all $k \in \{1, \cdots, n\}$. Hence, we can deduce that for all $x \in [0,1]^d$ and for $k \neq j$, where $j, k \in \{1, \cdots, d\}$,

$$A\psi(x) - B\phi(x) = C_k(x^{-k}) = C_j(x^{-j}). \tag{17}$$

Take the partial derivative of $C_k(x^{-k})$ and $C_j(x^{-j})$ w.r.t x^j to obtain, for all $x^j \in [0,1]$,

$$\frac{\partial C_k(x^{-k})}{\partial x^j} = \frac{\partial C_j(x^{-j})}{\partial x^j} = 0. \tag{18}$$

In the above simplification, we use the fact that $\forall x^j \in [0,1]$, $\frac{\partial C_j(x^{-j})}{\partial x^j} = 0$. Therefore, $C_k(x^{-k})$ cannot depend on x^j. We can apply the same condition on all $j \neq k$. As a result, $C_k(x^{-k})$ is a fixed constant vector denoted as C. We write this as

$$A\psi(x) = B\phi(x) + C. \tag{19}$$

Substitute the above into $A\sum_{j\leq T}\psi(x_j) = B\sum_{j\leq T}\phi(x_j)$ to obtain

$$B\sum_{j\leq T}\phi(x_j) + CT = B\sum_{j\leq T}\phi(x_j) \implies C = 0. \tag{20}$$

Therefore, we get

$$\forall x \in [0,1]^n, A\psi(x) = B\phi(x). \tag{21}$$

We now consider a sequence $x_{\leq \tilde{T}}$ from the test distribution $\mathbb{Q}_{X_{\leq \tilde{T}}}$. The prediction made by h is

$$h(x_{\leq \tilde{T}}) = \sigma\Big(A \sum_{j \leq \tilde{T}} \psi(x_j)\Big) = \sigma\Big(B \sum_{j \leq \tilde{T}} \phi(x_j)\Big) = f(x_{\leq \tilde{T}}). \tag{22}$$

We use equation 21 in the simplification above. From the above, we can conclude that h continues to be optimal for distribution $\mathbb{Q}_{X_{\leq \tilde{T}}}$.

□

Remarks A few remarks and observations from the proof are in order. First of all notice that we do not require ϕ and ψ to have the same output dimension. Suppose A has a left inverse. If that is the case, then we can simplify equation 21 to obtain $\psi(x) = A^{-1}B\phi(x), \forall x \in (0,1)^d$. This condition is known as linear representation identification in the literature (Roeder et al., 2021). We call this condition as necessary as it is implied under the condition that the model h achieves zero generalization error at a given length. In Theorem 1, we observe all the labels from $t = 1$ to T, i.e., y_1 to y_T. The result continues to hold if we only observe label at length T, i.e., y_T. This is to clarify that the proof does not rely on equating the true prediction and the predictions of the model for first token $t = 1$, which would make the analysis very similar to that of models with fixed dimensional inputs. Also, note that we would not need to separately require Assumption 2, if we equated the predictions at $t = 1$ because the support of each token's marginal distribution is $[0,1]^n$, which is regular closed. Since we want to emphasize that the result would continue to hold even if we only observed data for some length T and not all lengths leading up to T, we make the additional Assumption 2.

Theorem 2. *If $\mathcal{H}$ follows Assumption 3, the realizability condition holds, i.e., $f \in \mathcal{H}$, $\mathsf{supp}(X_i, X_j; \mathbb{P}) = [0,1]^{2n},\ \forall i \neq j \in \{1, \cdots, \infty\}$ and the regular closedness condition in Assumption 2 holds, then the model trained to minimize the risk in equation 1 (with $T \geq 3$) with ℓ_2 loss, achieves length and compositional generalization.*

Proof. Consider any h that solves equation 1. Since ℓ is ℓ_2 loss and realizability condition holds, f is one of the optimal solutions to equation 1. For all $x_{\leq i} \in \mathsf{supp}(X_{\leq i}; \mathbb{P})$ except over a set of measure zero the following condition holds

$$h(x_{\leq i}) = f(x_{\leq i}). \tag{23}$$

The above follows from the fact that h solves equation 1, i.e., $\mathbb{E}[\|h - f\|^2] = 0$ and from Theorem 1.6.6. (Ash and Doléans-Dade, 2000). Since $\mathsf{supp}(X_{\leq i})$ is regular closed, f, h are both continuously differentiable, we can use Lemma 1, it follows that the above equality holds for all $x_{\leq i} \in \mathsf{supp}(X_{\leq i})$. From realizability condition it follows that true $f(x_{\leq i}) = \rho\Big(\sum_{k \leq i} \phi(x_i, x_k)\Big)$. We substitute the parametric forms from Assumption 3 to get

$$\omega\Big(\sum_{k \leq i} \psi(x_i, x_k) + x_i\Big) = \rho\Big(\sum_{j \leq i} \phi(x_i, x_k) + x_i\Big). \tag{24}$$

Since ω and ρ are single layer perceptron with bijective activation σ. We substitute the parametric form of ω and ρ to obtain the following condition. For all $x_{\leq i} \in \mathsf{supp}(X_{\leq i})$,

$$\begin{aligned} &\sigma\Big(A \sum_{k \leq i} \psi(x_i, x_j) + Ax_i\Big) = \sigma\Big(B \sum_{k \leq i} \phi(x_i, x_j) + Bx_i\Big) \implies \\ &A \sum_{k \leq i} \psi(x_i, x_k) + Ax_i = B \sum_{k \leq i} \phi(x_i, x_k) + Bx_i. \end{aligned} \tag{25}$$

The second equality follows from the fact that the activation σ is bijective and hence the inputs to σ are equal. We take the derivative of the expressions above w.r.t x_j to get the following (follows

from Lemma 1). For $j < i$ (there exists a $j < i$ as $T \geq 3$ and we can set $i \geq 2$) and for all $x_j \in \mathsf{supp}(X_j; \mathbb{P})$, i.e., $x_j \in [0,1]^n$,

$$\begin{aligned} &\nabla_{x_j}\Big(A\sum_{k\leq i}\psi(x_i,x_k) + Ax_i\Big) = \nabla_{x_j}\Big(B\sum_{k\leq i}\phi(x_i,x_j) + Bx_i\Big) \implies \\ &A\nabla_{x_j}\psi(x_i,x_j) = B\nabla_{x_j}\phi(x_i,x_j), \end{aligned} \tag{26}$$

where $\nabla_{x_j}\psi(x_i,x_j), \nabla_{x_j}\phi(x_i,x_j)$ are the Jacobians of ψ and ϕ w.r.t x_j for a fixed x_i. Note that $A\nabla_{x_j}\psi(x_i,x_j) = B\nabla_{x_j}\phi(x_i,x_j)$ holds for all $x_i \in [0,1]^n, x_j \in [0,1]^n$ (here we use the fact that joint support of every pair of tokens spans $2n$ dimensional unit hypercube assumed in the Theorem 2). In this equality, we now consider the derivative w.r.t some component x_j^k of x_j. Denote the remaining components as x_j^{-k}. From the above condition it follows that for all $x_i \in [0,1]^n, x_j \in [0,1]^n$,

$$A\frac{\partial\psi(x_i,x_j)}{\partial x_j^k} = B\frac{\partial\phi(x_i,x_j)}{\partial x_j^k}. \tag{27}$$

Using fundamental theorem of calculus, we can integrate both sides for fixed x_j^{-k} and obtain the following for all $x_j^k \in [0,1]$,

$$\begin{aligned} &A\psi\big(x_i,[x_j^k,x_j^{-k}]\big) = B\phi\big(x_i,[x_j^k,x_j^{-k}]\big) + C_k\big(x_i,x_j^{-k}\big) = \\ &A\psi(x_i,x_j) = B\phi(x_i,x_j) + C_k(x_i,x_j^{-k}). \end{aligned} \tag{28}$$

The same condition is true of all k. Hence, $\forall x_i \in [0,1]^d, \forall x_j \in [0,1]^d$ and for $k \neq q$, where $q,k \in \{1,\cdots,d\}$,

$$A\psi(x_i,x_j) - B\phi(x_i,x_j) = C_k(x_i,x_j^{-k}) = C_q(x_i,x_j^{-q}). \tag{29}$$

Take the partial derivative of both sides w.r.t x_j^q to obtain, $\forall x_j^q \in [0,1]$,

$$\frac{\partial C_k(x_i,x_j^{-k})}{\partial x_j^q} = \frac{\partial C_q(x_i,x_j^{-q})}{\partial x_j^q} = 0. \tag{30}$$

Therefore, $C_k(x_i,x_j^{-k})$ cannot depend on x_j^q. We can apply the same condition on all $q \neq k$. As a result, $C_k(x_i,x_j^{-k})$ is only a function of x_i denoted as $C(x_i)$. Therefore, for $j < i$ and for all $x_i \in [0,1]^n, x_j \in [0,1]^n$

$$A\psi(x_i,x_j) = B\phi(x_i,x_j) + C(x_i). \tag{31}$$

If we substitute $x_i = x_j = x$, then the above equality extends for $i = j$. Substitute the above equation 31 into $A\sum_{j\leq i}\psi(x_i,x_j) + Ax_i = B\sum_{j\leq i}\phi(x_i,x_j) + Bx_i$ to obtain

$$\begin{aligned} &B\sum_{j\leq i}\phi(x_i,x_j) + (i)C(x_i) + Ax_i = B\sum_{j\leq i}\phi(x_i,x_j) + Bx_i \implies \\ &C(x_i) = (Bx_i - Ax_i)/(i). \end{aligned} \tag{32}$$

In the above i is at least two. Consider $i+1$ (since $T \geq 3$ we can consider this length)

$$C(x_{i+1}) = (Bx_{i+1} - Ax_{i+1})/(i+1) \tag{33}$$

Equating the above for same values of $x_i = x_{i+1} = x$, we obtain $(B - A)x = 0$ and thus $C(x) = 0$. Since this identity is valid for all $x \in [0, 1]^n$, we obtain that $A = B$. We can do this for all $x \in [0, 1]^n$ and thus $C(x) = 0, \forall x \in [0, 1]^n$. Thus we obtain

$$\forall x_i \in [0, 1]^n, x_j \in [0, 1]^n \;\; \psi(x_i, x_j) = \phi(x_i, x_j). \tag{34}$$

Since $A = B$ and $\phi = \psi$, $f = h$ and optimality on longer sequences trivially follows.

We now consider the case when there is no residual connection. The entire analysis is similar up to equation 31. Following similar steps as above, we get $C(x) = 0, \forall x \in [0, 1]^n$. From this it follows that for $j \leq i$

$$\forall x_i \in [0, 1]^n, x_j \in [0, 1]^n \;\; A\psi(x_i, x_j) = B\phi(x_i, x_j). \tag{35}$$

In the above, case A and B need not be equal. If A is left invertible, then this implies linear representation identification.

We now consider a sequence $x_{\leq \tilde{T}}$ from the test distribution $\mathcal{Q} \in \tilde{\mathcal{Q}}$. The prediction made by h is

$$h(x_{\leq \tilde{T}}) = \sigma\Big(A \sum_{j \leq \tilde{T}} \psi(x_{\tilde{T}}, x_j)\Big) = \sigma\Big(B \sum_{j \leq \tilde{T}} \phi(x_{\tilde{T}}, x_j)\Big) = f(x_{\leq \tilde{T}}) \tag{36}$$

We use equation 34 in the simplification above. From the above, we can conclude that h continues to be optimal for distribution $\mathbb{Q}_{X_{\leq \tilde{T}}}$.

□

Assumption 7. *The joint support* $\mathsf{supp}(X_{\leq i}; \mathbb{P})$ *is a regular closed set for all* $i \leq T$. *The support of* $[\phi(X_1, X_2), \phi(X_1, X_3)]$ *is* $\mathbb{R}^{2k}$.

We want to make two remarks on the above assumption. The assumption can be alternately stated in a relaxed form and require the support of $[\phi(X_1, X_2), \phi(X_1, X_3)]$ to be a linear subspace of $\mathbb{R}^{2k}$.

Theorem 3. *If* $\mathcal{H}$ *follows Assumption 4, the realizability condition holds, i.e.,* $\mathsf{supp}(X_i, X_j; \mathbb{P}) = [0, 1]^{2n}, \;\forall i \neq j \in \{1, \cdots, \infty\}$ *and mild assumption on the support (Assumption 7) holds, then the model trained to minimize the risk in equation 1 (with* $T \geq 3$*) with* ℓ_2 *loss, achieves length and compositional generalization.*

Proof. We start with the same steps as earlier proofs and equate the prediction of h and f. We first use the fact $h(x_{\leq i}) = f(x_{\leq i})$ almost everywhere in the support. We can use the continuity of h, f and regular closedness of the support to extend the equality to all points in the support (follows from the first part of Lemma 1) to obtain the following. For all $x_{\leq i} \in \mathsf{supp}(X_{\leq i})$.

$$\begin{aligned}
&\omega\Big(\sum_{j<i} \psi(x_i, x_j)\Big) = \rho\Big(\sum_{j<i} \phi(x_i, x_j)\Big), \\
&\sum_{j<i} \psi(x_i, x_j) = \omega^{-1} \circ \rho\Big(\sum_{j<i} \phi(x_i, x_j)\Big), \\
&\sum_{j<i} \psi(x_i, x_j) = a\Big(\sum_{j<i} \phi(x_i, x_j)\Big).
\end{aligned} \tag{37}$$

In the above simplification, we used the parametric form for the true labeling function and the learned labeling function and use the invertibility of ρ. Let us consider the setting when $i = 2$. In that case summation involves only one term. Substitute $x_1 = y$ and $x_2 = x$. We obtain $\forall x \in [0, 1]^n, y \in [0, 1]^n$,

$$\psi(x, y) = a(\phi(x, y)). \tag{38}$$

The above expression implies that ψ bijectively identifies ϕ. We substitute the above expression for a summation comprising of two terms, i.e., case when $i = 3$ (this is possible since $T \geq 3$), (x, y) and (x, z) to give

$$a(\phi(x, y)) + a(\phi(x, z)) = a\big(\phi(x, y) + \phi(x, z)\big). \tag{39}$$

We now use the that assumption $[\phi(x, y), \phi(x, z)]$ spans $\mathbb{R}^{2k}$ and substitute $\phi(x, y) = \alpha$ and $\phi(x, z) = \beta$

$$a(\alpha) + a(\beta) = a\big(\alpha + \beta\big). \tag{40}$$

Observe that $a(0) = 0$.

We want to show that a is linear. To show that, we need to argue that $a(c\alpha) = ca(\alpha)$ as we already know a satisfies additivity condition.

Suppose c is some rational number, i.e., $c = p/q$, where p and q are non-zero integers.

From the identity it is clear that $a(p\alpha) = pa(\alpha)$, where p is some integer. This follows directly.

$a(q\frac{1}{q}\alpha) = qa(\frac{1}{q}\alpha) \implies a(\frac{1}{q}\alpha) = \frac{1}{q}a(\alpha)$, where q is some integer.

Now combine these $a(p/q\alpha) = pa(1/q\alpha) = \frac{p}{q}a(\alpha)$. We have established the homogeneity condition for rationals.

We will now use the continuity of the function a and density of rationals to extend the claim for irrationals. Suppose c is some irrational. Define a sequence of rationals that approach c (this follows from that rationals are dense in $\mathbb{R}$).

$a(c\alpha) = a(\lim_{n\to\infty} q_n\alpha) = \lim_{n\to\infty} a(q_n\alpha)$.

In the second equality above, we use the definition of continuity (a is continuous since composition of continuous functions is continuous). We can also use the property that we already showed for rationals to further simplify

$\lim_{n\to\infty} a(q_n\alpha) = a(\alpha) \lim_{n\to\infty} q_n = ca(\alpha)$.

We have now established that a is linear. As a result, we can write

$$\begin{aligned} \psi(x, y) &= A(\phi(x, y)), \\ \phi(x, y) &= C(\psi(x, y)). \end{aligned} \tag{41}$$

$$\rho^{-1} \circ \omega(z) = Cz \implies \omega(z) = \rho(Cz)$$

We use the above to simplify $\omega\big(\sum_{j<i} \psi(x_i, x_j)\big) = \rho(C \sum_{j<i} \psi(x_i, x_j)) = \rho\big(\sum_{j<i} \phi(x_i, x_j)\big)$. This identity holds for all i. Thus we obtain length and compositional generalization.

□

Remark In the above proof, we did not have a residual connection or a term that involves $\phi(x, x)$. We believe that the stated result should continue to hold for the case when we include these terms and a different proof for that case is under active investigation. Our choice of the archictecture did not invoke multiple attention heads. If we include multiple attention heads, then also we can arrive at the same length generalization guarantees. However, an interesting consequence of multiple heads is that the linear identification is not required to hold on a per head level. We explain that below.

Following the same steps as the proof above, we get

$$\begin{aligned}
&\omega\Big(\sum_{j<i}\psi_1(x_i,x_j)+\psi_2(x_i,x_j)\Big)=\rho\Big(\sum_{j<i}\phi_1(x_i,x_j)+\phi_2(x_i,x_j)\Big),\\
&\sum_{j<i}\psi_1(x_i,x_j)+\psi_2(x_i,x_j)=\omega^{-1}\circ\rho\Big(\sum_{j<i}\phi_1(x_i,x_j)+\phi_2(x_i,x_j)\Big),\\
&\sum_{j<i}\psi_1(x_i,x_j)+\psi_2(x_i,x_j)=a\Big(\sum_{j<i}\phi_1(x_i,x_j)+\phi_2(x_i,x_j)\Big).
\end{aligned}\tag{42}$$

We can also show that a is linear. If a is linear, then sum of $\psi_1+\psi_2$ linearly identifies $\phi_1+\phi_2$. Therefore, this implies that each learned attention head need not correspond to one of the ground truth attention head.

Assumption 8. *The joint support* $\mathsf{supp}(X_{\leq i};\mathbb{P})$ *is a regular closed set for all* $i\leq T$*. For some length* $2\leq i\leq T$ *an there exists* in *sequences* $x_{\leq i}$ *such that their concatenation forms a* $in\times in$ *matrix of rank* in*.*

Theorem 4. *If* $\mathcal{H}$ *follows Assumption 5, and the realizability condition holds, i.e.,* $f\in\mathcal{H}$*, and a mild regularity condition on the support, i.e., Assumption 8, holds, then the model trained to minimize the risk in equation 1 with* ℓ_2 *loss* $(T\geq 2)$*, achieves length and compositional generalization.*

Proof. We start with the same steps as earlier proofs and equate the prediction of h and f. We first use the fact $h(x_{\leq i})=f(x_{\leq i})$ almost everywhere in the support. We can use the continuity of h,f and regular closedness of the support to extend the equality to all points in the support (from first part of Lemma 1) to obtain the following. For all $x_{\leq i}\in\mathsf{supp}(X_{\leq i})$.

$$\begin{aligned}
&f(x_{\leq i})=h(x_{\leq i})=\\
&\sigma(A\sum_{j=0}^{i-1}\Lambda^jBx_{i-j})=\sigma(\tilde{A}\sum_{j=0}^{i-1}\tilde{\Lambda}^j\tilde{B}x_{i-j})\implies\\
&A\sum_{j=0}^{i-1}\Lambda^jBx_{i-j}=\tilde{A}\sum_{j=0}^{i-1}\tilde{\Lambda}^j\tilde{B}x_{i-j}=\\
&[AB,A\Lambda B,A\Lambda^2B,\cdots,A\Lambda^{i-1}B]\begin{bmatrix}x_i\\x_{i-2}\\\vdots\\x_1\end{bmatrix}-[\tilde{A}\tilde{B},\tilde{A}\tilde{\Lambda}\tilde{B},\tilde{A}\tilde{\Lambda}^2\tilde{B},\cdots,\tilde{A}\tilde{\Lambda}^{i-1}\tilde{B}]\begin{bmatrix}x_i\\x_{i-2}\\\vdots\\x_1\end{bmatrix}=0\implies\\
&\Big[[AB,A\Lambda B,A\Lambda^2B,\cdots,A\Lambda^{i-1}B]-[\tilde{A}\tilde{B},\tilde{A}\tilde{\Lambda}\tilde{B},\tilde{A}\tilde{\Lambda}^2\tilde{B},\cdots,\tilde{A}\tilde{\Lambda}^{i-1}\tilde{B}]\Big]\boldsymbol{X}=0,
\end{aligned}\tag{43}$$

where $\boldsymbol{X}=\begin{bmatrix}x_i\\x_{i-2}\\\vdots\\x_1\end{bmatrix}$. In the above simplification, we used the fact that σ is a bijection.

Denote $R=\Big[[AB,A\Lambda B,A\Lambda^2B,\cdots,A\Lambda^{i-1}B]-[\tilde{A}\tilde{B},\tilde{A}\tilde{\Lambda}\tilde{B},\tilde{A}\tilde{\Lambda}^2\tilde{B},\cdots,\tilde{A}\tilde{\Lambda}^{i-1}\tilde{B}]$. We collect a set of points $\boldsymbol{X}^+=[\boldsymbol{X}^{(1)},\cdots,\boldsymbol{X}^{(l)}]$ where $l\geq ni$ and rank of $\boldsymbol{X}^+=ni$ (from Assumption 8). Since the matrix $\boldsymbol{X}^+$ is full rank, we have

$$R\boldsymbol{X}^+=0\implies R=0.$$

This yields

$$AB=\tilde{A}\tilde{B},A\Lambda B=\tilde{A}\tilde{\Lambda}\tilde{B},\cdots,A\Lambda^iB=\tilde{A}\tilde{\Lambda}^i\tilde{B}.\tag{44}$$

First equality is restated as $\tilde{A}^{-1}A = \tilde{B}B^{-1}$. Denote $\tilde{A}^{-1}A = C$. Thus $\tilde{B} = BC$. Observe that from the second equality, we get $\tilde{\Lambda} = C\Lambda C^{-1}$. Given the true data generation is parametrized by the triplet (A, Λ, B), the set of triplets $(\tilde{A}, \tilde{\Lambda}, \tilde{B})$ that solve the first two equalities are – $\{\tilde{A}$ is an arbitrary invertible matrix, $\tilde{\Lambda} = C\Lambda C^{-1},\ \tilde{B} = BC,$ where $C = \tilde{A}^{-1}A\}$.

Take any solution of the first two equalities and compute

$$\tilde{A}\tilde{\Lambda}^i\tilde{B} = \tilde{A}C\Lambda^i C^{-1}\tilde{B} = A\Lambda^i B$$

Since the above condition holds for all k, we obtain the following condition, $\forall i \geq 1$

$$h(x_{\leq i}) = \sigma(\tilde{A}\sum_{j=0}^{i-1}\tilde{\Lambda}^j\tilde{B}x_{i-j}) = \sigma(A\sum_{j=0}^{i-1}\Lambda^j Bx_{i-j}) = f(x_{\leq i}) \tag{45}$$

Hence, we know that equating the predictions up to length 2 suffices to learn parameters that generalize to all the lengths beyond.

□

Theorem 5. *If $\mathcal{H}$ follows Assumption 6, and the realizability condition holds, i.e., $f \in \mathcal{H}$ and regular closedness condition in Assumption 2 holds, then the model trained to minimize the risk in equation 1 with ℓ_2 loss (with $T \geq 2$), achieves length and compositional generalization.*

Proof. We start with the same steps as earlier proofs and equate the prediction of h and f everywhere in the support of the training distribution (using first part of Lemma 1). We start with equating label at length 1, i.e., y_1. For all $x_1 \in \mathsf{supp}(X_1; \mathbb{P})$

$$\begin{aligned}
&\sigma(A\sigma(Bx_1)) = \sigma(\tilde{A}\sigma(\tilde{B}x_1)) \implies \\
&A\sigma(Bx_1) = \tilde{A}\sigma(\tilde{B}x_1) \implies \\
&\sigma(B\tilde{B}^{-1}\tilde{B}x_1) = A^{-1}\tilde{A}\sigma(\tilde{B}x_1)
\end{aligned} \tag{46}$$

Say $y = \tilde{B}x_1$, $A^{-1}\tilde{A} = U$, $B\tilde{B}^{-1} = V$. We substitute these expressions in the simplificaction below. Observe that $y \in \tilde{B}\mathsf{supp}(X_1; \mathbb{P})$. We pick a y in the interior of $\tilde{B}\mathsf{supp}(X_1; \mathbb{P})$.

$$\sigma(Vy) = U\sigma(y) \tag{47}$$

Take the first row of V and U as $v^\top$ and $u^\top$ to obtain

$$\sigma(v^\top y) = u^\top\sigma(y) \tag{48}$$

Suppose there is some non-zero component of v say i but the corresponding component is zero in u.

$$\frac{\partial\sigma(v_i y_i + v_{-i}y_{-i})}{\partial y_i} = \sigma^{'}(v_i y_i + v_{-i}y_{-i})v_i = \frac{\partial u_{-i}^\top\sigma(y_{-i})}{\partial y_i} = 0 \tag{49}$$

From the above we get $\sigma^{'}(v^\top y) = 0$. But sigmoid is strictly monotonic on $\mathbb{R}$, $\sigma^{'}(x) > 0, \forall x \in \mathbb{R}$ and $v^\top y \in \mathbb{R}$. Hence, $\sigma^{'}(v^\top y) = 0$ is not possible. Similarly, suppose some component is non-zero in u and zero in v.

$$\frac{\partial\sigma(v_{-i}^\top y_{-i})}{\partial y_i} = 0 = \frac{\partial(u_i\sigma(y_i) + u_{-i}^\top\sigma(y_{-i}))}{\partial y_i} = u_i\sigma^{'}(y_i) \tag{50}$$

Since the derivative of σ cannot be zero, the above condition cannot be true.

From the above, we can deduce that both u and v have same non-zero components.

Let us start with the case where $p \geq 2$ components of u, v are non-zero. Below we equate the partial derivative w.r.t all components of y that have non-zero component in u (since y is in the interior of the image of $\tilde{B}x_1$, we can equate these derivatives).

$$\begin{aligned} \sigma(v^\top y) &= u^\top \sigma(y), \\ \frac{\partial^p \sigma(s)}{\partial s^p} \Pi_{u_i \neq 0} u_i = 0 \implies \frac{\partial^p \sigma(s)}{\partial s^p} &= 0. \end{aligned} \tag{51}$$

Since support X_1 has a non-empty interior, the set of values $v^\top y$ takes also has a non-empty interior in $\mathbb{R}$. From the above, we obtain $\sigma(s)$ is a polynomial of degree p. If $\sigma(s) = q(s)$ over a set of measure greater than zero in $\mathbb{R}$, $\sigma = q$ everywhere (from (Mityagin, 2015)), which cannot be true.

We are left with the case where u and v have one non-zero component each.

$\frac{1}{1+e^{-vy}} = \frac{u}{1+e^{-y}} \implies 1 + e^{-y} = u + ue^{-vy}$ In the simplification above, we take derivative w.r.t y to obtain $e^{-(v-1)y} = 1/uv$. We now again take derivative again w.r.t y to get $v = 1$ and substitute it back to get $u = 1$. Note that no other row of U or V can have same non-zero element because that would make matrix non invertible. From this we deduce that U and V are permutation matrices. From $\sigma(Vy) = U\sigma(y)$ it follows that $U = V = \Pi$. Thus $B = \Pi\tilde{B}$ and $\tilde{A} = A\Pi$.

Next, we equate predictions for y_2 to the ground truth. For all $x_1 \in \mathsf{supp}(X_1; \mathbb{P})$

$$\begin{aligned} &\sigma(A\sigma(\Lambda\sigma(Bx_1) + Bx_2)) = \sigma(\tilde{A}\sigma(\tilde{\Lambda}\sigma(\tilde{B}x_1) + \tilde{B}x_2)) \implies \\ &A\sigma(\Lambda\sigma(Bx_1) + Bx_2) = \tilde{A}\sigma(\tilde{\Lambda}\sigma(\tilde{B}x_1) + \tilde{B}x_2) \implies \\ &\tilde{A}\sigma(\tilde{\Lambda}\sigma(\tilde{B}x_1) + \tilde{B}x_2) = A\Pi\sigma(\tilde{\Lambda}\Pi^\top\sigma(Bx_1) + \Pi^\top Bx_2) = A\sigma(\Pi\tilde{\Lambda}\Pi^\top\sigma(Bx_1) + Bx_2). \end{aligned} \tag{52}$$

We use the simplification in the second step to equate to LHS in the first step as follows.

$$\begin{aligned} &A\sigma(\Pi\tilde{\Lambda}\Pi^\top\sigma(Bx_1) + Bx_2) = A\sigma(\Lambda\sigma(Bx_1) + Bx_2) \\ &\implies (\Pi\tilde{\Lambda}\Pi^\top - \Lambda)\sigma(Bx_1) = 0. \end{aligned} \tag{53}$$

Since $\sigma(Bx_1)$ spans a set that has a non-empty interior, we get that $\tilde{\Lambda} = \Pi^\top\Lambda\Pi$.

From the above conditions, we have arrived at $\tilde{\Lambda} = \Pi^\top\Lambda\Pi, \tilde{B} = \Pi^\top B, \tilde{A} = A\Pi$.

We want to show that for all $k \geq 1$

$$Ah_k = \tilde{A}\tilde{h}_k, \tag{54}$$

where $h_k = \sigma(\Lambda h_{k-1} + Bx_k)$ and $\tilde{h}_k = \sigma(\tilde{\Lambda}\tilde{h}_{k-1} + \tilde{B}x_k)$ and $h_0 = \tilde{h}_0 = 0$.

We show the above by principle of induction. Let us consider the base case below.

$$\tilde{A}\tilde{h}_1 = \tilde{A}\sigma(\tilde{B}x_1) = A\Pi\sigma(\Pi^\top Bx_1) = A\sigma(Bx_1) = Ah_1. \tag{55}$$

Suppose $\forall j \leq k$

$$Ah_j = \tilde{A}\tilde{h}_j. \tag{56}$$

We can simplify the above to

$$h_j = \Pi\tilde{h}_j, \forall j \in \{1, \cdots, k\}. \tag{57}$$

From this we get

$$\tilde{A}\tilde{h}_{k+1} = \tilde{A}\sigma(\tilde{\Lambda}\tilde{h}_k + \tilde{B}x_k) = A\Pi\sigma(\Pi^\top\Lambda\Pi\tilde{h}_k + \Pi^\top Bx_k) = A\sigma(\Lambda\Pi\tilde{h}_k + Bx_k) = A\sigma(\Lambda h_k + Bx_k) = Ah_{k+1}. \tag{58}$$

This completes the proof. □

Discussion on Assumption 3 We show how to translate models in Assumption 3 to position-wise non-linearity based transformers studied in the literature Wortsman et al. (2023). Standard attention operates on a set of keys, queries and values $\{k_i, q_i, v_i\}_{i=1}^L$, where $k_i = W_k x_i + b_k$, $q_i = W_q x_i + b_i$, $k_i = W_v x_i + b_v$, and (W_k, b_k), (W_q, b_q), (W_v, b_v) parametrize the key, query and value weight matrices and the biases respectively. The output of causal attention is computed as follows.

$$\begin{aligned} \alpha_i &= \theta([q_i^\top k_1, \cdots, q_i^\top k_i]), \\ h_i &= \sum_{j=1}^{i} \alpha_{ij} v_j, \\ y_i &= \rho(h_i + x_i), \end{aligned} \tag{59}$$

where h_i is the output of the attention, ρ is a non-linear transform (e.g., MLP) and y_i is the final output. Now suppose θ is a position-wise function such as ReLU, sigmoid that were studied in Wortsman et al. (2023), we obtain $h_i = \sum_{j=1}^{i} \alpha_{ij} v_j = \sum_{j=1}^{i} \theta(q_i^\top k_j) v_j$. Define $\phi(x_i, x_j) = \theta(q_i^\top k_j) v_j$. The final output of the one-layer transformer described above is stated as $\rho(\sum_{j \le i} \phi(x_i, x_j) + x_i)$, which is the same form that we assumed in Assumption 3.