On Provable Length and Compositional Generalization

Kartik Ahuja [1] **Amin Mansouri** [2 3]

Abstract

Out-of-distribution generalization capabilities of sequence-to-sequence models can be studied from the lens of two crucial forms of generalization: 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. In this work, we provide first provable guarantees on length and compositional generalization for common sequence-to-sequence models – deep sets, transformers, state space models, and recurrent neural nets – trained to minimize the prediction error. Taking a first principles perspective, we study the realizable case, i.e., the labeling function is realizable on the architecture. We show that limited capacity versions of these different architectures achieve both length and compositional generalization. Across different architectures, we also find that a linear relationship between the learned representation and the representation in the labeling function 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. The accounts of their successes (Bubeck et al., 2023; Gunasekar et al., 2023) as well as their failures, particularly in planning and reasoning (Bubeck et al., 2023; Stechly et al., 2023; Valmeekam et al., 2023), continue to rise. 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 and a superficial form of generalization (Dziri et al., 2024).

A model's ability to perform well across different distribution shifts highlights its ability to learn general algorithms. For models with fixed-dimensional inputs, considerable efforts have led to methods with provable out-of-distribution (OOD) generalization guarantees (Rojas-Carulla et al., 2018; Rame et al., 2022; Chaudhuri et al., 2023; Wiedemer et al., 2023b; Eastwood et al., 2024). For sequence-to-sequence models, a large body of empirical works have investigated OOD generalization (Anil et al., 2022; Jelassi et al., 2023) but we lack efforts that study provable OOD generalization guarantees for these models. These provable guarantees provide a stepping stone towards explaining the success of the existing paradigm and also shine a light on where the existing paradigm fails.

OOD generalization capabilities of sequence-to-sequence models can be studied from the lens of two forms of generalization: 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. While transformers (Vaswani et al., 2017) are the go-to sequence-to-sequence models for many applications, recently, alternative architectures based on state-space models, as noted by Gu et al. (2021), Orvieto et al. (2023b), and Gu & Dao (2023), have shown a lot of promise. This motivates us to study a range of natural sequence-to-sequence architectures, including deep sets (Zaheer et al., 2017), transformers, state space models (SSMs), and recurrent neural networks (RNNs). We focus on the realizable case, i.e., the labeling function is in the hypothesis class of the architecture. Our key contributions and insights are summarized below.

- Limited capacity versions of the different architectures namely deep sets, transformers, SSMs, and RNNs, provably achieve length and compositional generalization.
- A linear relationship between the learned representation and the representation in the labeling function, i.e., linear identification (Roeder et al., 2021), is necessary for length and compositional generalization.
- Through a range of experiments, we show the success of both forms of generalization, matching the predictions of the theory and even going beyond.

[1] Meta FAIR [2] Mila-Quebec AI Institute [3] EPFL. Correspondence to: Kartik Ahuja <kartikahuja@meta.com>.

Work presented at TF2M workshop at ICML 2024, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

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 $y_i \in \mathbb{R}^m$. Observe that this formulation includes both standard downstream tasks such as arithmetic tasks, e.g., $y_i = \sum_{j=1}^{i} x_j$, $y_i = \Pi_{j=1}^{i} x_j$ etc., as well as next-token prediction task, where $\{y_1, \cdots, y_t\} = \{x_2, \cdots, x_{t+1}\}$. We denote a sequence $\{s_1, \cdots, s_t\}$ as $s_{\leq t}$. Consider a sequence $\{x_j\}_{j=1}^{\infty}$, which is sampled from $\mathbb{P}_X$, and a subsequence of this sequence $x_{\leq t} = \{x_j\}_{j=1}^{t}$, whose distribution is denoted as $\mathbb{P}_{X_{\leq t}}$. The label $y_t = f(x_{\leq t})$, where f is the labeling function. The tuple of base distribution and the labeling function is denoted as $\mathcal{P} = \Big\{\mathbb{P}_X, f\Big\}$ and the tuple of base distribution up to length t is denoted as $\mathcal{P}(t) = \Big\{\mathbb{P}_{X_{\leq t}}, f\Big\}$. Define the support of k^{th} token X_k in the sequence sampled from $\mathbb{P}_X$ as $\mathsf{supp}(X_k)$. Given training sequences of length T from $\mathcal{P}(T)$, we are tasked to learn a model from the dataset that takes a 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 T, 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 next.

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, i.e., $h(x_{\leq t})$, 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 the ℓ_2 loss. Given sequences sampled from $\mathcal{P}(T)$, 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 learner aims to find an h^* that solves

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

where $\mathcal{H}$ is the hypothesis class of models. We seek to understand the properties of solutions to (1) through the lens of following questions.

> Can common sequence-to-sequence models $\mathcal{H}$ achieve length & compositional generalization? If so, when do they succeed and when do they fail?

Definition 2.1. 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 length generalizes w.r.t. $\mathcal{P}$.

In the above definition of length generalization, we simply ask if the model generalizes to longer sequences. We drop the phrase w.r.t $\mathcal{P}$ hereafter to avoid repetition. We now define a test distribution that evaluates compositional generalization capabilities. We consider sequences of fixed length T. Define a uniform distribution $\mathbb{Q}_{X_{\leq t}}$ such that the support of $\mathbb{Q}_{X_{\leq t}}$ equals the Cartesian product of the support of each token X_k from $\mathbb{P}_X$, we write this joint support as $\Pi_{j=1}^{t}\mathsf{supp}(X_j)$. In this case as well, the labeling function continues to be f. Hence, we obtain the tuple $\mathcal{Q}(T) = \{\mathbb{Q}_{X_{\leq T}}, f\}$.

Definition 2.2. 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 $\mathcal{Q}(T)$, then it achieves compositional generalization.

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. Basically, we show that at least one solution to (1) does not achieve the desired form of generalization – length or compositional. If there are no constraints on $\mathcal{H}$, we can always construct a function h that is equal to f on all sequences of length up to T but is equal to $f + c$ on sequences longer than T. Therefore, such a function will solve (1) but length generalization will not be achieved. The same argument extends to the case of compositional generalization.

RASP conjecture from (Zhou et al., 2023) (Zhou et al., 2023) propose a conjecture supported by empirical evidence, which delineates the conditions that suffice for length generalization for transformers. The conjecture places three requirements – a) realizability: the task of interest is realizable on the transformer, b) simplicity: the task can be expressed as a short program in RASP-L language, c) diversity: the dataset is sufficiently diverse such that there is no shorter program that achieves in-distribution generalization but not out-of-distribution generalization. In our analysis below, we use assumptions similar to a) and b) but weaker than c) to show length and compositional generalization.

2.1. Deep sets

Deep sets are a natural first choice of architecture to study here. These were introduced in (Zaheer et al., 2017). Informally stated, (Zaheer et al., 2017) show that a large family of permutation-invariant functions can be decomposed as $\rho\big(\sum_{x \in \mathcal{X}} \phi(x)\big)$. Consider the examples of the sum operator or the product operator, which take $\{x_1, x_2, \cdots, x_k\}$ as input, and return the sum $y = \sum_{j=1}^{k} x_j$ or the product $y = \Pi_{j=1}^{k} x_j$. These operations are permutation invariant and can be expressed using the decomposition above. For the sum operator ρ and ϕ are identity and for the product operator $\rho = \exp$ and $\phi = \log$.

Assumption 2.3. Each function in the hypothesis class $\mathcal{H}$ takes a sequence $\{x_1, \cdots, x_i\}$ as input and outputs $h(x_1, \cdots, x_i) = \omega\Big(\sum_{j \le i} \psi(x_j)\Big)$, where ω is a single layer perceptron with continuously differentiable bijective activation (e.g., sigmoid) and ψ is differentiable.

Assumption 2.4. The joint support $\mathsf{supp}(X_{\le i})$ is a regular closed set (in standard topology in $\mathbb{R}^{ni}$) for all $i \le T$.

Linear identification Each architecture that we study in this work relies on a hidden representation that is passsed on to a last non-linear layer to generate the label. Under the realizability condition for deep sets, the labeling function takes the form $f(\mathcal{X}) = \rho(\sum_{x\in\mathcal{X}} \phi(x))$, where $\phi(x)$ is the hidden representation. If the learned deep set is denoted by $\omega(\sum_{x\in\mathcal{X}} \psi(x))$, then the learned hidden representation is $\psi(x)$. If $\psi(x) = A\phi(x)$, then the learned representation is said to *linearly identify* the data generating representation $\phi(x)$. We borrow this definition from the identifiability literature (Khemakhem et al., 2020; Roeder et al., 2021).

Theorem 2.5. *If $\mathcal{H}$ follows Assumption 2.3, the realizability condition holds, i.e., $f \in \mathcal{H}$, $\mathsf{supp}(X_j) = [0, 1]^n$, $\forall j \ge 1$, and Assumption 2.4 holds, then the model trained to minimize the risk in (1) with ℓ_2 loss generalizes to all sequences in the hypercube $[0, 1]^{nt}$, $\forall t \ge 1$ and thus achieves length and compositional generalization.*

The proof is provided in Section A.2.1. In the above result, we work with ω represented by a single layer perceptron. In the above result, we require the support of the marginal distribution of each token to be $[0, 1]^n$. The support of T token length sequence under the joint training distribution can still be a much smaller subset of $[0, 1]^{nT}$. Despite this the model generalizes to all sequences in $[0, 1]^{nt}$ for all t. 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 is necessary for length and compositional generalization (Further details are in Section A.2.1). In Theorem A.4, we extend Theorem 2.5 to ω from C^1-diffeomorphisms.

High capacity deep sets The above results show that limited capacity constraints (Assumption 2.3) on deep sets suffice for length and compositional generalization. What about deep sets with arbitrary capacity, i.e., no constraints on ω and ψ? These express a large family of permutation invariant maps (Zaheer et al., 2017). Suppose $\mathcal{H}$ is the class of all permutation invariant maps and the labeling function $f \in \mathcal{H}$. Consider a map h such that $h = f$ for all sequences of length up to T, and $h = f + c$ otherwise. Observe that h is permutation invariant and also belongs to $\mathcal{H}$. h solves (1) but does not length generalize. Thus high capacity deep sets do not length generalize. A similar argument follows for compositional generalization as well.

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. 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 nonlinearity as ρ and attention layer as ϕ. We obtain the simplest form of causal transformer model as $\rho\Big(\sum_{j=1}^{i} \frac{1}{i} \cdot \phi(x_i, x_j)\Big)$. 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. 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 non-linear attention for a more tractable analysis.

Assumption 2.6. Each function in the hypothesis class $\mathcal{H}$ takes a sequence $\{x_1, \cdots, x_i\}$ as input and outputs $h(x_1, \cdots, x_i) = \omega\Big(\sum_{j\le i} \frac{1}{i} \cdot \psi(x_i, x_j)\Big)$, where ω is a single layer perceptron with continuously differentiable bijective activation (e.g., sigmoid) and ψ is differentiable.

Theorem 2.7. *If $\mathcal{H}$ follows Assumption 2.6, the realizability condition holds, i.e., $f \in \mathcal{H}$, $\mathsf{supp}(X_i, X_j) = [0, 1]^{2n}$, $\forall i \ne j$ and the regular closedness condition in Assumption 2.4 holds, then the model trained to minimize the risk in (1) (with $T \ge 2$) with ℓ_2 loss generalizes to all sequences in the hypercube $[0, 1]^{nt}$, $\forall t \ge 1$ and thus achieves length and compositional generalization.*

The proof is provided in Section A.2.2. We also obtain that a linear relationship between the learned attention representation denoted ψ and attention representation for the labeling function denoted ϕ is necessary for both length and compositional generalization (details in Section A.2.2). We provide extension of Theorem 2.7 from single layer perceptron ω to C^1-diffeomorphism in the Appendix.

High capacity transformers In the above results, we demonstrated that limited capacity transformers (Assumption 2.6, A.8) achieve length and compositional generalization. How about transformers with arbitrary capacity, i.e., no constraint on ω and ψ? If $\psi(x,y) = \psi(\tilde{x},y), \forall x \neq \tilde{x}$, then the decomposition for the causal transformer $\omega\Big(\sum_{j=1}^{i}\frac{1}{i}\cdot\psi(x_i,x_j)\Big)$ becomes $\omega\Big(\sum_{j=1}^{i}\frac{1}{i}\cdot\psi(x_j)\Big)$, which is very similar to deep sets. In such a case, we can use arguments similar to that of arbitrary capacity deep sets and argue that compositional generalization is impossible.

On positional encoding The discussion so far uses the current query and compares it to keys from the past, it does not distinguish the keys based on their positions. For many arithmetic tasks such as computing the median, maximum etc., the positions of keys do not matter but for other downstream tasks such as sentiment classification, the position of the words can be important. In Section A.2.2, we adapt the architecture to incorporate relative positional encodings and show how some of the results extend. We modify the model as $\rho(\sum_{j=1}^{i}\frac{1}{i}\phi_{i-j}(x_i,x_j))$, where $\phi_{i-j}(x_i,x_j)$ computes the query key inner product while taking the relative position $i-j$ into account. We show that if $\phi_{i-j}=0$ for $i-j > T_{\max}$, i.e., two tokens sufficiently far apart do not impact the data generation, then length generalization and compositional generalization are achievable.

2.3. State space models

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 used the lens of linear recurrent layer followed by position-wise non-linearities as the main building block to understand these models. We illustrate the dynamics of these models to show the generation of $x_{\leq t}$ and $y_{\leq t}$ next.

$$\begin{aligned} &h_1 = Bx_1, \cdots, h_t = \Lambda h_{t-1} + Bx_t, \\ &y_1 = \rho(h_1), \cdots, \quad y_t = \rho(h_t), \end{aligned} \tag{2}$$

where $h_t \in \mathbb{R}^k$ is hidden state at t, $\Lambda \in \mathbb{R}^{k\times k}, B \in \mathbb{R}^{k\times n}$ and $\rho: \mathbb{R}^k \rightarrow \mathbb{R}^m$. Observe that $h_t = \sum_{j=0}^{t-1}\Lambda^j Bx_{t-j}$.

Assumption 2.8. Each function in the hypothesis class $\mathcal{H}$ takes a sequence $\{x_1,\cdots,x_i\}$ as input and outputs $h(x_1,\cdots,x_i) = \omega\Big(\sum_{j=0}^{i-1}\Lambda^j Bx_{i-j}\Big)$, where $\omega: \mathbb{R}^k \rightarrow \mathbb{R}^m$ is a C^1-diffeomorphism, B and Λ are square invertible.

Theorem 2.9. *If $\mathcal{H}$ follows Assumption 2.8, and the realizability condition holds, i.e., $f \in \mathcal{H}$, and a further condition on the support, i.e., Assumption A.13, holds, then the model trained to minimize the risk in (1) with ℓ_2 loss ($T \geq 2$) achieves length and compositional generalization.*

The proof is provided in Section A.2.3. In the above result as well, linear representation identification, i.e., the predicted hidden state $\tilde{h}_t$ and the true hidden state h_t bear a linear relationship, turns out to be necessary for length and compositional generalization (See Section A.2.3 for details).

High capacity SSMs In the above result, we have shown that SSMs with limited capacity can achieve length generalization. How about SSMs with arbitrary capacity, i.e., no constraint on Λ, B and ω? (Orvieto et al., 2023a) showed that SSMs with arbitrary capacity (i.e., with appropriately large Λ and B matrices) can approximate a sequence-to-sequence mapping up to some length with arbitrary precision. Consider the true labeling function f and another function h, which is equal to f for all sequences of length up to T and $f+c$ for larger lengths. As a result, h is a solution to (1) for SSMs with arbitrary capacity and it does not achieve length generalization. The same argument extends to compositional generalization.

2.4. Vanilla recurrent neural networks

Standard RNNs have a non-linear recurrence unlike the linear recurrence studied in the previous section. We use the same notation as the previous section and only add an activation for non-linear recurrence. We illustrate the dynamics to show the generation of $x_{\leq t}$ and $y_{\leq t}$ below.

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

Assumption 2.10. Each function in $\mathcal{H}$ used by the learner is a vanilla RNN of the form (3), where the position-wise non-linearity is a single layer perceptron $\sigma \circ A$, and Λ, B govern the hidden state dynamics (as in (3)). A, Λ, B are square invertible matrices, and σ is the sigmoid activation.

Theorem 2.11. *If $\mathcal{H}$ follows Assumption 2.10, and the realizability condition holds, i.e., $f \in \mathcal{H}$ and regular closedness condition in Assumption 2.4 holds, then the model trained to minimize the risk in (1) with ℓ_2 loss (with $T \geq 2$) achieves length and compositional generalization.*

The proof is provided in Section A.2.4. Here also we find that a linear relationship between predicted hidden state $\tilde{h}_t$ and true hidden state h_t is necessary for both length and compositional generalization (See Section A.2.4 for details). In fact, the relationship is a permutation map. Similar to previous sections, we can show that high capacity RNNs cannot achieve length and compositional generalization. Finally, we point the reader to Section A.3 for our experimental findings, which is not included here due to space limit.

3. Conclusion

In this work, we formalized first provable length generalization and compositional generalization in sequence-to-sequence models. This effort gives way to a foundation for the recently proposed RASP conjecture (Zhou et al., 2023).

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

Abbe, E., Bengio, S., Lotfi, A., and Rizk, K. Generalization on the unseen, logic reasoning and degree curriculum. *arXiv preprint arXiv:2301.13105*, 2023.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Anil, C., Wu, Y., Andreassen, A., Lewkowycz, A., Misra, V., Ramasesh, V., Slone, A., Gur-Ari, G., Dyer, E., and Neyshabur, B. Exploring length generalization in large language models. *Advances in Neural Information Processing Systems*, 35:38546–38556, 2022.

Arora, S. and Goyal, A. A theory for emergence of complex skills in language models. *arXiv preprint arXiv:2307.15936*, 2023.

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

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

Brady, J., Zimmermann, R. S., Sharma, Y., Schölkopf, B., Von Kügelgen, J., and Brendel, W. Provably learning object-centric representations. In *International Conference on Machine Learning*, pp. 3038–3062. PMLR, 2023.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Chaudhuri, K., Ahuja, K., Arjovsky, M., and Lopez-Paz, D. Why does throwing away data improve worst-group error? In *International Conference on Machine Learning*, pp. 4144–4188. PMLR, 2023.

Dziri, N., Lu, X., Sclar, M., Li, X. L., Jiang, L., Lin, B. Y., Welleck, S., West, P., Bhagavatula, C., Le Bras, R., et al. Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems*, 36, 2024.

Eastwood, C., Singh, S., Nicolicioiu, A. L., Vlastelica Pogančić, M., von Kügelgen, J., and Schölkopf, B. Spuriosity didn't kill the classifier: Using invariant predictions to harness spurious features. *Advances in Neural Information Processing Systems*, 36, 2024.

Gordon, J., Lopez-Paz, D., Baroni, M., and Bouchacourt, D. Permutation equivariant models for compositional generalization in language. In *International Conference on Learning Representations*, 2019.

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

Gu, A., Goel, K., and Ré, C. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C. C. T., Del Giorno, A., Gopi, S., Javaheripi, M., Kauffmann, P., de Rosa, G., Saarikivi, O., et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023.

Hua, W., Dai, Z., Liu, H., and Le, Q. Transformer quality in linear time. In *International Conference on Machine Learning*, pp. 9099–9117. PMLR, 2022.

Hupkes, D., Dankers, V., Mul, M., and Bruni, E. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67: 757–795, 2020.

Jelassi, S., d'Ascoli, S., Domingo-Enrich, C., Wu, Y., Li, Y., and Charton, F. Length generalization in arithmetic transformers. *arXiv preprint arXiv:2306.15400*, 2023.

Kazemnejad, A., Padhi, I., Natesan Ramamurthy, K., Das, P., and Reddy, S. The impact of positional encoding on length generalization in transformers. *Advances in Neural Information Processing Systems*, 36, 2024.

Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pp. 2207–2217. PMLR, 2020.

Kim, N. and Linzen, T. Cogs: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9087–9105, 2020.

Lachapelle, S., Mahajan, D., Mitliagkas, I., and Lacoste-Julien, S. Additive decoders for latent variables identification and cartesian-product extrapolation. *arXiv preprint arXiv:2307.02598*, 2023.

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

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

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

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

Orvieto, A., De, S., Gulcehre, C., Pascanu, R., and Smith, S. L. On the universality of linear recurrences followed by nonlinear projections. *arXiv preprint arXiv:2307.11888*, 2023a.

Orvieto, A., Smith, S. L., Gu, A., Fernando, A., Gulcehre, C., Pascanu, R., and De, S. Resurrecting recurrent neural networks for long sequences. *arXiv preprint arXiv:2303.06349*, 2023b.

Rame, A., Kirchmeyer, M., Rahier, T., Rakotomamonjy, A., Gallinari, P., and Cord, M. Diverse weight averaging for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 35:10821–10836, 2022.

Roeder, G., Metz, L., and Kingma, D. On linear identifiability of learned representations. In *International Conference on Machine Learning*, pp. 9030–9039. PMLR, 2021.

Rojas-Carulla, M., Schölkopf, B., Turner, R., and Peters, J. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018.

Shaw, P., Uszkoreit, J., and Vaswani, A. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018.

Shen, K., Guo, J., Tan, X., Tang, S., Wang, R., and Bian, J. A study on relu and softmax in transformer. *arXiv preprint arXiv:2302.06461*, 2023.

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

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

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

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

Wiedemer, T., Brady, J., Panfilov, A., Juhos, A., Bethge, M., and Brendel, W. Provable compositional generalization for object-centric learning. *arXiv preprint arXiv:2310.05327*, 2023a.

Wiedemer, T., Mayilvahanan, P., Bethge, M., and Brendel, W. Compositional generalization from first principles. *arXiv preprint arXiv:2307.05596*, 2023b.

Wortsman, M., Lee, J., Gilmer, J., and Kornblith, S. Replacing softmax with relu in vision transformers. *arXiv preprint arXiv:2309.08586*, 2023.

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

Xu, Z., Niethammer, M., and Raffel, C. A. Compositional generalization in unsupervised compositional representation learning: A study on disentanglement and emergent language. *Advances in Neural Information Processing Systems*, 35:25074–25087, 2022.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., and Smola, A. J. Deep sets. *Advances in neural information processing systems*, 30, 2017.

Zhang, M., He, J., Lei, S., Yue, M., Wang, L., and Lu, C.-T. Can llm find the green circle? investigation and human-guided tool manipulation for compositional generalization. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11996–12000. IEEE, 2024.

Zhou, H., Bradley, A., Littwin, E., Razin, N., Saremi, O., Susskind, J., Bengio, S., and Nakkiran, P. What algorithms can transformers learn? a study in length generalization. *arXiv preprint arXiv:2310.16028*, 2023.

A. Appendix

Contents

We organize the Appendix as follows.

A.1. Related Works

Length generalization In the field of length generalization, many important empirical insights have been synthesized over the last few years. (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 different settings for transformer-based models. 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. In (Kazemnejad et al., 2024), the authors did a comprehensive study of different positional embeddings and provided evidence to show that explicit use of positional encodings is perhaps not essential. Our work is both related and inspired by some of the recent findings in (Zhou et al., 2023). In this work, the authors proposed the RASP conjecture. The conjecture delineates the tasks where transformers excel or fall short in length generalization, emphasizing the necessity of task simplicity and data diversity. While (Zhou et al., 2023) provide empirical evidence for the conjecture, our work formalizes and proves simpler versions of the conjecture for a range of architectures with minimal assumptions on data diversity.

On the theoretical side of length generalization, 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 & Liu, 2023), the authors leverage directed acyclic graphs (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 & Baroni, 2018; Loula et al., 2018; Gordon et al., 2019; Hupkes et al., 2020; Kim & Linzen, 2020; Xu et al., 2022; Arora & Goyal, 2023; Zhang et al., 2024), is too expansive to address comprehensively here. However, we reference several pertinent works from whom we borrow the formal definition of compositionality. Recent studies, notably those by (Wiedemer et al., 2023a;b; Lachapelle et al., 2023; Brady 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 similar to these studies but diverges by centering on common sequence-to-sequence architectures as the main 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 $\forall t$. 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 training sequence length. We remind the reader of 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 A.1. *Let $\mathcal{X} \subseteq \mathbb{R}^n$. If $f : \mathcal{X} \rightarrow \mathbb{R}^m$ and $g : \mathcal{X} \rightarrow \mathbb{R}^m$ 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 converges 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

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

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 $\lim_{k\rightarrow\infty} x_k = x$, 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

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

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, $\lim_{k\rightarrow\infty} x_k = x$. 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

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

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

□

A.2.1. Deep sets

In this section, we provide the proofs for length and compositional generalization for deep sets. We restate the theorems from the main body for convenience of the reader. In what follows, we remind the reader that we denote the labeling function $f(\mathcal{X}) = \rho(\sum_{x\in\mathcal{X}} \phi(x))$ and the function learned is denoted as $h(\mathcal{X}) = \omega(\sum_{x\in\mathcal{X}} \psi(x))$.

Theorem 2.5. *If $\mathcal{H}$ follows Assumption 2.3, the realizability condition holds, i.e., $f \in \mathcal{H}$, $\mathsf{supp}(X_j) = [0,1]^n,\ \forall j \geq 1$, and Assumption 2.4 holds, then the model trained to minimize the risk in (1) with ℓ_2 loss generalizes to all sequences in the hypercube $[0,1]^{nt},\ \forall t \geq 1$ and thus achieves length and compositional generalization.*

Proof. Consider any h that solves (1). Since ℓ is ℓ_2 loss and realizability condition holds, f is one of the optimal solutions to (1). For all $x_{\leq T} \in \mathsf{supp}(X_{\leq T})$ 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 (1), i.e., $\mathbb{E}[\|h - f\|^2] = 0$ and from Theorem 1.6.6. (Ash & Doléans-Dade, 2000). Since $\mathsf{supp}(X_{\leq T})$ is regular closed, f, h are both continuously differentiable, we can use Lemma A.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 2.3 to get

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

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

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

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 A.1). For all $x_r \in \mathsf{supp}(X_r)$, 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. We now take the derivative w.r.t some component x^k of vector $x = [x^1, \cdots, x^n]$. Denote the components other than k as $x^{-k} = x \setminus x^k$. 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 $x^k \in [0,1]$,

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

The above condition is true of all $k \in \{1, \cdots, n\}$. Hence, we can deduce that for all $x \in [0,1]^n$ 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 any sequence $x_{\leq \tilde{T}}$ from $[0, 1]^{n\tilde{T}}$. 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 (21) in the simplification above. From the above, we can conclude that h continues to be optimal for distribution $\mathbb{P}_{X_{\leq \tilde{T}}}$.

□

Linear identification From the fact that model achieves generalization on sequences of length T, we obtain (21). As a result, we can state that relationship in (21) is necessary for length generalization and compositional generalization. Suppose the output layer matrix A has a left inverse. If that is the case, then we can simplify (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 (Khemakhem et al., 2020; Roeder et al., 2021). As a result, this condition as necessary for both length and compositional generalization.

Remarks A few remarks and observations from the proof are in order. Firstly, observe that we do not require ϕ and ψ to have the same output dimension for the above to proof to go through. Secondly, in Theorem 2.5, 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. Finally, we make an observation in this result, which would apply to all the subsequent theorems. The definition of compositional generalization requires generalization to the Cartesian product over sequences of length T, where T is the training length. Since our model generalizes to the hypercube $[0, 1]^{nt}, \forall t$, we achieve compositional generalization even beyond the training lengths.

Extending Theorem 2.5 to ω from C^1-diffeomorphisms class

Assumption A.2. Each function in $\mathcal{H}$ is expressed as $h(x_1, \cdots, x_i) = \omega(\sum_{j=1}^{i}\psi(x_j))$, where ω is a C^1-diffeomorphism.

Assumption A.3. The joint support $\mathsf{supp}(X_{\leq i})$ is a regular closed set for all $i \leq T$. The support of all tokens is equal, i.e., $\mathsf{supp}(X_j) = [0, 1]^n$, where $j \geq 1$. The support of $[\phi(X_1), \phi(X_2)]$ is $\mathbb{R}^{2m}$, where ϕ is the embedding function for the labeling function $f(\mathcal{X}) = \rho(\sum_{x\in\mathcal{X}}\phi(x))$.

We provide a remark on the assumption and where it is used following the proof of the next theorem.

Theorem A.4. *If $\mathcal{H}$ follows Assumption A.2, the realizability condition holds, i.e., $f \in \mathcal{H}$, and a further assumption on the support (Assumption A.3) holds, then the model trained to minimize the risk in (1) (with $T \geq 2$) with ℓ_2 loss generalizes to all sequences in $[0,1]^{nt}, \forall t \geq 1$ and thus 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 A.1) to obtain the following. For all $x_{\leq i} \in \mathsf{supp}(X_{\leq i})$

$$\begin{aligned}&\omega\Big(\sum_{j\leq i}\psi(x_j)\Big) = \rho\Big(\sum_{j\leq i}\phi(x_j)\Big) \implies \sum_{j\leq i}\psi(x_j) = \omega^{-1}\circ\rho\Big(\sum_{j\leq i}\phi(x_j)\Big) \implies \\ &\sum_{j\leq i}\psi(x_j) = a\Big(\sum_{j\leq i}\phi(x_j)\Big),\end{aligned} \tag{23}$$

where $a = \omega^{-1}\circ\rho$. In the above simplification, we used the parametric form for the true labeling function and the learned labeling function and use the invertibility of ω. Let us consider the setting when $i = 1$. In that case summation involves only one term. Substitute $x_1 = x$. We obtain $\forall x \in [0,1]^n$,

$$\psi(x) = a(\phi(x)). \tag{24}$$

The above expression implies that ψ bijectively identifies ϕ. Let us consider the setting when $i = 2$. Substitute $x_1 = x$ and $x_2 = y$. We obtain

$$a(\phi(x)) + a(\phi(y)) = a\big(\phi(x) + \phi(y)\big). \tag{25}$$

We now use the that assumption $[\phi(x), \phi(y)]$ spans $\mathbb{R}^{2m}$, where $\phi(x)$ and $\phi(y)$ individually span $\mathbb{R}^m$. Substitute $\phi(x) = \alpha$ and $\phi(y) = \beta$. We obtain $\forall \alpha \in \mathbb{R}^m, \forall \beta \in \mathbb{R}^m$

$$a(\alpha) + a(\beta) = a\big(\alpha + \beta\big). \tag{26}$$

Observe that $a(0) = 0$ (substitute $\alpha = \beta = 0$ in the above).

We use (26) 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.

From the identity above, we want to show that (51) $a(p\alpha) = pa(\alpha)$, where p is some integer.

Substitute $\beta = -\alpha$ in $a(\alpha+\beta) = a(\alpha) + a(\beta)$. We obtain $a(0) = a(\alpha) + a(-\alpha) \implies a(-\alpha) = -a(\alpha)$. Suppose p is a positive integer. We simplify $a(p\alpha)$ as follows $a(\alpha + (p-1)\alpha) = a(\alpha) + a((p-1)\alpha)$. Repeating this simplification, we get $a(p\alpha) = pa(\alpha)$. Suppose p is a negative integer. We can write $a(p\alpha) = a(-p \times -\alpha) = -pa(-\alpha)$. Since $a(-\alpha) = -a(\alpha)$, we get $a(p\alpha) = pa(\alpha)$.

Suppose c is some rational number, i.e., $c = p/q$, where p and q are non-zero integers. We already know $a(p\alpha) = pa(\alpha)$. Further, we obtain

$a(q\frac{1}{q}\alpha) = qa(\frac{1}{q}\alpha) \implies a(\frac{1}{q}\alpha) = \frac{1}{q}a(\alpha)$, 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 the fact 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

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

Observe that $a : \mathbb{R}^m \rightarrow \mathbb{R}^m$ and for any $\alpha, \beta \in \mathbb{R}^m$ $a(\alpha+\beta) = a(\alpha) + a(\beta)$ and $a(c\alpha) = ca(\alpha)$. From the definition of a linear map it follows that a is linear. As a result, we can write $\forall x \in [0,1]^n$

$$\psi(x) = A(\phi(x)) \tag{27}$$

Observe that a is invertible because both ρ and ω are invertible. As a result, we know that A is an invertible matrix. From this we get

$$\phi(x) = A^{-1}\psi(x) = C(\psi(x)) \tag{28}$$

For all $z \in \mathbb{R}^m$, we obtain

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

Let us consider any sequence $x_{\leq \tilde{T}} \in [0,1]^{n\tilde{T}}$. We use the above conditions

$$\omega\Big(\sum_{j\leq \tilde{T}} \psi(x_j)\Big) = \rho(C\sum_{j\leq \tilde{T}} \psi(x_j)) = \rho\Big(\sum_{j\leq \tilde{T}} \phi(x_j)\Big)$$

Thus we obtain length and compositional generalization.

□

Remark on Assumption A.3 In Assumption A.3, we require that the support of $[\phi(X_1), \phi(X_2)]$ is $\mathbb{R}^{2m}$. This assumption is used in the proof in equation (26). We used this assumption to arrive at $a(\alpha+\beta) = a(\alpha) + a(\beta), \forall \alpha, \beta \in \mathbb{R}^m$. We then used continuity of a to conclude a is linear. Now suppose $[\phi(X_1), \phi(X_2)]$ is some subset $\mathcal{Z} \subseteq \mathbb{R}^{2m}$. We believe that it is possible to extend the result to more general $\mathcal{Z}$, it might still be possible to arrive at a is linear. We leave this investigation to future work.

Remark on expressivity under Assumption A.2 and Assumption A.3 Assumption A.3 requires ω is a C^1-diffeomorphism. Suppose the label is one dimensional, i.e., $m = 1$. From Assumption A.3 output dimension of ϕ is restricted to be one dimensional. Consider the map $h(x_1, \cdots, x_i) = \rho(\sum_{j\leq i} \phi(x_j))$. The output dimension of ϕ is required to grow with sequence length to express all permutation invariant maps (See Theorem 7 in (Zaheer et al., 2017)). Thus by restricting the output dimension of ϕ to one, we cannot express all the permutation invariant maps.

Product operator Consider the product operator $y_i = \Pi_{j=1}^{i} x_i$, where each $x_i > 0$. Observe that we can rewrite this as $y_i = \exp(\sum_{j=1}^{i} \log(x_j))$. This operator is realizable on deep sets from hypothesis class described by Assumption A.2 with $\omega = \exp$ and $\psi = \log$. In Assumption A.3, we require the support of $[\phi(X_1), \phi(X_2)]$ to be $\mathbb{R}^2$. We let the support of X_1 and X_2 be $(0, \infty)$. In Assumption A.3 we require that the support of each token was equal to $[0, 1]$. However, the proof of Theorem A.4 still goes through even if support is $(0, \infty)$. Hence, we can use Theorem A.4 to conclude that deep sets trained to predict the output of multiplication can multiply longer sequences and also multiply new token combinations.

A.2.2. Transformers

In this section, we provide the proofs for length and compositional generalization for transformers.

We restate the theorems from the main body for convenience of the reader. In what follows, we want to remind the reader we denote the labeling function $f(x_1, \cdots, x_i) = \rho(\sum_{j\leq i} \phi(x_i, x_j))$ and the function learned is denoted as $h(x_1, \cdots, x_i) = \omega(\sum_{j\leq i} \psi(x_i, x_j))$. Theorem 2.7 presents the results for generalization where ω is a single layer perceptron. Theorem A.6

adapts it to incorporate positional encodings. Theorem A.9 extends the setting of Theorem 2.7 to ω that come from class of C^1-diffeomorphisms. Theorem A.12 adapts Theorem A.9 to incorporate positional encodings.

Theorem 2.7. *If $\mathcal{H}$ follows Assumption 2.6, the realizability condition holds, i.e., $f \in \mathcal{H}$, $\mathsf{supp}(X_i, X_j) = [0,1]^{2n},\ \forall i \neq j$ and the regular closedness condition in Assumption 2.4 holds, then the model trained to minimize the risk in (1) (with $T \geq 2$) with ℓ_2 loss generalizes to all sequences in the hypercube $[0,1]^{nt},\ \forall t \geq 1$ and thus achieves length and compositional generalization.*

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

$$h(x_{\leq i}) = f(x_{\leq i}). \tag{29}$$

The above follows from the fact that h solves (1), i.e., $\mathbb{E}[\|h - f\|^2] = 0$ and from Theorem 1.6.6. (Ash & Doléans-Dade, 2000). Since $\mathsf{supp}(X_{\leq i})$ is regular closed, f, h are both continuously differentiable, we can use Lemma A.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 2.6 to get

$$\omega\Big(\sum_{k\leq i} \frac{1}{i} \cdot \psi(x_i, x_k)\Big) = \rho\Big(\sum_{k\leq i} \frac{1}{i} \cdot \phi(x_i, x_k)\Big). \tag{30}$$

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 $x_{\leq i} \in \mathsf{supp}(X_{\leq i})$,

$$\sigma\Big(A\sum_{k\leq i} \frac{1}{i} \cdot \psi(x_i, x_k)\Big) = \sigma\Big(B\sum_{k\leq i} \frac{1}{i} \cdot \phi(x_i, x_k)\Big) \implies A\sum_{k\leq i} \psi(x_i, x_k) = B\sum_{k\leq i} \phi(x_i, x_k). \tag{31}$$

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

$$\begin{aligned} &\nabla_{x_j}\Big(A\sum_{k\leq i} \psi(x_i, x_k)\Big) = \nabla_{x_j}\Big(B\sum_{k\leq i} \phi(x_i, x_k)\Big) \implies \\ &A\nabla_{x_j}\psi(x_i, x_j) = B\nabla_{x_j}\phi(x_i, x_j), \end{aligned} \tag{32}$$

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 A.6). 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 $x_i \in [0,1]^n, x_j \in [0,1]^n$,

$$A\frac{\partial \psi(x_i, x_j)}{\partial x_j^k} = B\frac{\partial \phi(x_i, x_j)}{\partial x_j^k}. \tag{33}$$

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

$$\begin{aligned} &A\psi\big(x_i, [x_j^k, x_j^{-k}]\big) = B\phi\big(x_i, [x_j^k, x_j^{-k}]\big) + C_k\big(x_i, x_j^{-k}\big) = \\ &A\psi(x_i, x_j) = B\phi(x_i, x_j) + C_k(x_i, x_j^{-k}). \end{aligned} \tag{34}$$

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 $q, k \in \{1, \cdots, d\}$,

$$A\psi(x_i, x_j) - B\phi(x_i, x_j) = C_k(x_i, x_j^{-k}) = C_q(x_i, x_j^{-q}). \tag{35}$$

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

$$\frac{\partial C_k(x_i, x_j^{-k})}{\partial x_j^q} = \frac{\partial C_q(x_i, x_j^{-q})}{\partial x_j^q} = 0. \tag{36}$$

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 $x_i \in [0,1]^n, x_j \in [0,1]^n$

$$A\psi(x_i, x_j) = B\phi(x_i, x_j) + C(x_i). \tag{37}$$

If we substitute $x_i = x_j = x$, then the above equality extends for $i = j$ and thus we get

$$A\psi(x_i, x_i) = B\phi(x_i, x_i) + C(x_i). \tag{38}$$

Substitute the above (37), (38) into $A\sum_{k\leq i}\psi(x_i, x_k) = B\sum_{k\leq i}\phi(x_i, x_k)$ to obtain

$$B\sum_{k\leq i}\phi(x_i, x_k) + (i)C(x_i) = B\sum_{k\leq i}\phi(x_i, x_k) \implies C(x_i) = 0. \tag{39}$$

Thus we obtain

$$\forall x_i \in [0,1]^n, x_j \in [0,1]^n \;\; A\psi(x_i, x_j) = B\phi(x_i, x_j). \tag{40}$$

We now consider any sequence $x_{\leq \tilde{T}} \in [0,1]^{n\tilde{T}}$. The prediction made by h is

$$h(x_{\leq \tilde{T}}) = \sigma\Big(A\sum_{j\leq \tilde{T}}\psi(x_{\tilde{T}}, x_j)\Big) = \sigma\Big(B\sum_{j\leq \tilde{T}}\phi(x_{\tilde{T}}, x_j)\Big) = f(x_{\leq \tilde{T}}) \tag{41}$$

We use (40) in the simplification above. From the above, we can conclude that h continues to be optimal for all sequences in $[0,1]^{n\tilde{T}}$.

□

Linear identification Observe that we arrive at equation (40) by starting from the condition that the model generalizes on sequences of length up to T. If the weight matrix in the output layer A is left invertible, then we obtain that $\psi(x_i, x_j) = A^{-1}B\phi(x_i, x_j)$, which implies that linear representation identification is necessary for both compositional and length generalization (this argument is based on the same reasoning as the previous proof of Theorem 2.5).

On the absence of labels at all lengths from $t = 1$ to $t = T$ A few important remarks are to follow. In the proof above, we do not require to observe all the labels from $t = 1$ to $t = T$, where $T \geq 2$. The proof goes through provided we observe data at two different lengths.

Positional encoding In what follows, we extend the above result (Theorem 2.7) to incorporate positional encoding. We start with extension of the hypothesis class to incorporate positional encoding.

Assumption A.5. Each function in the hypothesis class $\mathcal{H}$ used by the learner is given as $h(x_1, \cdots, x_i) = \omega\Big(\sum_{j\leq i} \frac{1}{i}\psi_{i-j}(x_i, x_j)\Big)$, where ω is a single layer perceptron with continuously differentiable bijective activation (e.g., sigmoid) and each ψ_k is a map that is differentiable. Also, $\psi_k = 0$ for $k \geq T_{\mathsf{max}}$, i.e., two tokens that are sufficiently far apart do not interact.

In the above assumption, we incorporate relative positional encodings by making the function ψ_{i-j} depend on the relative positional difference between token x_i and token x_j. We would like to emphasize the reasons why we assume that the tokens that are sufficiently far apart do not interact. Suppose $T_{\mathsf{max}} = \infty$, which implies tokens at all positions interact. As a result, during training since we only see sequences of finite length T, we will not see the effect of interactions of tokens that are separated at a distance larger than T on the data generation, which makes it impossible to learn anything about ϕ_{i-j}, where $i - j \geq T - 1$.

In the theorem that follows, we show that we can achieve length and compositional generalization for the above hypothesis class.

Theorem A.6. *If $\mathcal{H}$ follows Assumption A.5, the realizability condition holds, i.e., $f \in \mathcal{H}$, $\mathsf{supp}(X_i, X_j) = [0, 1]^{2n}$, $\forall i \neq j \in \{1, \cdots, \infty\}$, the regular closedness condition in Assumption 2.4 holds and $T \geq T_{\mathsf{max}} \geq 2$, then the model trained to minimize the risk in (1) with ℓ_2 loss generalizes to all sequences in the hypercube $[0, 1]^{nt}$, $\forall t$ and thus achieves length and compositional generalization.*

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

$$h(x_{\leq i}) = f(x_{\leq i}). \tag{42}$$

The above follows from the fact that h solves (1), i.e., $\mathbb{E}[\|h-f\|^2] = 0$ and from Theorem 1.6.6. (Ash & Doléans-Dade, 2000). Since $\mathsf{supp}(X_{\leq i})$ is regular closed, f, h are both continuously differentiable, we can use Lemma A.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 $f(x_{\leq i}) = \rho\Big(\sum_{k\leq i} \phi_{i-k}(x_i, x_k)\Big)$. We substitute the parametric forms from Assumption 2.6 to get

$$\omega\Big(\sum_{k\leq i} \frac{1}{i} \cdot \psi_{i-k}(x_i, x_k)\Big) = \rho\Big(\sum_{k\leq i} \frac{1}{i} \cdot \phi_{i-k}(x_i, x_k)\Big). \tag{43}$$

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 $x_{\leq i} \in \mathsf{supp}(X_{\leq i})$,

$$\begin{aligned} &\sigma\Big(A\sum_{k\leq i} \frac{1}{i} \cdot \psi_{i-k}(x_i, x_k)\Big) = \sigma\Big(B\sum_{k\leq i} \frac{1}{i} \cdot \phi_{i-k}(x_i, x_k)\Big) \implies \\ &A\sum_{k\leq i} \psi_{i-k}(x_i, x_k) = B\sum_{k\leq i} \phi_{i-k}(x_i, x_k). \end{aligned} \tag{44}$$

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 A.1). The equality holds true for all $i \leq T$.

From the above, we can use $i = 1$ and obtain

$$A\psi_0(x_1, x_1) = B\phi_0(x_1, x_1), \forall x_1 \in [0, 1]^n.$$

From $i = 2$, we obtain

$$A\psi_0(x_2, x_2) + A\psi_1(x_2, x_1) = B\phi_0(x_2, x_2) + B\phi_1(x_2, x_1), \forall x_1 \in [0,1]^n, x_2 \in [0,1]^n$$

Combining the two conditions we get

$$A\psi_1(x_2, x_1) = B\phi_1(x_2, x_1), \forall x_1 \in [0,1]^n, x_2 \in [0,1]^n.$$

We can use this argument and arrive at

$$A\psi_{i-1}(x_i, x_1) = B\phi_{i-1}(x_i, x_1), \forall x_i \in [0,1]^n, x_1 \in [0,1]^n, \forall i \leq T.$$

Thus we obtain

$$\forall i - j \leq T - 1, \forall x_i \in [0,1]^n, x_j \in [0,1]^n, \quad A\psi_{i-j}(x_i, x_j) = B\phi_{i-j}(x_i, x_j). \tag{45}$$

From Assumption A.5 and $T \geq T_{\mathsf{max}}$, we already know that

$$\forall i - j \geq T, \forall x_i \in [0,1]^n, x_j \in [0,1]^n, \quad A\psi_{i-j}(x_i, x_j) = B\phi_{i-j}(x_i, x_j) = 0. \tag{46}$$

If A is left invertible, then the above condition implies that linear representation identification is necessary for both compositional and length generalization.

We now consider any sequence $x_{\leq \tilde{T}} \in [0,1]^{n\tilde{T}}$. The prediction made by h is

$$h(x_{\leq \tilde{T}}) = \sigma\Big(A \sum_{j \leq \tilde{T}} \psi_{\tilde{T}-j}(x_{\tilde{T}}, x_j)\Big) = \sigma\Big(B \sum_{j \leq \tilde{T}} \phi_{\tilde{T}-j}(x_{\tilde{T}}, x_j)\Big) = f(x_{\leq \tilde{T}}) \tag{47}$$

We use (40) in the simplification above. From the above, we can conclude that h continues to be optimal for all sequences in $[0,1]^{n\tilde{T}}$.

□

Assumption A.7. The joint support $\mathsf{supp}(X_{\leq i})$ is a regular closed set for all $i \leq T$. The support of all pairs of tokens is equal, i.e., $\mathsf{supp}(X_i, X_j) = [0,1]^{2n}$, where $i \neq j$, $i \geq 1, j \geq 1$. The support of $[\phi(X_1, X_2), \phi(X_1, X_3)]$ is $\mathbb{R}^{2m}$, where ϕ is the embedding function for the labeling function $\rho(\sum_{j \leq i} \phi(x_i, x_j))$.

Assumption A.8. Each function in $\mathcal{H}$ takes a sequence $\{x_1, \cdots, x_i\}$ as input and outputs $h(x_1, \cdots, x_i) = \omega(\sum_{j=1}^{i-1} \frac{1}{i-1} \psi(x_i, x_j))$, where $\omega : \mathbb{R}^m \rightarrow \mathbb{R}^m$ is a C^1-diffeomorphism, $\omega(0) = 0$.

Theorem A.9. *If $\mathcal{H}$ follows Assumption A.8, the realizability condition holds, i.e., $f \in \mathcal{H}$, and a further assumption on the support (Assumption A.7) holds, then the model trained to minimize the risk in (1) (with $T \geq 3$) with ℓ_2 loss generalizes to all sequences in $[0,1]^{nt}, \forall t \geq 1$ and thus 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 $h(x_{\leq i}) = f(x_{\leq i}), \forall i \leq T$ 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 A.1) to obtain the following. For all $x_{\leq i} \in \mathsf{supp}(X_{\leq i})$

$$\begin{aligned}
&\omega\Big(\sum_{j<i} \frac{1}{i-1} \cdot \psi(x_i, x_j)\Big) = \rho\Big(\sum_{j<i} \frac{1}{i-1} \cdot \phi(x_i, x_j)\Big) \implies \\
&\sum_{j<i} \frac{1}{i-1} \psi(x_i, x_j) = \omega^{-1} \circ \rho\Big(\sum_{j<i} \frac{1}{i-1} \cdot \phi(x_i, x_j)\Big) \implies \\
&\sum_{j<i} \frac{1}{i-1} \psi(x_i, x_j) = a\Big(\sum_{j<i} \frac{1}{i-1} \phi(x_i, x_j)\Big),
\end{aligned} \tag{48}$$

where $a = \omega^{-1} \circ \rho$. In the above simplification, we used the parametric form for the true labeling function and the learned labeling function and use the invertibility of ω. 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 $\forall x \in [0, 1]^n, y \in [0, 1]^n$,

$$\psi(x, y) = a(\phi(x, y)). \tag{49}$$

The above expression implies that ψ bijectively identifies ϕ. Let us consider the setting when $i = 3$ (this is possible since $T \geq 3$). We substitute $x_3 = x$, $x_2 = y$, $x_1 = z$ and obtain

$$\frac{1}{2}\Big[a(\phi(x, y)) + a(\phi(x, z))\Big] = a\big(\frac{1}{2}\big(\phi(x, y) + \phi(x, z)\big)\big). \tag{50}$$

Substitute $\phi(x, y) = \alpha$ and $\phi(x, z) = \beta$. In the simplifcation that follows, we use the that assumption $[\phi(x, y), \phi(x, z)]$ spans $\mathbb{R}^{2m}$, where $\phi(x, y)$ and $\phi(x, z)$ individually span $\mathbb{R}^m$.

$$\frac{1}{2}(a(\alpha) + a(\beta)) = a\big(\frac{1}{2}(\alpha + \beta)\big). \tag{51}$$

Observe that $a(0) = 0$ because $\omega^{-1} \circ \rho(0) = 0$ because $\omega^{-1}(0) = \rho(0) = 0$.

$$\begin{aligned} &\frac{1}{2}(a(2\alpha) + a(0)) = a\big(\frac{1}{2}(2\alpha + 0)\big) \\ &a(2\alpha) = 2a(\alpha) \end{aligned} \tag{52}$$

Next, substitute α with 2α and β with 2β in (51) to obtain

$$\begin{aligned} &\frac{1}{2}(a(2\alpha) + a(2\beta)) = a\big(\frac{1}{2}(2\alpha + 2\beta)\big) \\ &a(\alpha + \beta) = a(\alpha) + a(\beta) \end{aligned} \tag{53}$$

We use (53) 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.

$a(q\frac{1}{q}\alpha) = qa(\frac{1}{q}\alpha) \implies a(\frac{1}{q}\alpha) = \frac{1}{q}a(\alpha)$, 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 the fact 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

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

Observe that $a : \mathbb{R}^m \to \mathbb{R}^m$ and for any $\alpha, \beta \in \mathbb{R}^m$ $a(\alpha + \beta) = a(\alpha) + a(\beta)$ and $a(c\alpha) = ca(\alpha)$. From the definition of a linear map it follows that a is linear. As a result, we can write $\forall x \in [0, 1]^n, y \in [0, 1]^n$

$$\psi(x, y) = A(\phi(x, y)) \tag{54}$$

Observe that a is invertible because both ρ and ω are invertible. As a result, we know that A is an invertible matrix. From this we get

$$\phi(x, y) = A^{-1}\psi(x, y) = C(\psi(x, y)) \tag{55}$$

For all $z \in \mathbb{R}^m$, we obtain

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

Let us consider any sequence $x_{\leq \tilde{T}} \in [0, 1]^{n\tilde{T}}$. We use the above conditions

$$\omega\big(\sum_{j<\tilde{T}} \psi(x_{\tilde{T}}, x_j)\big) = \rho(C \sum_{j<\tilde{T}} \psi(x_{\tilde{T}}, x_j)) = \rho\big(\sum_{j<\tilde{T}} \phi(x_{\tilde{T}}, x_j)\big).$$

Thus we obtain length and compositional generalization.

□

Linear identification Observe that we arrive at equation (54) by starting from the condition that the model generalizes on sequences of length up to T. The condition directly implies that ψ linearly identifies ϕ.

On absence of labels at all lengths from 1 to T We argue that the above proof can be adapted to the setting where we do not observe labels at all lengths from 1 to T. Suppose we only observe label at length T. Take equation (48) and substitute $x_i = x$ and $x_j = y$ for all $j < i$ to obtain the same condition as equation (49). Suppose T is odd and larger than or equal to 3. Fix $x_i = x$, $x_{2j-1} = y, \forall j \in \{1, \cdots, (T-1)/2\}$, $x_{2j} = z, \forall j \in \{1, \cdots, (T-1)/2\}$. We obtain the same condition as equation (50). Rest of the proof can be adapted using a similar line of reasoning.

Remark on Assumption A.8 We require that the support of $[\phi(X_1, X_2), \phi(X_1, X_3)]$ is $\mathbb{R}^{2m}$. This assumption is used in the proof in equation (53). We used this assumption to arrive at $a(\alpha + \beta) = a(\alpha) + a(\beta), \forall \alpha, \beta \in \mathbb{R}^m$. We then used continuity of a to conclude a is linear. Now suppose $[\phi(X_1, X_2), \phi(X_1, X_3)]$ is some subset $\mathcal{Z} \subseteq \mathbb{R}^{2m}$. We believe that it is possible to extend the result to more general $\mathcal{Z}$, it might still be possible to arrive at a is linear. We leave this investigation to future work.

Multiple attention heads 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. The model class with two attention heads ψ_1, ψ_2 can be stated as follows $\omega\Big(\sum_{j<i} A[\psi_1(x_i, x_j), \psi_2(x_i, x_j)]^\top\Big)$, where A combines the outputs of the attention heads linearly. Following the same steps of proof of Theorem A.9, we obtain the following.

$$\begin{aligned}
&\omega\Big(\sum_{j<i} A[\psi_1(x_i, x_j), \psi_2(x_i, x_j)]^\top\Big) = \rho\Big(\sum_{j<i} B[\phi_1(x_i, x_j), \phi_2(x_i, x_j)]^\top\Big), \\
&\omega\Big(\sum_{j<i} \tilde{\psi}(x_i, x_j)\Big) = \rho\Big(\sum_{j<i} \tilde{\phi}(x_i, x_j)\Big), \\
&\sum_{j<i} \tilde{\psi}(x_i, x_j = a\Big(\sum_{j<i} \tilde{\phi}(x_i, x_j)\Big),
\end{aligned} \tag{56}$$

where $a = \omega^{-1} \circ \rho$. In the above simplification, the RHS shows the labeling function and the RHS is the function that is learned. We can follow the same strategy as the proof of Theorem A.9 for the rest of the proof. We set $i = 2$ and obtain a condition similar to (49) and for $i = 3$ we obtain a condition similar to (50). Following a similar proof technique, we obtain a is linear and the proof extends to multiple attention heads.

Positional encoding We next present the result when ω is continuously differentiable and invertible.

Assumption A.10. Each function in the hypothesis class $\mathcal{H}$ used by the learner is given as $h(x_1, \cdots, x_i) = \omega\Big(\sum_{j\leq i} \psi_{i-j}(x_i, x_j)\Big)$, where ω is a C^1-diffeomorphism. Also, $\psi_{i-j} = 0$ for $i - j > T_{\mathsf{max}} - 1$, i.e., two tokens that are sufficiently far apart do not interact. For all $k \leq T_{\mathsf{max}} - 1$ each $x \in [0,1]^n$, $\exists\ y \in [0,1]^n$ we $\psi_k(x, y) = 0$.

In the theorem that follows, we require the support of training distribution under consideration is already sufficiently diverse and hence we only seek to prove length generalization guarantees.

Assumption A.11. The joint support $\mathsf{supp}(X_{\leq T}) = [0,1]^T$. The support of $[\phi_1(X_1, X_2), \phi_2(X_1, X_3)]$ is $\mathbb{R}^{2k}$, where ϕ_{i-j} is the embedding function for the labeling function $\rho(\sum_{j\leq i} \phi_{i-j}(x_i, x_j))$.

Theorem A.12. *If $\mathcal{H}$ follows Assumption A.10, the realizability condition holds, i.e., $f \in \mathcal{H}$, Assumption A.11 holds and $T \geq T_{\mathsf{max}}$, then the model trained to minimize the risk in (1) (with $T \geq 2$) with ℓ_2 loss achieves length 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 A.1) to obtain the following. For all $x_{\leq i} \in \mathsf{supp}(X_{\leq i})$

$$\begin{aligned}
&\omega\Big(\sum_{j<i} \frac{1}{i-1}\psi_{i-j}(x_i, x_j)\Big) = \rho\Big(\sum_{j<i} \frac{1}{i-1}\phi_{i-j}(x_i, x_j)\Big), \\
&\sum_{j<i} \frac{1}{i-1}\psi_{i-j}(x_i, x_j) = \omega^{-1} \circ \rho\Big(\sum_{j<i} \frac{1}{i-1}\phi_{i-j}(x_i, x_j)\Big), \\
&\sum_{j<i} \frac{1}{i-1}\psi_{i-j}(x_i, x_j) = a\Big(\sum_{j<i} \frac{1}{i-1}\phi_{i-j}(x_i, x_j)\Big),
\end{aligned} \tag{57}$$

where $a = \omega^{-1} \circ \rho$. In the above simplification, we used the parametric form for the true labeling function and the learned labeling function. We also used 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 $\forall x \in [0,1]^n, y \in [0,1]^n$,

$$\psi_1(x, y) = a(\phi_1(x, y)). \tag{58}$$

For $i = 3$, substitute $x_1 = x$, $x_3 = z$ and set $x_2 = y$ in such a way that $\phi_1(x, y) = 0$ (follows from Assumption A.10). Thus we obtain

$$\psi_2(x, y) = a(\phi_2(x, y)). \tag{59}$$

Similarly, we can obtain the following. For all $k \leq T_{\mathsf{max}}$

$$\psi_k(x, y) = a(\phi_k(x, y)). \tag{60}$$

The above expression implies that ψ bijectively identifies ϕ. Let us consider the setting when $i = 3$ (this is possible since $T \geq 3$). We substitute $x_3 = x$, $x_2 = y$, $x_1 = z$ to give

$$\frac{1}{2}\big(a(\phi_1(x, y)) + a(\phi_2(x, z))\big) = a\Big(\frac{1}{2}(\phi_1(x, y) + \phi_2(x, z))\Big). \tag{61}$$

We now use the that assumption $[\phi_1(x, y), \phi_2(x, z)]$ spans $\mathbb{R}^{2k}$ and substitute $\phi_1(x, y) = \alpha$ and $\phi_2(x, z) = \beta$

$$\frac{1}{2}(a(\alpha) + a(\beta)) = a\Big(\frac{1}{2}(\alpha + \beta)\Big). \tag{62}$$

Rest of the proof follows the same strategy as proof of Theorem A.9. □

A.2.3. State space models

In this section, we discuss SSMs and provide the proofs of Theorem 2.9.

Assumption A.13. The joint support $\mathsf{supp}(X_{\leq i})$ is a regular closed set for all $i \leq T$. The support of X_1 is $\mathbb{R}^n$. 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 2.9. *If $\mathcal{H}$ follows Assumption 2.8, and the realizability condition holds, i.e., $f \in \mathcal{H}$, and a further condition on the support, i.e., Assumption A.13, holds, then the model trained to minimize the risk in (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 $h(x_{\leq i}) = f(x_{\leq i}), \forall i \leq T$ 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 A.1) to obtain the following. For all $x_{\leq i} \in \mathsf{supp}(X_{\leq i})$.

$$\begin{aligned}
&f(x_{\leq i}) = h(x_{\leq i}) = \\
&\rho(\sum_{j=0}^{i-1} \Lambda^j B x_{i-j}) = \omega(\sum_{j=0}^{i-1} \tilde{\Lambda}^j \tilde{B} x_{i-j}) \implies \\
&\omega^{-1} \circ \rho(\sum_{j=0}^{i-1} \Lambda^j B x_{i-j}) = \sum_{j=0}^{i-1} \tilde{\Lambda}^j \tilde{B} x_{i-j} = \\
&c(\sum_{j=0}^{i-1} \Lambda^j B x_{i-j}) = \sum_{j=0}^{i-1} \tilde{\Lambda}^j \tilde{B} x_{i-j}
\end{aligned} \tag{63}$$

For $i = 1, \forall x_1 \in \mathbb{R}^n, c(Bx_1) = \tilde{B}x_1$. Substitute $Bx_1 = x$, we obtain $\forall x \in \mathbb{R}^n, c(x) = \tilde{B}B^{-1}x = Cx$, where we use the fact that Bx_1 spans $\mathbb{R}^n$ as B is invertible.

From linearity of c, we obtain

$$\omega^{-1} \circ \rho(z) = Cz \implies \rho(z) = \omega(Cz), \forall z \in \mathbb{R}^n \tag{64}$$

We use this linearity of c to simplify

$$\begin{aligned}
&c(\sum_{j=0}^{i-1} \Lambda^j B x_{i-j}) = \sum_{j=0}^{i-1} \tilde{\Lambda}^j \tilde{B} x_{i-j} \implies \\
&C(\sum_{j=0}^{i-1} \Lambda^j B x_{i-j}) = \sum_{j=0}^{i-1} \tilde{\Lambda}^j \tilde{B} x_{i-j} \implies \\
&[CB, C\Lambda B, C\Lambda^2 B, \cdots, C\Lambda^{i-1}B] \begin{bmatrix} x_i \\ x_{i-2} \\ \vdots \\ x_1 \end{bmatrix} - [\tilde{B}, \tilde{\Lambda}\tilde{B}, \tilde{\Lambda}^2\tilde{B}, \cdots, \tilde{\Lambda}^{i-1}\tilde{B}] \begin{bmatrix} x_i \\ x_{i-2} \\ \vdots \\ x_1 \end{bmatrix} = 0 \implies \\
&\Big[[CB, C\Lambda B, C\Lambda^2 B, \cdots, C\Lambda^{i-1}B] - [\tilde{B}, \tilde{\Lambda}\tilde{B}, \tilde{\Lambda}^2\tilde{B}, \cdots, \tilde{\Lambda}^{i-1}\tilde{B}]\Big] \boldsymbol{X} = 0,
\end{aligned} \tag{65}$$

where $\boldsymbol{X} = \begin{bmatrix} x_i \\ x_{i-2} \\ \vdots \\ x_1 \end{bmatrix}$.

Denote $R = \Big[[CB, C\Lambda B, C\Lambda^2 B, \cdots, C\Lambda^{i-1}B] - [\tilde{B}, \tilde{\Lambda}\tilde{B}, \tilde{\Lambda}^2\tilde{B}, \cdots, \tilde{\Lambda}^{i-1}\tilde{B}]\Big]$. 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 A.13). Since the matrix $\boldsymbol{X}^+$ is full rank, we have

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

This yields

$$CB = \tilde{B}, C\Lambda B = \tilde{\Lambda}\tilde{B}, \cdots, C\Lambda^i B = \tilde{\Lambda}^i\tilde{B}. \tag{66}$$

Observe that from the second equality, we get $\tilde{\Lambda} = C\Lambda C^{-1}$. Given the parameters (Λ, B), the set of parameters $(\tilde{\Lambda}, \tilde{B})$ that solve the first two equalities are – $\{\tilde{B}$ is an arbitrary invertible matrix, $\tilde{\Lambda} = C\Lambda C^{-1}$, where $C = \tilde{B}B^{-1}\}$.

Take any solution of the first two equalities and compute

$$\tilde{\Lambda}^i\tilde{B} = C\Lambda^i C^{-1}\tilde{B} = C\Lambda^i B, \forall i \geq 1 \tag{67}$$

From (67) and (64), we obtain that for all $x_{\leq i} \in \mathbb{R}^{ni}$

$$h(x_{\leq i}) = \omega(\sum_{j=0}^{i-1}\tilde{\Lambda}^j\tilde{B}x_{i-j}) = \omega(C\sum_{j=0}^{i-1}\Lambda^j Bx_{i-j}) = \rho(\sum_{j=0}^{i-1}\Lambda^j Bx_{i-j}) = f(x_{\leq i}) \tag{68}$$

This establishes both compositional and length generalization.

□

Linear identification Observe that equation (63) is arrived at under the condition that the model generalizes at all lengths up to T. From (63) and linearity of $c(\cdot)$ it follows that $\sum_{j=0}^{i-1}\tilde{\Lambda}^j\tilde{B}x_{i-j} = C(\sum_{j=0}^{i-1}\Lambda^j Bx_{i-j})$. Recall that $\sum_{j=0}^{i-1}\tilde{\Lambda}^j\tilde{B}x_{i-j} = \tilde{h}_j$ and $\sum_{j=0}^{i-1}\Lambda^j Bx_{i-j} = h_j$. From this it follows that $\tilde{h}_j = Ch_j$, which proves that learned hidden state are a linear transform of the hidden state underlying the labeling function.

A.2.4. Vanilla RNNs

In this section, we discuss RNNs and present the proof of Theorem 2.11.

Lemma A.14. *The k^{th} derivative of sigmoid function denoted $\frac{\partial^k\sigma(s)}{\partial s^k}$ is not zero identically.*

Proof. The first derivative of the sigmoid function $\frac{\partial\sigma(s)}{\partial s} = \sigma(s)(1-\sigma(s))$. We argue that the $\frac{\partial^k\sigma(s)}{\partial s^k}$ is a polynomial in $\sigma(s)$ with degree $k+1$. Consider the base case of $k=1$. This condition is true as $\frac{\partial\sigma(s)}{\partial s} = \sigma(s)(1-\sigma(s))$. Now let us assume that $\frac{\partial^k\sigma(s)}{\partial s^k}$ is a polynomial of degree at most $k+1$ denoted as $P_{k+1}(\sigma(s))$. We simplify

$$\frac{\partial^k\sigma(s)}{\partial s^k} = P_{k+1}(\sigma(s)) = \sum_{j=1}^{k+1} a_j(\sigma(s))^j$$

We take another derivative of the term above as follows.

$$\frac{\partial^{k+1}\sigma(s)}{\partial s^{k+1}} = \frac{\partial P_{k+1}(\sigma(s))}{\partial s} = \sum_{j=1}^{k+1} a_j\frac{\partial(\sigma(s))^j}{\partial s} = \sum_{j=1}^{k+1} a_j j\sigma(s)^{j-1}(\sigma(s)(1-\sigma(s)))$$

Observe that the $\frac{\partial^{k+1}\sigma(s)}{\partial s^{k+1}}$ is also a polynomial in $\sigma(s)$. Observe that the degree $k+2$ term has one term with coefficient $-a_{k+1}\cdot(k+1)$. Since $a_{k+1} \neq 0$, the coefficient of degree $k+2$, $-a_{k+1}\cdot(k+1)$, is also non-zero. Since $\frac{\partial^k\sigma(s)}{\partial s^k}$ is a polynomial in $\sigma(s)$ with degree $k+1$ and hence, it cannot be zero identically.

□

Lemma A.15. *Let $x \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$. Suppose $Ax = 0, \forall x \in \mathcal{X}$, where $\mathcal{X}$ has a non-empty interior. Under these conditions $A = 0$.*

Proof. Since $\mathcal{X}$ has a non-empty interior, we can construct a ℓ_∞ ball centered on θ, defined as follows – $\tilde{\mathcal{X}} = \{\theta + \sum_{j=1}^{n} \alpha_j e_j \;|\|\alpha\|_\infty \leq \alpha_{\mathsf{max}}\ \}$, where e_j is a vector that is zero in all components and one on the j^{th} component. Suppose A was non-zero. One of the columns say a_j is non-zero. Consider two points in the ball $\tilde{\mathcal{X}}$ such that j^{th} coefficients are non-zero but rest of the coefficients are zero. We denote the j^{th} components for the two components as α_j and $\tilde{\alpha}_j$, where $\alpha_j \neq \tilde{\alpha}_j$. We now plug these two points into the condition that $Ax = 0$

$$\begin{aligned} A(\theta + \alpha_j e_j) = 0 \implies A\theta = \alpha_j a_j, \\ A(\theta + \tilde{\alpha}_j e_j) = 0 \implies A\theta = \tilde{\alpha}_j a_j, \end{aligned} \tag{69}$$

We take a difference of the two steps above and obtain

$$(\alpha_j - \tilde{\alpha_j})a_j = 0 \implies a_j = 0$$

This is a contradiction. Hence, $A = 0$. □

Theorem 2.11. *If $\mathcal{H}$ follows Assumption 2.10, and the realizability condition holds, i.e., $f \in \mathcal{H}$ and regular closedness condition in Assumption 2.4 holds, then the model trained to minimize the risk in (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 A.1). We start with equating label at length 1, i.e., y_1. For all $x_1 \in \mathsf{supp}(X_1)$

$$\begin{aligned} &\sigma(A\sigma(Bx_1)) = \sigma(\tilde{A}\sigma(\tilde{B}x_1)) \implies A\sigma(Bx_1) = \tilde{A}\sigma(\tilde{B}x_1) \implies \\ &\sigma(B\tilde{B}^{-1}\tilde{B}x_1) = A^{-1}\tilde{A}\sigma(\tilde{B}x_1) \end{aligned} \tag{70}$$

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

$$\sigma(Vy) = U\sigma(y) \tag{71}$$

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

$$\sigma(v^\top y) = u^\top \sigma(y) \tag{72}$$

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

$$\frac{\partial \sigma(v_i y_i + v_{-i} y_{-i})}{\partial y_i} = \sigma^{'}(v_i y_i + v_{-i} y_{-i})v_i = \frac{\partial u_{-i}^\top \sigma(y_{-i})}{\partial y_i} = 0 \tag{73}$$

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.

$$\frac{\partial \sigma(v_{-i}^\top y_{-i})}{\partial y_i} = 0 = \frac{\partial (u_i \sigma(y_i) + u_{-i}^\top \sigma(y_{-i}))}{\partial y_i} = u_i \sigma^{'}(y_i) \tag{74}$$

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).

$$\begin{aligned} \sigma(v^\top y) &= u^\top \sigma(y), \\ \frac{\partial^p \sigma(s)}{\partial s^p}\Pi_{u_i\neq 0}u_i = 0 &\implies \frac{\partial^p \sigma(s)}{\partial s^p} = 0. \end{aligned} \tag{75}$$

Since support X_1 has a non-empty interior, the set of values $v^\top y$ takes also has a non-empty interior in $\mathbb{R}$. Hence, the above equality is true over a set of values s, which have a non-empty interior. Since $\sigma(s)$ is analytic, $\frac{\partial^p \sigma(s)}{\partial s^p}$ is also analytic. From (Mityagin, 2015), it follows that $\frac{\partial^p \sigma(s)}{\partial s^p} = 0$ everywhere. From Lemma A.14, we know this condition 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 (label y_2 exists as $T \geq 2$). For all $x_1 \in \mathsf{supp}(X_1)$

$$\begin{aligned} &\sigma(A\sigma(\Lambda\sigma(Bx_1) + Bx_2)) = \sigma(\tilde{A}\sigma(\tilde{\Lambda}\sigma(\tilde{B}x_1) + \tilde{B}x_2)) \implies \\ &A\sigma(\Lambda\sigma(Bx_1) + Bx_2) = \tilde{A}\sigma(\tilde{\Lambda}\sigma(\tilde{B}x_1) + \tilde{B}x_2) \implies \\ &\tilde{A}\sigma(\tilde{\Lambda}\sigma(\tilde{B}x_1) + \tilde{B}x_2) = A\Pi\sigma(\tilde{\Lambda}\Pi^\top\sigma(Bx_1) + \Pi^\top Bx_2) = A\sigma(\Pi\tilde{\Lambda}\Pi^\top\sigma(Bx_1) + Bx_2). \end{aligned} \tag{76}$$

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

$$\begin{aligned} &A\sigma(\Pi\tilde{\Lambda}\Pi^\top\sigma(Bx_1) + Bx_2) = A\sigma(\Lambda\sigma(Bx_1) + Bx_2) \\ &\implies (\Pi\tilde{\Lambda}\Pi^\top - \Lambda)\sigma(Bx_1) = 0. \end{aligned} \tag{77}$$

Since $\sigma(Bx_1)$ spans a set that has a non-empty interior, we get that $\tilde{\Lambda} = \Pi^\top\Lambda\Pi$ (from Lemma A.15).

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 $k \geq 1$

$$h_k = \Pi\tilde{h}_k, \tag{78}$$

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 $h_0 = \tilde{h}_0 = 0$. In other words, we define T_k as a mapping that takes $x_{\leq k}$ as input and outputs h_k, i.e., $T(x_{\leq k}) = h_k$. Similarly, we write $\tilde{T}(x_{\leq k}) = \tilde{h}_k$. We want to show

$$T_k = \Pi\tilde{T}_k, \forall k \tag{79}$$

We show the above by principle of induction. Let us consider the base case below. For all $x_1 \in \mathbb{R}^n$

$$\tilde{A}\sigma(\tilde{B}x_1) = A\Pi\sigma(\Pi^\top Bx_1) = A\sigma(Bx_1) = Ah_1 \implies h_1 = \Pi\tilde{h}_1 \implies T_1(x_1) = \Pi\tilde{T}_1(x_1) \tag{80}$$

Suppose $\forall j \leq k, T_j = \Pi\tilde{T}_j$.

Having shown the base case and assumed the condition for $j \leq k$, we now consider the mapping $\tilde{T}_{k+1}$

$$\Pi\tilde{T}_{k+1}(x_{\leq k+1}) = \Pi\sigma(\tilde{\Lambda}\tilde{h}_k + \tilde{B}x_{k+1}) = \Pi\sigma(\Pi^\top\Lambda\Pi\tilde{h}_k + \Pi^\top Bx_k) = \sigma(\Lambda h_k + Bx_k) = T_{k+1}(x_{\leq k+1}). \tag{81}$$

The prediction from the model $(\tilde{A}, \tilde{\Lambda}, \tilde{B})$ at a time step k is denoted as $\tilde{y}_k$ and it relates to $\tilde{h}_k$ as follows $\tilde{y}_k = \sigma(\tilde{A}\tilde{h}_k)$. We use the above condition in equation (79) to arrive at the following result. For all $x_{\leq k} \in \mathbb{R}^{nk}$

$\tilde{y}_k = \sigma(\tilde{A}\tilde{h}_k) = \sigma(\tilde{A}\tilde{T}(x_{\leq k})) = \sigma(A\Pi\tilde{T}(x_{\leq k})) = \sigma(AT(x_{\leq k})) = y_k$

This completes the proof. □

Permutation identification In the previous proof we start with condition that the model generalizes at all lengths up to T and arrive at equation (78). This condition directly implies that $\tilde{h}_j$ linearly identifies h_j. Since the two are related to each other through a permutation matrix, we state that $\tilde{h}_j$ satisfies permutation identification w.r.t h_j (following the notion of permutation identification (Khemakhem et al., 2020)).

A.3. Experiments

Here we present the empirical evaluation of compositional and length generalization capabilities of the architectures studied in Sections 2.1- 2.4. All the experiments are carried out in the realizable case where $f \in \mathcal{H}$. More specifically, depending on the architecture in question, we use a random instance of the architecture to generate the labels. We train a model h from the same architecture class to minimize the ℓ_2 loss between h and f. Under different scenarios, we ask if h achieves length generalization and compositional generalization. We also seek to understand the relationship between the hidden representations of h and hidden representations of f.

Length generalization We sample sequences $x_{\leq t}$ of varying length with a maximum length of $T = 10$. Each token $x_i \sim \text{Uniform}[0,1]^n$, where $n = 20$. The sequences are then fed to the labeling f, which comes from the hypothesis class of the architecture, to generate the labels. We minimize the empirical risk version of (1) over the same hypothesis class with ℓ_2 loss. For evaluation, we present the ℓ_2 loss on the test datasets. We also evaluate R^2 of linear regression between the learned hidden representations denoted $\psi(x_i)$ and true hidden representations $\phi(x_i)$ for all $x_i \in x_{\leq t}$ from the test dataset sequences. This metric is often used to evaluate the claims of linear identification (Khemakhem et al., 2020), i.e., the higher this value, the closer the linear relationship.

In Figure 1 we present results averaged over five seeds for models with one hidden layer MLP for ρ (ϕ is one hidden layer MLP for deep sets). Figure 1 shows a very small test loss of models on increasing sequence lengths when only trained with sequences of up to length $T = 10$, which is in agreement with Theorem 2.5-2.11. Further, in Figures 2, 3, 4, 5 we show an exemplar sequence from test set and how the trained model from each architecture tracks it. Table 1 shows the average of R^2 score of $\psi(x_i), \phi(x_i)$ across different positions i at test time. These results demonstrate a linear relationship between learned and true hidden representations, which agrees with our claims of linear identification. In Section A.3, we show that when realizability condition does not hold, i.e., $f \notin \mathcal{H}$, then length generalization is not achieved.

Additionally, to support the theory on other types of attention, Figures 6, 7 demonstrate the loss and prediction of a Transformer with ReLU attention and one hidden layer MLPs for ω, ψ trained on output sequences of a Transformer with ReLU attention and one hidden layer MLP for ρ, ϕ. Similarly, all these models were trained to predict sequences of length up to $T = 10$ output by a true labeling function f in their respective hypothesis classes $\mathcal{H}$, and were tested with sequences of length up to 100. As a reminder, the output tokens $y_i \in \mathbb{R}^m$, where $m = 20$, and the figures below show only one representative dimension for illustration. All models demonstrate strong length generalization capacity.

In Figure 8, we present additional findings for length generalization capability of all architectures when both the learner and the generating process MLPs all consist of two hidden layers with input, output, and hidden size matching $n = m = k = 20$.

Figures 9, 10, 11 present the prediction behaviour of Transformer with softmax attention, SSM, and RNN architectures with two hidden layers in ρ (and two hidden layer MLPs for the learner ω). Training procedure remains the same. We can observe that all models length generalize.

Model	R^2 ($t = 20$)	R^2 ($t = 100$)
Deep set	0.97 ± 0.01	0.97 ± 0.01
Transformer	0.99 ± 0.01	0.99 ± 0.01
SSM	0.99 ± 0.01	0.99 ± 0.01
RNN	0.99 ± 0.01	0.99 ± 0.01

Table 1. Average test R^2 of true and learned hidden representations $\psi(x_i), \phi(x_i)$ across all positions i at various lengths unseen during training. A strong linear relationship is observed for all models across lengths.

Model	Test Loss $\times 10^6$	R^2
Deep set	1.27 ± 0.24	0.96 ± 0.01
Transformer	4.50 ± 3.28	1.00 ± 0.00
SSM	11.00 ± 10.92	1.00 ± 0.00
RNN	1.22 ± 0.12	0.99 ± 0.00

Table 2. Compositional generalization: Test ℓ_2 loss and R^2 score for models on sequences of length $T = 10$. A strong linear relationship is observed for all models for new sequences made of unseen token combinations.

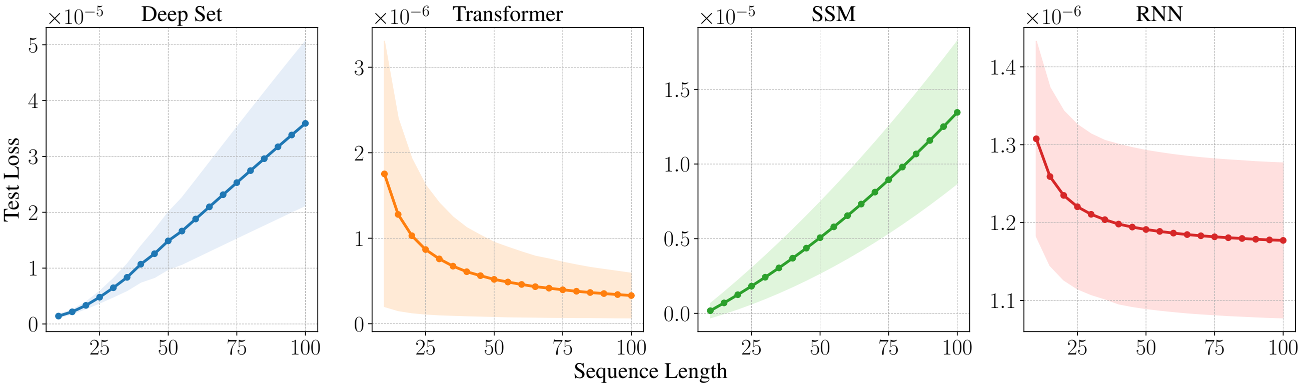

Figure 1. Length generalization: Test ℓ_2 loss on sequences of different lengths. The models are trained only on sequences of length up to $T = 10$. All models achieve small error values $\approx 10^{-5} - 10^{-6}$ at all sequence lengths and thus length generalize. Since the error values are already quite small, the increasing or decreasing trends are not numerically significant.

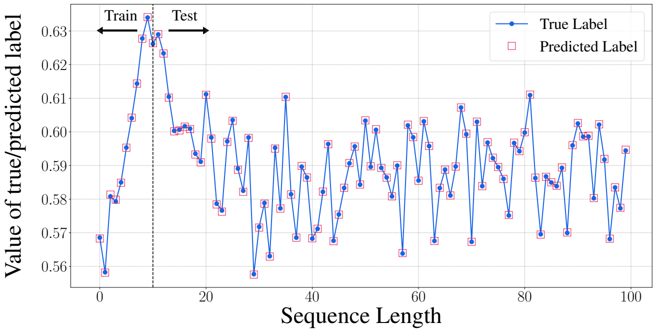

Figure 2. A transformer model with softmax attention and one hidden layer MLP trained on sequences of length up to $T = 10$ shows perfect generalization to sequences of length up to 100.

Compositional generalization During training, we sample each component k of a token from Uniform$[0, 1]$ and accept the sampled sequences that satisfy the following for all components i: $-0.5 \leq \sum_{j=1}^{T}(x_j^k - 0.5) \leq 0.5 \ \ \forall k$, where x_j^k is the k^{th} component of token j. During testing, we sample $x_{\leq t}$ from the complementary set of the training set, i.e., corners of hypercube $[0, 1]^{nt}$. We present the ℓ_2 loss on the test dataset, as well as the mean R^2, where the results are averaged over 5 seeds. The rest of the details are the same as the previous section, i.e., $T = 10$, $n = 20$, ρ is one hidden layer MLP (ϕ is one hidden layer MLP for deep sets). Table 2 shows the test ℓ_2 loss of models and the R^2 scores for linear identification.

Additionally we present the prediction behavior of different architectures on the test sequences that consist of unseen token combinations during training. This helps us better interpret qualitatively how the model actually performs in following the true labels. Figures 13- 16 show the prediction trajectories for different architectures. We can observe that not only do these models perform quite well on unseen sequences of length up to $T = 10$, but they also length generalize and continue to remain consistent with the true labels on unseen combinations at longer lengths than the training.

Figures 17, 18, 19, 20 present the prediction behaviour of deep set, transformer with softmax attention, SSM, and RNN architectures with *two* hidden layers in ρ (and two hidden layer MLPs for the learner ω) when trained on sequences of length up to $T = 10$ sampled from the red region in Figure 12. We can observe that all models continue to generalize to unseen combinations beyond their training length. Table 3 presents the test loss and R^2 on the test set when the model is only trained on the red region in Figure 12. All models generalize to unseen combination of tokens and the learned representations linearly identify the true hidden representations

Failure Cases Although most of our focus has been on the success scenarios for length and compositional generalization, here we provide an example to show how a model might fail. In Figure 21, we present the prediction of a deep set model

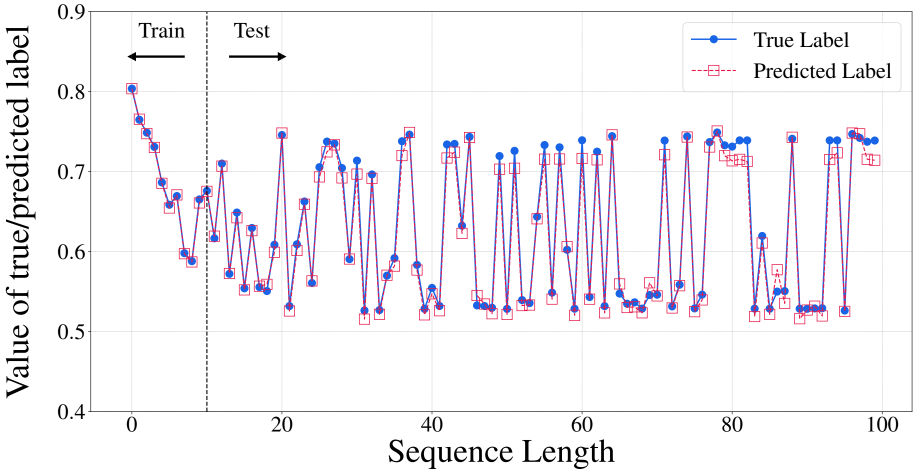

Figure 3. A deep set model with one hidden layer MLP for ψ, ω trained on sequences of length up to $T = 10$ shows perfect generalization to sequences of length up to 100.

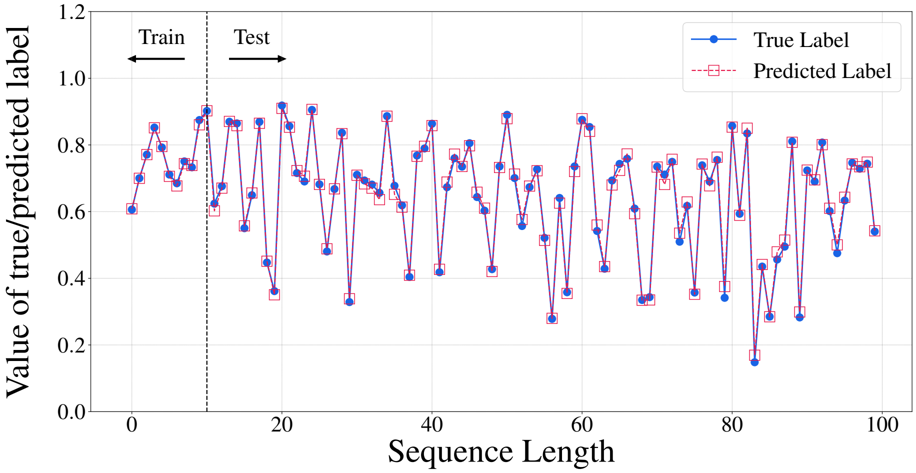

Figure 4. A SSM model with one hidden layer MLP for ω trained on sequences of length up to $T = 10$ length generalizes to sequences of length up to 100.

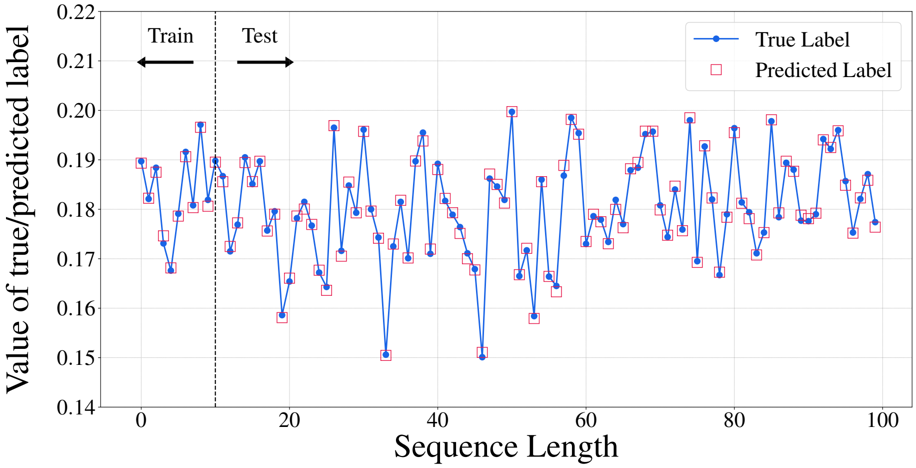

Figure 5. A RNN model with one hidden layer MLP for ω trained on sequences of length up to $T = 10$ length generalizes to sequences of length up to 100.

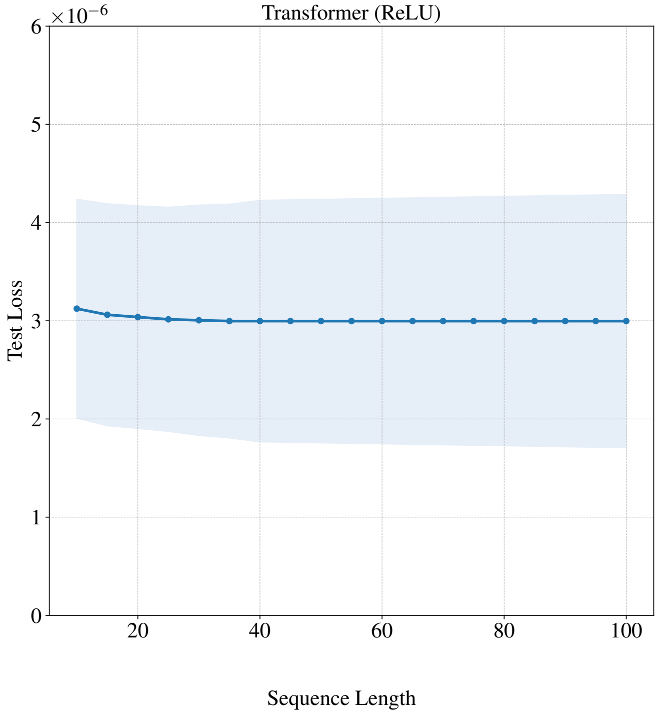

Figure 6. Tess loss of a transformer model with ReLU attention and one hidden layer MLP for ω, ψ trained on sequences of length up to $T = 10$ length generalizes to sequences of length up to 100. The results are averaged over five seeds.

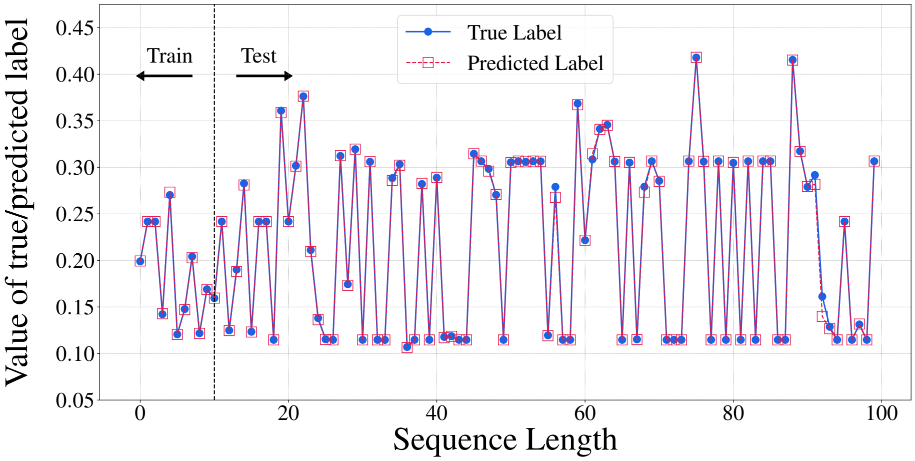

Figure 7. A transformer model with ReLU attention and one hidden layer MLP for ω, ψ trained on sequences of length up to $T = 10$ length generalizes to sequences of length up to 100.

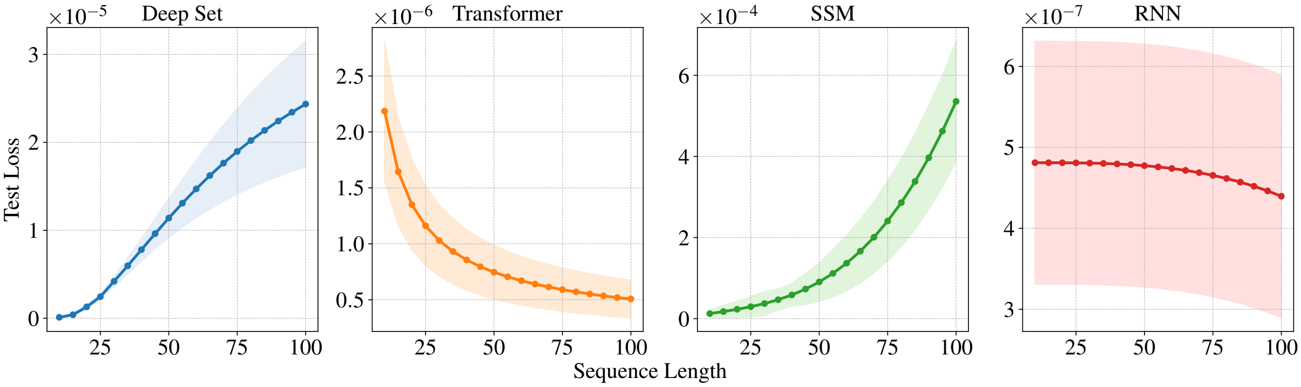

Figure 8. Length generalization: Test ℓ_2 loss on sequences of different lengths. The models are trained only on sequences of length up to $T = 10$. All models achieve small error values $\approx 10^{-4} - 10^{-7}$ at all sequence lengths and thus length generalize. Since the error values are already quite small, the increasing or decreasing trends are not numerically significant.

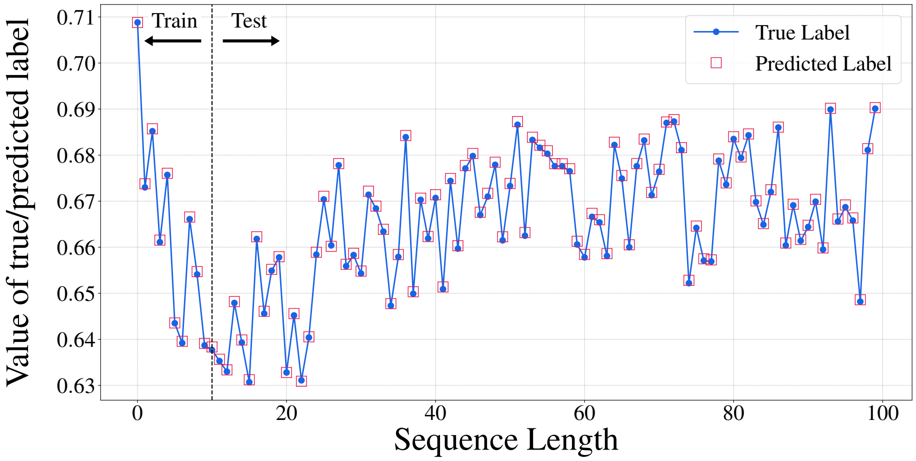

Figure 9. A transformer model with softmax attention with *two* hidden layer MLP for ω, ψ trained on sequences of length up to $T = 10$ length generalizes to sequences of length up to 100.

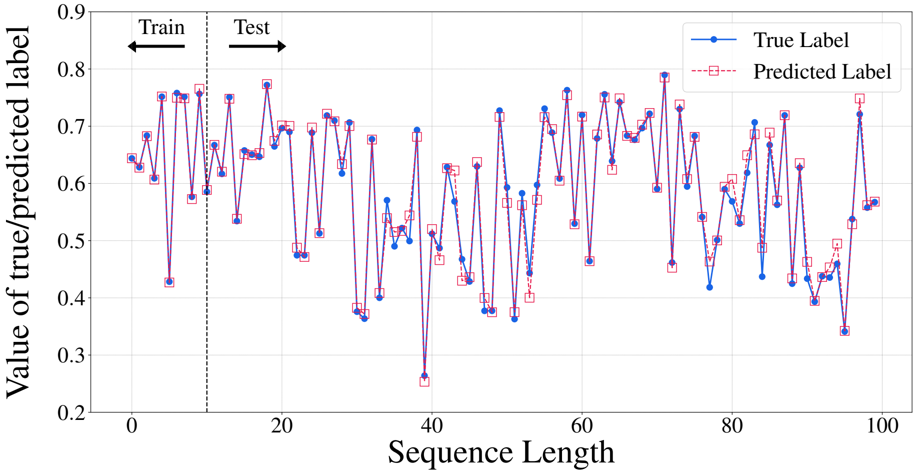

Figure 10. A SSM model with *two* hidden layer MLP for ω trained on sequences of length up to $T = 10$ length generalizes to sequences of length up to 100.

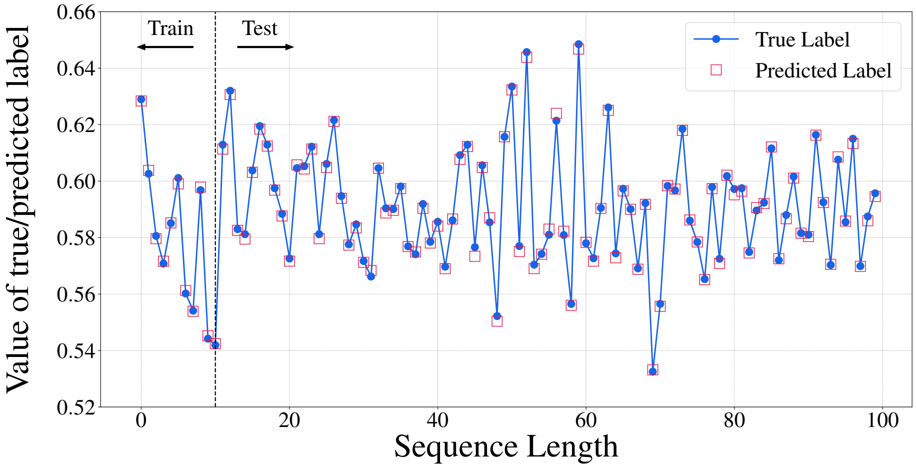

Figure 11. A RNN model with *two* hidden layer MLP for ω trained on sequences of length up to $T = 10$ length generalizes to sequences of length up to 100.

that is sampled from a hypothesis class $\mathcal{H}$ that does not contain the true data generating function. In particular, the true labeling function f is a deep set with one hidden layer MLPs for ρ, ϕ, but the learner uses h, a deep set model for which the MLPs ψ, ω have no hidden layers. We can observe that the model can predict the test sequence well up to the length it has learned during training, but starts to diverge from the true labels beyond that. This demonstrates a failure case in which the realizability condition is violated.

Next we provide additional experimental results as well as the training details.

Model Architecture In all the architectures, there are two types of non-linearities, ω that generates the target label, ψ that operates on inputs (used in deep sets and transformers). We use MLPs to implement these non-linearities. We instantiate MLPs with k hidden layers, and the input, output, and hidden dimensions are all the same $m = n = k$. Recall that under the realizability assumption $f \in \mathcal{H}$. Therefore, we need to select the labeling function from $\mathcal{H}$. To do so, the weights of MLP are initialized according to $\mathcal{N}(\mu, \sigma^2)$, where $\mu = 0.0, \sigma = 0.6$. For RNNs and SSMs, A, B, Λ are initialized separately for the learner and true generating process as orthogonal matrices. All hidden layers, as well as the output layer are followed by a sigmoidal activation function.

Training Details and Hyperparameter Selection We train all models with AdamW optimizer (Loshchilov & Hutter, 2019) with a learning rate of 10^{-3}, weight decay of 0.01, $\epsilon = 10^{-8}, \beta_1 = 0.9, \beta_2 = 0.95$. We reduce the learning rate by a factor of 0.8 if the validation loss is not improved more than 10^{-6} for 1 epochs. This drop is followed by a cool-down period of 1 epochs, and the learning rate cannot decrease to lower than 10^{-7}. For all datasets we use a streaming dataset where each epoch contains 100 batches of size 256 sampled online from the specified training and test distributions, and we train all models for 100 epochs. Therefore, the size of the training dataset is 256×10^4 and the size of the testing dataset is 256×10^2. Since our models are generally small, running the experiments is rather inexpensive, and we carried out each experiment on 4 CPU cores using 20 GB of RAM. For inference, specially for SSM and RNN with very long sequences, we use RTX8000 GPUs.

Practical Considerations For training and evaluating compositional aspect of generalization, we follow the sampling procedure described in A.3 with a slight modification that allows for faster sampling and easier training. This procedure is illustrated in Figure 12, and results in a more difficult testing strategy, as the test set spans a smaller area than the complement of the training set.

We opted for such a procedure because rejection sampling from the complement of the training set given A.3 is extremely slow. In particular, given our batch size of 256, token dimension $n = 20$, and having 100 batches per epoch, constructing the full test set requires finding $256 \times 100 \times 20$ sequences of length $t \leq T$ that are rejected by the original constraints. This becomes quite inefficient and slow specially in higher dimensions as the sum of the sequence along each component tends to concentrate more around $t/2$, therefore it becomes harder to find such sequences (the sum follows Irwin-Hall

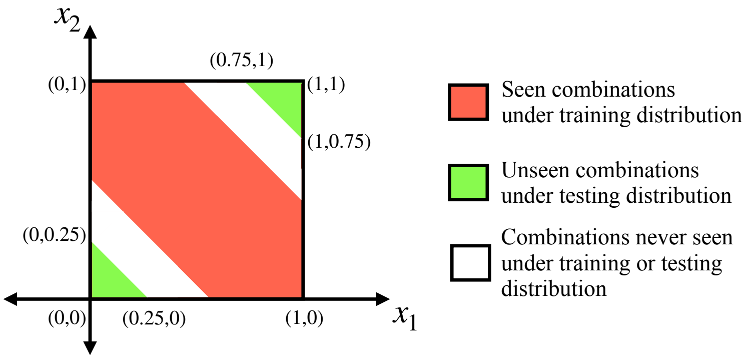
x_2
(0.75,1)
(0,1)
(1,1)
(1,0.75)
(0,0.25)
(0,0)
(0.25,0)
(1,0)
x_1
Seen combinations under training distribution
Unseen combinations under testing distribution
Combinations never seen under training or testing distribution

Figure 12. Illustrating the modified support of train vs test distribution for compositional generalization. This enables speed up in the sampling procedure, while keeping the challenging aspect of generalization to the corners.

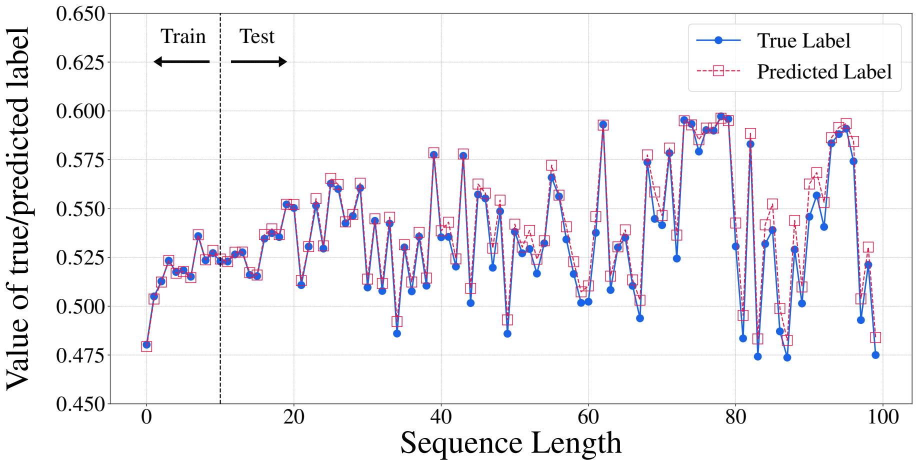

Figure 13. A deep set model with one hidden layer MLP for ω, ψ trained on sequences of length up to $T = 10$ sampled according to Figure 12 can generalize to unseen test sequences (Figure 12). Additionally, the compositional generalization holds even beyond the training length.

distribution since the components come from the Uniform distribution). To improve the speed of sampling the test dataset, we sample token dimensions x_i^k from the smaller corners shown in Figure 12 which allows for parallel sampling. These corners correspond to sampling $x_i^k \sim \text{Uniform}[0, 1/2T]$ or $x_i^k \sim \text{Uniform}[1/2 + 1/2T, 1]$. This way we can sample token components independently and in parallel without having to reject any samples, since by construction no test sequence coincides with the training set. This procedure leaves a gap (see Figure 12) that will not be sampled neither during training nor testing.

A.4. Broader Impacts

When machine learning models are deployed to assist in making decisions in safety-critical applications (e.g., self-driving cars, healthcare, etc.), we want to ensure that they make decisions that can be trusted well beyond the regime of the training data that they are exposed to. In this work, we took some steps towards building a well-founded theory that helps us establish guarantees well beyond the training data regime. At this point, we do not anticipate a negative impact specifically of this work.

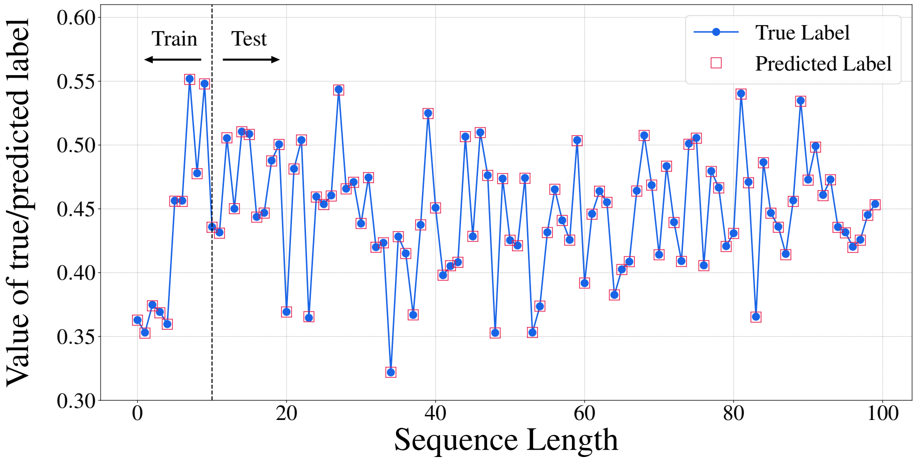

Figure 14. A Transformer model with softmax attention and one hidden layer MLP for ω trained on sequences of length up to $T = 10$ sampled according to Figure 12 can generalize to unseen test sequences (Figure 12). Additionally, the compositional generalization holds even beyond the training length.

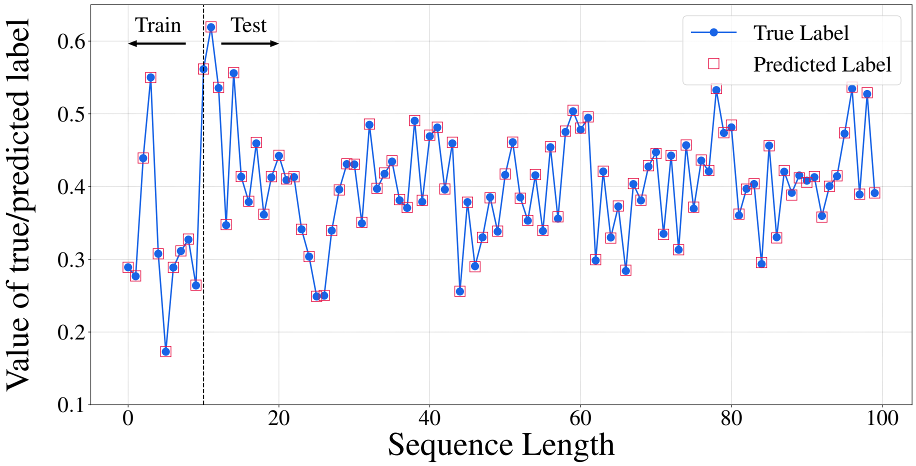

Figure 15. A SSM model with one hidden layer MLP for ω trained on sequences of length up to $T = 10$ sampled according to Figure 12 can generalize to unseen test sequences (Figure 12). Additionally, the compositional generalization holds even beyond the training length.

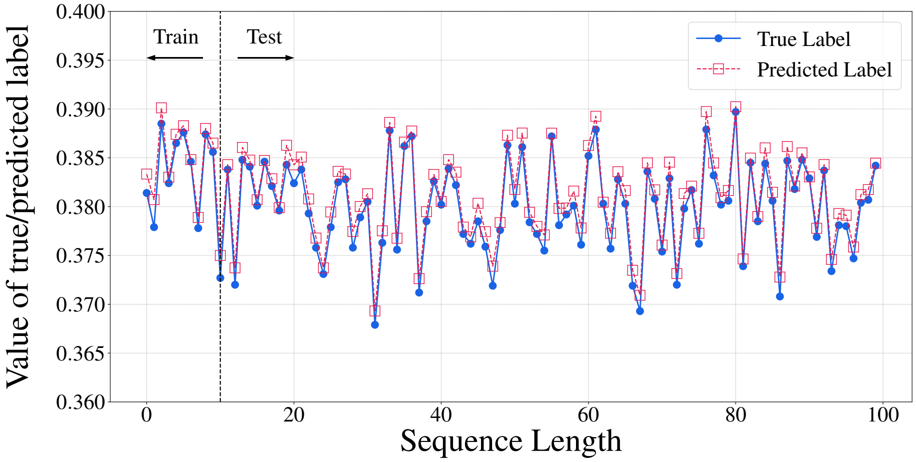

Figure 16. A RNN model with one hidden layer MLP for ω trained on sequences of length up to $T = 10$ sampled according to Figure 12 can generalize to unseen test sequences (Figure 12). Additionally, the compositional generalization holds even beyond the training length.

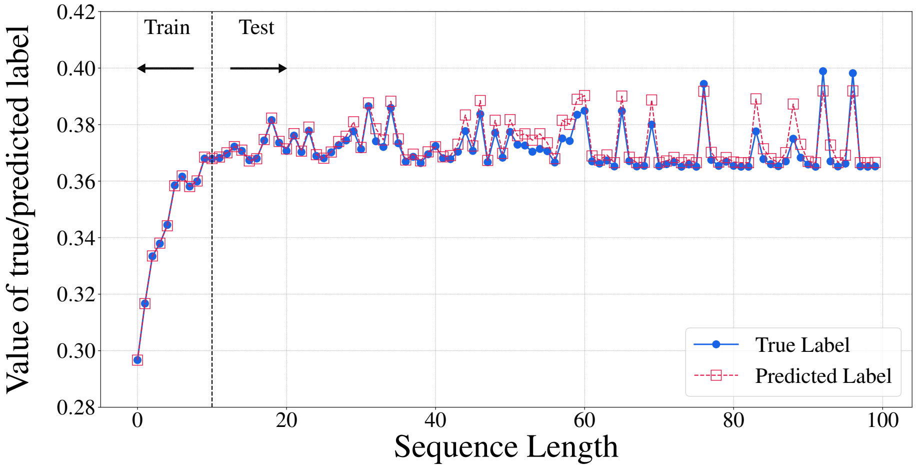

Figure 17. A deep set model with *two* hidden layer MLP for ω, ψ trained on sequences of length up to $T = 10$ sampled according to Figure 12 can generalize to unseen test sequences. Additionally, the compositional generalization holds even beyond the training length.

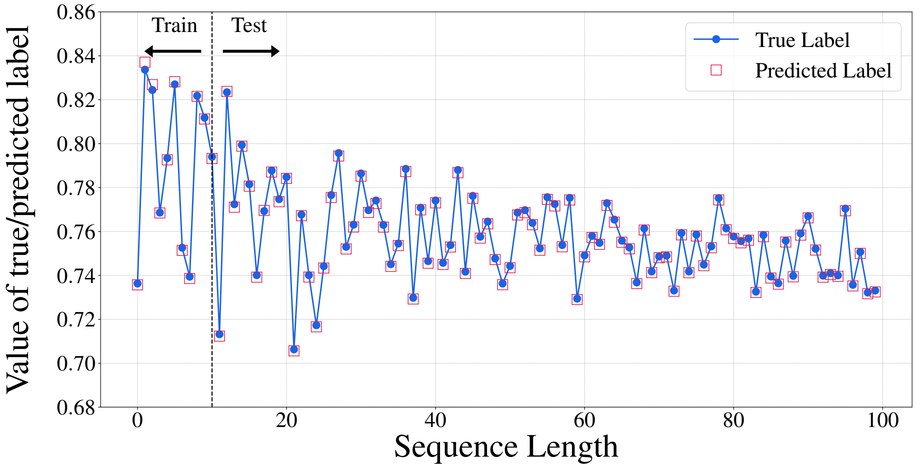

Figure 18. A Transformer model with softmax attention and *two* hidden layer MLP for ω trained on sequences of length up to $T = 10$ sampled according to Figure 12 can generalize to unseen test sequences. Additionally, the compositional generalization holds even beyond the training length.

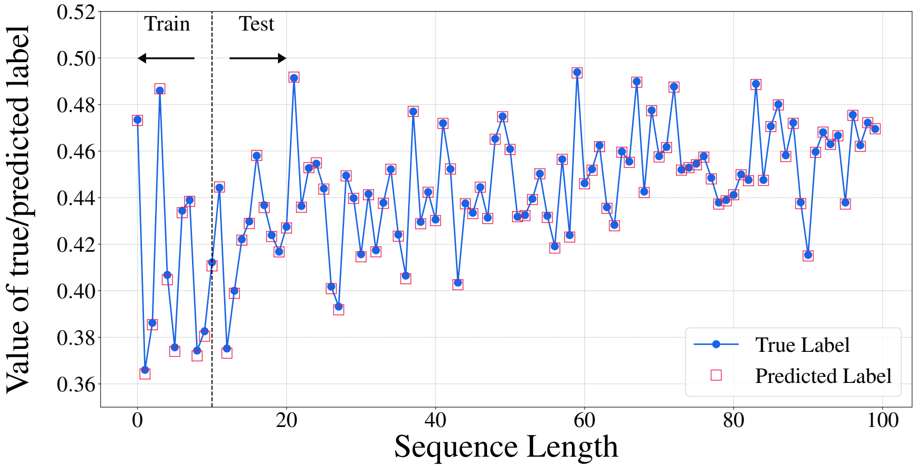

Figure 19. A SSM model with *two* hidden layer MLP for ω trained on sequences of length up to $T = 10$ sampled according to Figure 12 can generalize to unseen test sequences. Additionally, the compositional generalization holds even beyond the training length.

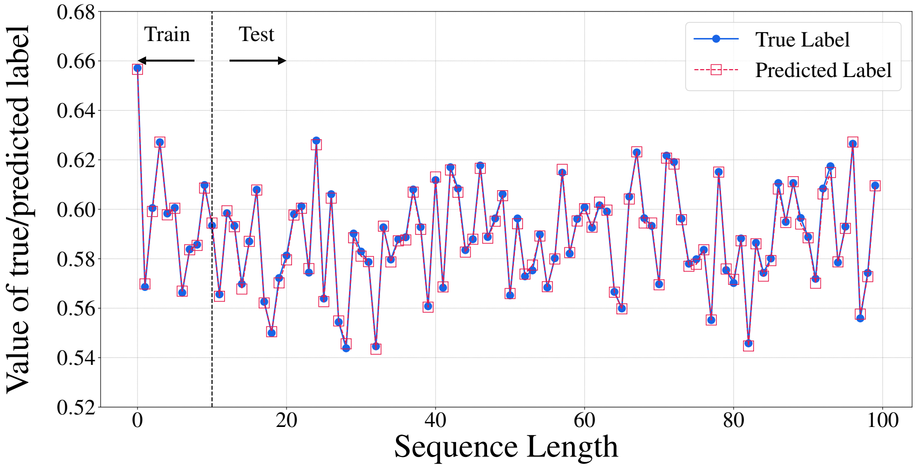

Figure 20. A RNN model with *two* hidden layer MLP for ω trained on sequences of length up to $T = 10$ sampled according to Figure 12 can generalize to unseen test sequences. Additionally, the compositional generalization holds even beyond the training length.

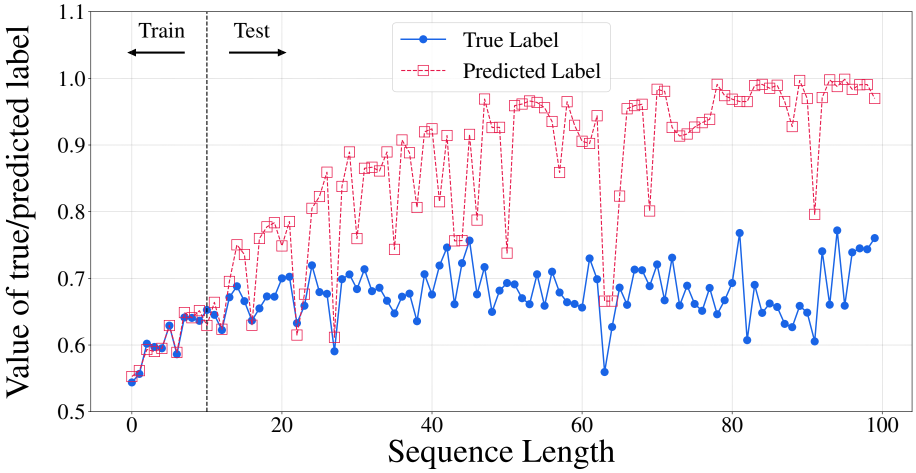

Figure 21. A failure case of length generalization: The predictions come from a deep set with linear layers for ψ, ω trained to predict the sequences (of length up to T) output by a deep set with 1 hidden layer MLPs for ϕ, ρ. In this case the realizability condition does not hold, and the learner fails to length generalize.

Model	Test Loss $\times 10^6$	R^2
Deep set	0.08 ± 0.02	0.96 ± 0.01
Transformer	3.06 ± 1.11	1.00 ± 0.00
SSM	5.92 ± 2.47	1.00 ± 0.00
RNN	0.35 ± 0.17	0.96 ± 0.01

Table 3. Compositional generalization: Test ℓ_2 loss and R^2 score for models with *two* hidden layers on sequences of length $T = 10$. A strong linear relationship is observed for all models for new sequences made of unseen token combinations.