Non-Asymptotic Length Generalization

Thomas Chen[†], Tengyu Ma[†], Zhiyuan Li[‡]
[†]Stanford University [‡]Toyota Technological Institute at Chicago

Length Generalization

- Length generalization is a phenomenon where a model trained on shorter length instances of a task performs well on longer length instances.
- When training transformers in a supervised setting, length generalization is empirically observed for some ground-truth functions but not others Nogueira et al. (2021); Nye et al. (2021); Shaw et al. (2021); Anil et al. (2022); Delétang et al. (2023); Ruoss et al. (2023); Zhou et al. (2023); Jelassi et al. (2023); Zhou et al. (2024)
- RASP-L Conjecture: when training transformers, ground-truth functions expressible by a short RASP program are usually learned by the transformer in a length generalizable way Zhou et al. (2023); Huang et al. (2024)
- **This work:** In an abstract setting, when can we prove that there is a concrete function $F : \mathbb{N} \to \mathbb{N}$ where if the ground-truth has description length c, then training inputs of length at most $F(c)$ are sufficient to ensure length generalization?

Setup

- Hypothesis class $\mathcal{F}$, subset of computable functions.
- Encoding system $\mathcal{R} : \{0,1\}^* \to \mathcal{F}$ is a computable mapping from descriptions to functions. E.g. $\mathcal{R}_{\text{DFA}}$ maps descriptions of DFAs under the standard encoding to the Turing Machine (TM) computing that DFA. Let $\mathcal{F}^{\mathcal{R}} := \{\mathcal{R}(p) : p \in \{0,1\}^*\}$.
- Length N Training Dataset: $D_N(f_*) := \{(x, f_*(x)) : x \in \{0,1\}^*, |x| \leq N\}$
- Learning Algorithm $\mathcal{A} : D_N(f_*) \to \{0,1\}^*$ for any $N \geq 0, f_* \in \mathcal{F}$.
- We say a learning algorithm $\mathcal{A}$ *length-generalizably learns* a function f_* at input length N w.r.t. encoding system $\mathcal{R}$ iff $\mathcal{R}(\mathcal{A}(D_N(f_*))) = f_*$.

Prior (Asymptotic) Results

Definition (Adapted from Gold (1967))

A function class $\mathcal{F} \subseteq \mathcal{F}^{\mathcal{R}}$ admits *length generalization in the limit* w.r.t. encoding system $\mathcal{R}$ if there exists a learning algorithm $\mathcal{A}$ such that for all $f_* \in \mathcal{F}$, there exists a natural number N such that for all $N' \geq N$, $\mathcal{A}$ length-generalizably learns f_* at input length N'.

Theorem (Adapted from Theorem I.4 of Gold (1967))

For all encoding systems $\mathcal{R}$, the function class $\mathcal{F}^{\mathcal{R}}$ admits length generalization in the limit

- Length generalization in the limit is an asymptotic notion and Theorem 2 predicts length generalization in the limit for the class of primitive-recursive functions (a very large class)
- We seek a finer grained ("non-asymptotic") result of how hard it is to length generalize for various function classes

Non-Asymptotic Length Generalization

- Complexity Measure $\mathcal{C} : \{0,1\}^* \to \mathbb{N}$, which is s.t. there exists a Turing Machine E which enumerates descriptions $p \in \{0,1\}^*$ in an non-decreasing order of $\mathcal{C}(p)$. E.g. $\mathcal{C}(p) = |p|$. Let $\mathcal{C}^{\mathcal{R}}(f_*) := \min_{p:\mathcal{R}(p)=f_*} \mathcal{C}(p)$.

Definition

A function class $\mathcal{F} \subseteq \mathcal{F}^{\mathcal{R}}$ admits *non-asymptotic length generalization* w.r.t. encoding system $\mathcal{R}$ and complexity measure $\mathcal{C}$ if there exists a learning algorithm $\mathcal{A}$ and a computable function $\widehat{N}_{\mathcal{A}}^{\mathcal{R},\mathcal{F}} : \mathbb{N} \to \mathbb{N}$ such that for all $f_* \in \mathcal{F}$ and for all $N' \geq \widehat{N}_{\mathcal{A}}^{\mathcal{R},\mathcal{F}}(\mathcal{C}^{\mathcal{R}}(f_*))$, $\mathcal{A}$ length-generalizably learns f_* at input length N'.

Proposition

Regular languages $\mathcal{F}^{\mathcal{R}}$ admits non-asymptotic length generalization w.r.t. encoding system $\mathcal{R} = \mathcal{R}_{\mathsf{DFA}}$ and complexity measure $\mathcal{C}_{\mathsf{DFA}}$. More specifically, there exists a learning algorithm $\mathcal{A}$ such that $N_{\mathcal{A}}^{\mathcal{R}}(c) \leq 2c - 2$ for all $c \in \mathbb{N}$.[1]

[1] For Learning Algorithm $\mathcal{A}$, let $N_{\mathcal{A}}^{\mathcal{R}}(c) := N_{\mathcal{A}}^{\mathcal{R},\mathcal{F}^{\mathcal{R}}}(c)$

Minimum Complexity Interpolator

Hyperparameters: Complexity measure $\mathcal{C}$, encoding system $\mathcal{R}$
Input : Training set S for some $f_* \in \mathcal{F}^{\mathcal{R}}$, $S \subseteq \{(x, f_*(x)) \mid x \in \{0,1\}^*\}$
Output : $\arg \min_{\substack{p \in \{0,1\}^* \\ \forall (x,y) \in S,\ y = \mathcal{R}(p)(x)}} \mathcal{C}(p)$

Algorithm 1: Minimum-Complexity Interpolator ($\mathcal{A}^{\mathcal{R},\mathcal{C}}_{\mathsf{mci}}$)

Theorem

Given any encoding system $\mathcal{R}$ and complexity measure $\mathcal{C}$, for all $c \in \mathbb{N}$, it holds that

$$N^{\mathcal{R}}_{\mathcal{A}^{\mathcal{R},\mathcal{C}}_{\mathsf{mci}}}(c) = \min_{\mathcal{A}} N^{\mathcal{R}}_{\mathcal{A}}(c) = \min\{n \in \mathbb{N} : \forall f \neq f' \in \mathcal{F}^{\mathcal{R}}, \exists x \in \{0,1\}^{\leq n} \ s.t.\ f(x) \neq f'(x)\}.$$

We term the latter quantity as the length complexity for c w.r.t. $\mathcal{R}$.

Equivalent Definitions of Non-Asymptotic Length Generalization

Definition

The *Language Equivalence Problem* for encoding system $\mathcal{R}$ is the computational problem where given any $p, q \in \{0,1\}^*$, determine whether $\mathcal{R}(p) = \mathcal{R}(q)$.

Lemma

For any encoding system $\mathcal{R}$ and complexity measure $\mathcal{C}$ satisfying the aforementioned conditions, the Language Equivalence problem for $\mathcal{R}$ is decidable if and only if $\mathcal{F}^{\mathcal{R}}$ admits non-asymptotic length generalization w.r.t. $\mathcal{R}, \mathcal{C}$.

Proposition

Let $\mathcal{R}_{\mathsf{CFG}}$ be the encoding system for CFGs and $\mathcal{C}_{\mathsf{CFG}}(\langle G \rangle)$ is the complexity measure that maps a CFG $G = (N, T, P, S = \{0,1\})$ to $|N| + |T| + |P|$. Then for any learning algorithm $\mathcal{A}$, the length complexity, $N_{\mathcal{A}}^{\mathcal{R}_{\mathbf{CFG}}} : \mathbb{N} \to \mathbb{N}$, is not computably bounded.

C-RASP Yang & Chiang (2024)

Definition (C-RASP, (Yang & Chiang, 2024))

Boolean-Valued Operations

Initial	$h_j^{(i)} := 1[x_j = a]$ for $a \in \{0, 1\}$
Boolean	$h_j^{(i)} := \neg h_j^{(i')}$ $h_j^{(i)} := h_j^{(i')} \wedge h_j^{(i'')}$
Sign	$h_j^{(i)} := 1[h_j^{(i')} > 0]$
Constant	$h_j^{(i)} := 1$

Count-Valued Operations

Partial Sum	$h_j^{(i)} := \mathsf{ps}(h^{(i')})_j$
Conditional	$h_j^{(i)} := h_j^{(i')} \; ? \; h_j^{(i'')} : h_j^{(i''')}$
Addition	$h_j^{(i)} := h_j^{(i')} + h_j^{(i'')}$
Subtraction	$h_j^{(i)} := h_j^{(i')} - h_j^{(i'')}$
Min/Max	$h_j^{(i)} := \min(h_j^{(i')}, h_j^{(i'')})$ $h_j^{(i)} := \max(h_j^{(i')}, h_j^{(i'')})$
Constant	$h_j^{(i)} := 1$

C-RASP^{1}

Definition (C-RASP^{1})

With integer T, let C-RASP^{1,T} be the set of C-RASP programs where each program f has parameters $a, b, d \in [-T, T]$, $a > 0$. For any $n > 0$, on input $x \in \{0,1\}^n$, f computes:

$$f(x) = 1[a \cdot \mathsf{ps}(x)_n - b \cdot n - d > 0]$$

Theorem

For any $T \in \mathbb{N}$ and ground-truth function $f_* \in$ C-RASP^{1,T}, the Minimum-Complexity Interpolator $\mathcal{A}_{\mathsf{mci}}^{\mathcal{R},\mathcal{C}}$ can length generalize given $O(T^2)$ inputs.

C-RASP^2

Definition (C-RASP^2)

With integers T and $1 \leq K \leq T^2$, let C-RASP^{2,K,T} be the set of programs where each program f has parameters $0 < z \leq T$, $\forall i \in [K]$, $a^{(i)}, b^{(i)}, \lambda_i \in \{-T, \ldots, T\}$, with $a^{(i)} > 0$ and $\frac{b^{(i)}}{a^{(i)}} \in (0, 1)$. [2] For any $n > 0$, on input $x \in \{0,1\}^n$, the first layer computes the values of K heads, $\{h^{(i)}\}_{i \in [K]}$, on the n prefixes of x:

$$\forall j \in [n], \forall i \in [K], h_j^{(i)} = 1[\mathsf{ps}(x)_j > \frac{b^{(i)}}{a^{(i)}} j]$$

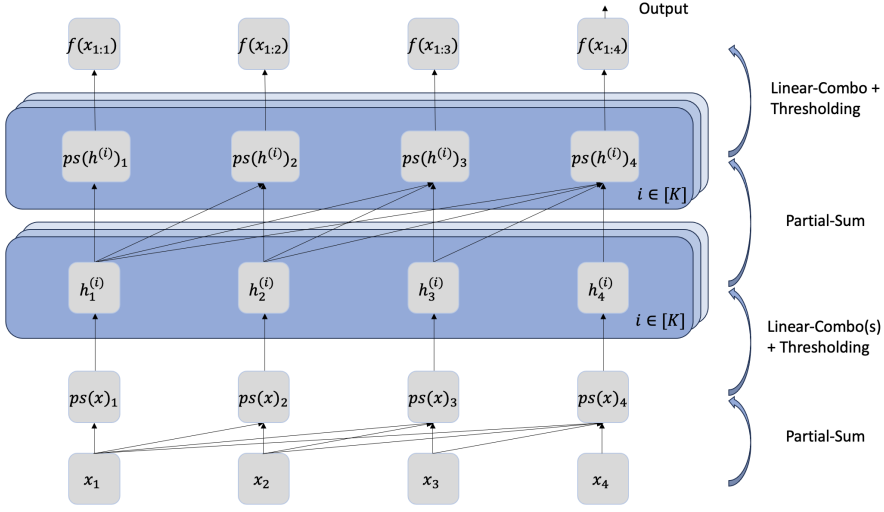
The second layer computes the output: $f(x) = 1[\sum_{i \in [K]} \lambda_i \mathsf{ps}(h^{(i)})_n > z \cdot n]$.

Theorem (Main Result)

For any $T \in \mathbb{N}, K \leq T^2$, and ground-truth function $f_* \in$ C-RASP^{2,K,T}, the Minimum-Complexity Interpolator $\mathcal{A}_{\mathsf{mci}}^{\mathcal{R},\mathcal{C}}$ can length generalize given length $O(T^{O(K)})$ inputs.

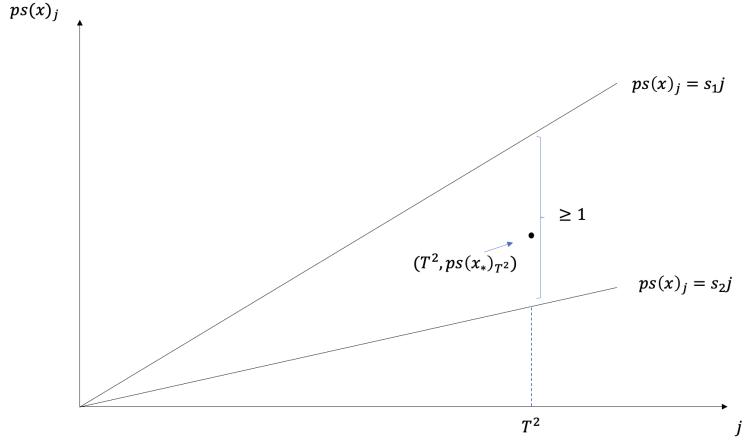
[2] We place some additional technical assumptions on these parameters.

C-RASP[2]


Output
$f(x_{1:1})$
$f(x_{1:2})$
$f(x_{1:3})$
$f(x_{1:4})$
Linear-Combo + Thresholding
$ps(h^{(i)})_1$
$ps(h^{(i)})_2$
$ps(h^{(i)})_3$
$ps(h^{(i)})_4$
$i \in [K]$
Partial-Sum
$h_1^{(i)}$
$h_2^{(i)}$
$h_3^{(i)}$
$h_4^{(i)}$
$i \in [K]$
Linear-Combo(s) + Thresholding
$ps(x)_1$
$ps(x)_2$
$ps(x)_3$
$ps(x)_4$
Partial-Sum
x_1
x_2
x_3
x_4

(Warmup) Proof Sketch of C-RASP[1] Length Generalization

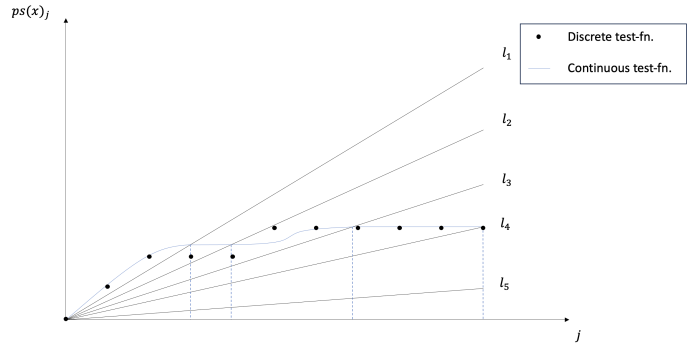
- By the characterization of $\mathcal{A}^{\mathcal{R},\mathcal{C}}_{\text{mci}}$, we want to show that for any $f \neq f' \in$ C-RASP^{1,T}, there is a *short* string x ($|x| \leq T^2$) where $f(x) \neq f'(x)$.
- Suppose $f(x) = 1[\text{ps}(x)_{|x|} > \frac{b}{a}|x|]$, $f'(x) = 1[\text{ps}(x)_{|x|} > \frac{b'}{a'}|x|]$ and $\frac{b}{a} \neq \frac{b'}{a'}$.



Discrete and Continuous Test-Functions

Definition

A continuous test-function $\mathcal{Y}$, with respect to $\{s_i\}_{i\in[k]} \subset \mathbb{Q}^k$, is a 1-Lipschitz, monotone non-decreasing continuous function $[0,1] \to [0,1]$, with $\mathcal{Y}(0) = 0$. The induced activations $(B_1(\mathcal{Y}), \ldots, B_k(\mathcal{Y}))$ of $\mathcal{Y}$ w.r.t. $\{s_i\}_{i\in[k]}$ are: $\forall i \in [k], \quad B_i(\mathcal{Y}) := \int_0^1 1[\mathcal{Y}(j) > s_i \cdot j]dj$.



Note: Discrete Test-Functions have a one-to-one correspondence to strings $x \in \{0,1\}^*$.

(Main Result) Proof Sketch of C-RASP^2 Length Generalization

Key Observation: C-RASP^2 programs are linear threshold functions over "first-layer activations," $(\frac{\mathsf{ps}(h^{(1)})_n}{n}, \frac{\mathsf{ps}(h^{(2)})_n}{n}, \ldots, \frac{\mathsf{ps}(h^{(K)})_n}{n})$.

$$\forall j \in [n], \forall i \in [K], h_j^{(i)} = 1[\mathsf{ps}(x)_j > \frac{b^{(i)}}{a^{(i)}} j]$$

$$f(x) = 1[\sum_{i \in [K]} \lambda_i \frac{\mathsf{ps}(h^{(i)})_n}{n} > z]$$

Suppose $f \neq f' \in$ C-RASP^{2,K,T}. Proof Plan:

1. Characterize the set of all possible activations $\mathbb{A} \subset [0,1]^{2K}$ w.r.t. the first layers of both f and f'.
2. Let $P \subset \mathbb{A}$ be the subset of activations which distinguish f and f'. Show P is *at least a minimal size*.
3. "Discretize" a particular continuous test-function $\mathcal{Y}_{\mathsf{center}}$ which corresponds to P's "center", to get a discrete test-function (string) $\mathcal{X}$ of short length that distinguishes f, f'.

Step 1: Characterize the set of all possible activations $\mathbb{A} \subset [0,1]^{2K}$ w.r.t. the first layers of both f and f'.

Characterizing the Set of Possible First-Layer Activations

- Let $\{s_i\}_{i\in[k]} = \{\frac{b^{(i)}}{a^{(i)}}\}_{i\in[K]} \cup \{\frac{(b^{(i)})'}{(a^{(i)})'}\}_{i\in[K]} \subset (0,1)$, where $K \le k \le 2K$.
- Let $\mathbb{A}(\{s_i\}_{i\in[k]}) := \{(B_1(\mathcal{Y}), \ldots, B_k(\mathcal{Y})) : \mathcal{Y}$ continuous test-function w.r.t $\{s_i\}_{i\in[k]}\}$.

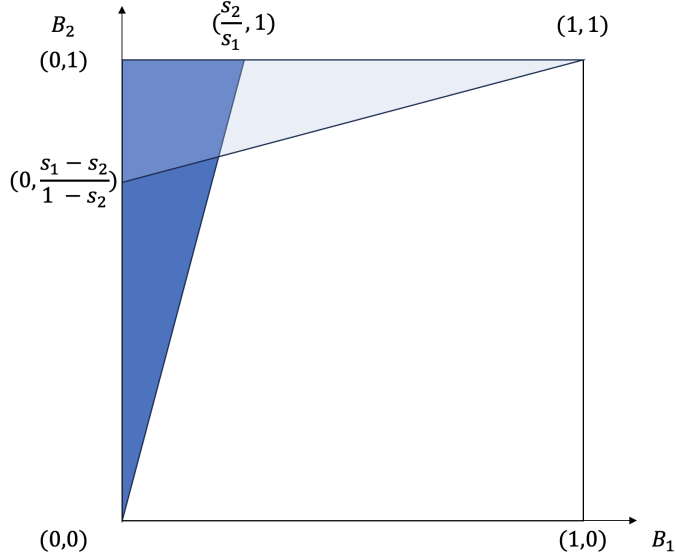
Lemma (Characterization of $\mathbb{A}(\{s_i\}_{i\in[k]})$)

For any $\{s_i\}_{i\in[k]} \subset (0,1)$, there are a finite number of k-dimensional convex polytopes $\{A_j\}_{j\in[2^{k-1}]}$, such that:

$$\mathbb{A}(\{s_i\}_{i\in[k]}) = \bigcup_{j\in[N_k]} A_j$$

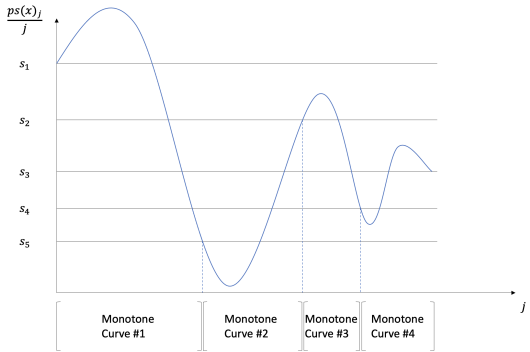
Moreover, if $\{s_i\}_{i\in[k]} \subset (0,1)$ are rational numbers with maximum denominator T, the faces of each polytope A_j is given by a precision $\mathrm{poly}(K,T)$-linear inequality.

Depiction of $\mathbb{A}(\{s_1, s_2\})$ for the $k = 2$ Case



Proof Idea of Lemma Characterizing $\mathbb{A}(\{s_i\}_{i\in[k]})$

- Show that for every continuous test-function $\mathcal{Y}$, there exists another continuous test-function $\mathcal{Y}'$ where $\forall i \in [k], B_i(\mathcal{Y}) = B_i(\mathcal{Y}')$ and $\mathcal{Y}'$ follows one of 2^{k-1} possible *schema* (blueprints).
- Each schema is characterized by the order which it crosses the k *lines*. Each is a concatenation of curves where later curves may not cross as many lines as earlier curves.



Step 2: Show that the subset of activations, $P \subset [0,1]^{2K}$, which cause f and f' to disagree is *at least a minimal size*.

Lower-bound on Size of Polytope P

Let $H_1^+ := \{(B_1, \ldots, B_k) : \sum_{i \in [K]} \lambda_i B_{\mathsf{ord}(1,i)} > z\}$ (activations causing f to return 1)
Let $H_2^+ := \{(B_1, \ldots, B_k) : \sum_{i \in [K]} \lambda'_i B_{\mathsf{ord}(2,i)} < z'\}$ (activations causing f' to return 0).
Suppose $\mathbb{A}(\{s_i\}_{i \in [k]}) \cap H_1^+ \cap H_2^+ \neq \emptyset$. Then there exists $j_* \in [2^{k-1}]$ where
$P := A_{j_*} \cap H_1^+ \cap H_2^+ \neq \emptyset, P \subset [0,1]^k$, and where P satisfies:

Lemma

Consider a nonempty k-dimensional polytope $P \subset \mathbb{R}^k$ with vertices V and N faces. Suppose the faces of P are each defined by a linear inequality over variables $\{B_i\}_{i \in [k]}$, with integer coefficients of magnitude at most p_{face}, where points on the face satisfy the linear inequality with equality. For $j \in [N]$, define L_j as the linear inequality for the jth face of P. Then, for any $j \in [N]$, for any vertex $x \in V$ which does not lie on the jth face of P, we have the following lower bound on the margin of x on the jth face of P.

$$L_j(x) \gtrapprox \frac{1}{(p_{\mathsf{face}}\sqrt{k})^k}$$

Step 3: "Discretize" a particular continuous test-function $\mathcal{Y}_{\mathsf{center}}$ which corresponds to P's "center", to get a discrete test-function (string) $\mathcal{X}$ of short length that distinguishes f, f'.

Discretization of Continuous Test-Functions

- Suppose polytope $P := A_{j_*} \cap H_1^+ \cap H_2^+ \subset [0,1]^k, P \neq \emptyset$ and has vertices V.
- There is a continuous test-function $\mathcal{Y}_*$ whose activations are $(B_i(\mathcal{Y}_*))_{i\in[k]} = \frac{1}{|V|}\sum_{x\in V} x$.
- The margin of $(B_i(\mathcal{Y}_*))_{i\in[k]}$ to any face of P is at least $\gamma = \frac{1}{|V|}\frac{1}{(p_{\text{face}}\sqrt{k})^k}$, with $p_{\text{face}} = \text{poly}(K,T)$.
- We can discretize $\mathcal{Y}_*$ into a discrete test-function $\mathcal{X}$ of length $n = O(\frac{1}{\gamma}\cdot T^k)$ via the following lemma:

Lemma

For any continuous test-function $\mathcal{Y}$ w.r.t. $\{s_i\}_{i\in[k]}$ with p-precision activations, there exists an $n_0 \leq O(p\cdot T^k)$ so that for any positive integer multiple n of n_0, there exists a discrete test-function $\mathcal{X}$ of length n so that:

$$\forall i \in [k], |B_i(\mathcal{Y}) - B_i(\mathcal{X})| \leq \frac{\text{poly}(K,T)}{n}$$

- For n large enough (i.e. $O(\frac{1}{\gamma}\cdot T^k)$), $\mathcal{X}$ corresponds to string $x \in \{0,1\}^n$ where $f(x) \neq f'(x)$, since $\mathcal{X}$ *behaves similarly enough* to $\mathcal{Y}_*$ and $\mathcal{Y}_*$ distinguishes f, f'.

Summary

Fn. Class	Complexity of Ground-Truth Function	Length of Training Data Suff. to Generalize
DFAs	number of states, c	$2c - 2$
CFGs	description length, c	no computable bound in c exists
C-RASP[1]	precision, T	$O(T^2)$
C-RASP[2]	precision, T, and number of heads, K	$O(T^{O(K)})$

Table: Summary of results: upper bounds on minimum length of binary strings in training data which suffices for length generalization.

Future Directions

- Extend results to 3-layer C-RASP functions
- Extend results to C-RASP functions with bias terms (i.e. extension to Dyck-1)
- Frameworks of Partial Length Generalization/Statistical Frameworks (E.g. Golowich et al. (2025))

References I

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, 2022. URL https://openreview.net/forum?id=zSkYVeX7bC4.

Delétang, G., Ruoss, A., Grau-Moya, J., Genewein, T., Wenliang, L. K., Catt, E., Cundy, C., Hutter, M., Legg, S., Veness, J., and Ortega, P. A. Neural networks and the chomsky hierarchy, 2023. URL https://openreview.net/pdf?id=WbxHAzkeQcn.

Gold, E. M. Language identification in the limit, 1967. URL https://www.sciencedirect.com/science/article/pii/S0019995867911655.

Golowich, N., Jelassi, S., Brandfonbrener, D., Kakade, S. M., and Malach, E. The role of sparsity for length generalization in transformers, 2025. URL https://arxiv.org/abs/2502.16792.

References II

Huang, X., Yang, A., Bhattamishra, S., Sarrof, Y., Krebs, A., Zhou, H., Nakkiran, P., and Hahn, M. A formal framework for understanding length generalization in transformers, 2024. URL https://openreview.net/forum?id=U49N5V51rU.

Jelassi, S., d'Ascoli, S., Domingo-Enrich, C., Wu, Y., Li, Y., and Charton, F. Length generalization in arithmetic transformers, 2023. URL https://arxiv.org/abs/2306.15400.

Nogueira, R., Jiang, Z., and Lin, J. Investigating the limitations of transformers with simple arithmetic tasks, 2021. URL https://arxiv.org/abs/2102.13019.

Nye, M., Andreassen, A. J., Gur-Ari, G., Michalewski, H., Austin, J., Bieber, D., Dohan, D., Lewkowycz, A., Bosma, M., Luan, D., Sutton, C., and Odena, A. Show your work: Scratchpads for intermediate computation with language models, 2021. URL https://arxiv.org/abs/2112.00114.

References III

Ruoss, A., Delétang, G., Genewein, T., Grau-Moya, J., Csordás, R., Bennani, M., Legg, S., and Veness, J. Randomized positional encodings boost length generalization of transformers, 2023. URL https://aclanthology.org/2023.acl-short.161/.

Shaw, P., Chang, M.-W., Pasupat, P., and Toutanova, K. Compositional generalization and natural language variation: Can a semantic parsing approach handle both?, 2021. URL https://aclanthology.org/2021.acl-long.75/.

Yang, A. and Chiang, D. Counting like transformers: Compiling temporal counting logic into softmax transformers, 2024. URL https://openreview.net/forum?id=FmhPg4UJ9K#discussion.

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, 2023. URL https://openreview.net/forum?id=AssIuHnmHX.

References IV

Zhou, Y., Alon, U., Chen, X., Wang, X., Agarwal, R., and Zhou, D. Transformers can achieve length generalization but not robustly, 2024. URL https://openreview.net/pdf?id=DWkWIh3vFJ.