

efficiently as any model, up to constant factors. Since Dyck- (k,m) is a regular (finite-state) language, the application of general-purpose RNN constructions trivially proves that RNNs can generate the language (Merrill, 2019). However, the best construction we are aware of uses $O(k^{\frac{m}{2}})$ hidden units (Horne and Hush, 1994; Indyk, 1995), where k is the vocabulary size and m is the nesting depth, which is exponential.¹ We provide an explicit construction proving that a Simple (Elman; Elman (1990)) RNN can generate any Dyck- (k,m) using only $6m\lceil\log k\rceil - 2m = O(m\log k)$ hidden units, an exponential improvement. This is not just a strong result relative to RNNs’ general capacity; we prove that even computationally unbounded models generating Dyck- (k,m) require $\Omega(m\log k)$ hidden units, via a simple communication complexity argument.

Our proofs provide two explicit constructions, one for the Simple RNN and one for the LSTM, which detail how these networks can use their hidden states to simulate stacks in order to efficiently generate Dyck- (k,m) . The differences between the constructions exemplify how LSTMs can use *exclusively* their gates, ignoring their Simple RNN subcomponent entirely, to reduce the memory by a factor of 2 compared to the Simple RNN.²

We prove these results under a theoretical setting that aims to reflect the realistic settings in which RNNs have excelled in NLP. First, we assume *finite precision*; the value of each hidden unit is represented by $p = O(1)$ bits. This has drastic implications compared to existing work (Siegelmann and Sontag, 1992; Weiss et al., 2018; Merrill, 2019; Merrill et al., 2020). It implies that **only regular languages** can be generated by any machine with d hidden units, since they can take on only 2^{pd} states (Korsky and Berwick, 2019). This points us to focus on whether languages can be implemented memory-efficiently. Second, we consider networks as language *generators*, not acceptors;³ informally, RNNs’ practical successes have been primarily as generators, like in language modeling and machine translation (Karpathy et al., 2015; Wu et al., 2016).

Finally, we include a preliminary study in learn-

¹For hard-threshold neural networks, the lower-bound would be $\Omega(k^{\frac{m}{2}})$ if Dyck- (k,m) were an arbitrary regular language.

²We provide implementations at <https://github.com/john-hewitt/dyckkm-constructions/>.

³Acceptors consume a whole string and then decide whether the string is in the language; generators must decide which tokens are possible continuations at each timestep.

ing Dyck- (k,m) with LSTM LMs from finite samples, finding for a range of k and m that learned LSTM LMs extrapolate well given the hidden sizes predicted by our theory.

In summary, we prove that RNNs are memory optimal in generating a family of bounded hierarchical languages that forms the scaffolding of natural language syntax by describing mechanisms that allow them to do so; this provides theoretical insight into their empirical success.

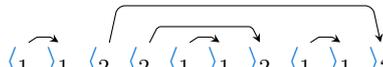
1.1 Motivating bounded hierarchical structure

Hierarchical structure is central to human language production and comprehension, showing up in grammatical constraints and semantic composition, among other properties (Chomsky, 1956). Agreement between subject and verb in English is an intuitive example:

Laws the lawmaker the reporter questions writes are (1)

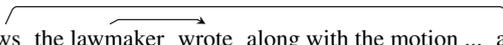


The Dyck- k languages—well-nested brackets of k types—are the prototypical languages of hierarchical structure; by the Chomsky-Schützenberger Theorem (Chomsky and Schützenberger, 1959), they form the scaffolding for any context-free language. They have a simple structure:



However, human languages are unlike Dyck- k and other context-free languages in that they exhibit *bounded memory requirements*. Dyck- k requires storage of an unboundedly long list of open brackets in memory. In human language, as the center-embedding depth grows, comprehension becomes more difficult, like in our example sentence above (Miller and Chomsky, 1963). Empirically, center-embedding depth of natural language is rarely greater than 3 (Jin et al., 2018; Karlsson, 2007). However, it does exhibit long-distance, shallow hierarchical structure:

Laws the lawmaker wrote along with the motion ... are



Our Dyck- (k,m) language puts a bound on depth in Dyck- k , capturing the long-distance hierarchical structure of natural language as well as its bounded memory requirements.⁴

⁴For further motivation, we note that center-embedding directly implies bounded memory requirements in arc-eager

2 Related Work

This work contributes primarily to the ongoing theoretical characterization of the expressivity of RNNs. Siegelmann and Sontag (1992) proved that RNNs are Turing-complete if provided with infinite precision and unbounded computation time. Recent work in NLP has taken an interest in the expressivity of RNNs under conditions more similar to RNNs’ practical uses, in particular assuming one “unrolling” of the RNN per input token, and precision bounded to be logarithmic in the sequence length. In this setting, Weiss et al. (2018) proved that LSTMs can implement simplified counter automata; the implications of which were explored by Merrill (2019, 2020). In this same regime, Merrill et al. (2020) showed a strict hierarchy of RNN expressivity, proving among other results that RNNs augmented with an external stack (Grefenstette et al., 2015) can recognize hierarchical (Context-Free) languages like Dyck- k , but LSTMs and RNNs cannot.

Korsky and Berwick (2019) prove that, given infinite precision, RNNs can recognize context-free languages. Their proof construction uses the floating point precision to simulate a stack, e.g., implementing `push` by dividing the old floating point value by 2, and `pop` by multiplying by 2. This implies that one can recognize any language requiring a *bounded* stack, like our Dyck- (k,m) , by providing the model with precision that scales with stack depth. In contrast, our work assumes that the precision cannot scale with the stack depth (or vocabulary size); in practice, neural networks are used with a fixed precision (Hubara et al., 2017).

Our work also connects to empirical studies of what RNNs can *learn* given finite samples. Considerable evidence has shown that LSTMs can learn languages requiring counters (but Simple RNNs do not) (Weiss et al., 2018; Sennhauser and Berwick, 2018; Yu et al., 2019; Suzgun et al., 2019), and neither Simple RNNs nor LSTMs can learn Dyck- k . In our work, this conclusion is foregone because Dyck- k requires unbounded memory while RNNs have finite memory; we show that LSTMs extrapolate well on Dyck- (k,m) , the memory-bounded variant of Dyck- k . Once augmented with an external (unbounded) memory, RNNs have been shown to learn hierarchical languages (Suzgun et al., 2019; Hao et al., 2018; Grefenstette et al., 2015; Joulin and Mikolov, 2015). Finally, considerable study

left-corner parsers (Resnik, 1992).

has gone into what RNN LMs learn about natural language syntax (Lakretz et al., 2019; Khandelwal et al., 2018; Gulordava et al., 2018; Linzen et al., 2016).

3 Preliminaries and definitions

3.1 Formal languages

A formal language \mathcal{L} is a set of strings $\mathcal{L} \subseteq \Sigma^*\omega$ over a fixed vocabulary, Σ (with the end denoted by special symbol ω). We denote an arbitrary string as $w_{1:T} \in \Sigma^*\omega$, where T is the string length.

The Dyck- k language is the language of nested brackets of k types, and so has $2k$ words in its vocabulary: $\Sigma = \{\langle_i, \rangle_i\}_{i \in [k]}$. Any string in which brackets are well-nested, i.e., each \langle_i is closed by its corresponding \rangle_i , is in the language. Formally, we can write it as the strings generated by the following context-free grammar:

$$\begin{aligned} X &\rightarrow | \langle_i X \rangle_i X \\ &| \epsilon, \end{aligned}$$

where ϵ is the empty string.⁵ The memory necessary to generate any string in Dyck- k is proportional to the number of *unclosed* open brackets at any time. We can formalize this simply by counting how many more open brackets than close brackets there are at each timestep:

$$d(w_{1:t}) = \text{count}(w_{1:t}, \langle) - \text{count}(w_{1:t}, \rangle)$$

where $\text{count}(w_{1:t}, a)$ is the number of times a occurs in $w_{1:t}$. We can now define Dyck- (k,m) by combining Dyck- k with a depth bound, as follows:

Definition 1 (Dyck- (k,m)). For any positive integers k, m , Dyck- (k,m) is the set of strings

$$\{w_{1:T} \in \text{Dyck-}k \mid \forall t=1, \dots, T, d(w_{1:t}) \leq m\}$$

3.2 Recurrent neural networks

We now provide definitions of recurrent neural networks as probability distributions over strings, and define what it means for an RNN to generate a formal language. We start with the most basic form of RNN we consider, the Simple (Elman) RNN:

Definition 2 (Simple RNN (generator)). A Simple RNN (generator) with d hidden units is a probabil-

⁵And ω appended to the end.

ity distribution f_θ of the following form:

$$\begin{aligned} h_0 &= \mathbf{0} \\ h_t &= \sigma(W h_{t-1} + U x_t + b) \\ w_t | w_{1:t-1} &\sim \text{softmax}(g_\theta(h_{t-1})) \end{aligned}$$

where $h_t \in \mathbb{R}^d$ and the function g has the form $g_\theta(h_{t-1}) = V h_{t-1} + b_v$. The input $x_t = E w_t$; overloading notation, w_t is the one-hot (indicator) vector representing the respective word. θ is the set of trainable parameters, W, U, b, V, b_v, E .

The Long Short-Term Memory (LSTM) model (?) is a popular extension to the Simple RNN, intended to ease learning by resolving the vanishing gradient problem. In this work, we're not concerned with learning but with expressivity. We study whether the LSTM's added complexity enables it to generate Dyck- (k, m) using less memory.

Definition 3 (LSTM (generator)). An LSTM (generator) with d hidden units is a probability distribution f_θ of the following form:

$$\begin{aligned} h_0, c_0 &= \mathbf{0} \\ f_t &= \sigma(W_f h_{t-1} + U_f x_t + b_f) \\ i_t &= \sigma(W_i h_{t-1} + U_i x_t + b_i) \\ o_t &= \sigma(W_o h_{t-1} + U_o x_t + b_o) \\ \tilde{c}_t &= \tanh(W_{\tilde{c}} h_{t-1} + U_{\tilde{c}} x_t + b_{\tilde{c}}) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\ h_t &= o_t \odot \tanh(c_t) \\ w_t | w_{1:t-1} &\sim \text{softmax}(g_\theta(h_{t-1})) \end{aligned}$$

where $h_t, c_t \in \mathbb{R}^d$, the function g has the form $g_\theta(h_{t-1}) = V h_{t-1} + b$, and $x_t = E w_t$, where w_t is overloaded as above. θ is the set of trainable parameters: all W, U, b , as well as V, E .

Notes on finite precision. Under our finite precision setting, each hidden unit is a rational number specified using p bits; hence it can take on any value in $\mathbb{P} \subset \mathbb{Q}$, where $|\mathbb{P}| = 2^p$. Each construction is free to choose its specific subset.⁶ A machine with d such hidden units thus can take on any of 2^{dp} configurations.

Our constructions require the sigmoid ($\sigma(x) = \frac{1}{1+e^{-x}}$) and tanh nonlinearities to saturate (that is, take on the values at the bounds of their ranges) to ensure arbitrarily long-distance dependencies. Under standard definitions, these functions approach

⁶The \mathbb{P} for our constructions is provided in Appendix G.3.

but never take on their bounding values. Fortunately, under finite precision, we can provide non-standard definitions under which, if provided with large enough inputs, the functions saturate.⁷ Let there be $\beta \in \mathbb{R}$ such that $\sigma(x) = 1$ if $x > \beta$, and $\sigma(x) = 0$ if $x < -\beta$. Likewise for hyperbolic tangent, $\tanh(x) = 1$ if $x > \beta$, and $\tanh(x) = -1$ if $x < -\beta$. This reflects empirical behavior in toolkits like PyTorch (Paszke et al., 2019).

3.3 Formal language generation

With this definition of RNNs as generators, we now define what it means for an RNN (a distribution) to generate a language (a set). Intuitively, since a formal language is a set of strings \mathcal{L} , our definition should be such that a distribution generates \mathcal{L} if its probability mass on the set of all strings $\Sigma^* \omega$ is concentrated on the set \mathcal{L} . So, we first define the set of strings on which a probability distribution concentrates its mass. The key intuition is to control the local token probabilities $f_\theta(w_t | w_{1:t-1})$, not the global $f_\theta(w_{1:T})$, which must approach zero with sequence length.

Definition 4 (locally ϵ -truncated support). Let f_θ be a probability distribution over $\Sigma^* \omega$, with conditional probabilities $f_\theta(w_t | w_{1:t-1})$. Then the locally ϵ -truncated support of the distribution is the set

$$\{w_{1:T} \in \Sigma^* \omega : \forall t \in 1 \dots T, f_\theta(w_t | w_{1:t-1}) \geq \epsilon\}.$$

This is the set of strings such that the model assigns at least ϵ probability mass to each token conditioned on the prefix leading up to that token. A distribution generates a language, then, if there exists an ϵ such that the locally truncated support of the distribution is equal to the language:⁸

Definition 5 (generating a language). A probability distribution f_θ over Σ^* generates a language $\mathcal{L} \subseteq \Sigma^*$ if there exists $\epsilon > 0$ such that the locally ϵ -truncated support of f_θ is \mathcal{L} .

4 Formal results

We now state our results. We provide intuitive proof sketches in the next section, and leave the full proofs to the Appendix. We start with an application of known work to prove that Dyck- (k, m) can be generated by RNNs.

⁷For example, because the closest representable number (in \mathbb{P}) to the true value of $\sigma(x)$ for some x is 1 instead of some number < 1 .

⁸We also note that any f_θ generates multiple languages, since one can vary the parameter ϵ ; for example, any softmax-defined distribution must generate Σ^* with ϵ small because they assign positive mass to all strings.

Theorem 1 (Naive generation of Dyck- (k,m)). For any $k, m \in \mathbb{Z}^+$, there exists a Simple RNN f_θ with $O(k^{m+1})$ hidden units that generates Dyck- (k,m) .

The proof follows by first recognizing that there exists a deterministic finite automaton (DFA) with $O(k^{m+1})$ states that generates Dyck- (k,m) . Each state of the DFA is a sequence of up to m unclosed brackets (of k possible types), implying $k^{m+1} - 1$ total states. Then, one applies a general-purpose algorithm for implementing DFAs with $|Q|$ states using an RNN with $O(|Q|)$ hidden units (Omlin and Giles, 1996; Merrill, 2019). Intuitively, this construction assigns a separate hidden unit to each state.⁹

For our results, we first present two theorems for the Simple RNN and LSTM that use $O(mk)$ hidden units by simulating a stack of m $O(k)$ -dimensional vectors, which are useful for discussing the constructions. Then we show how to reduce to $O(m \log k)$ via an efficient encoding of k symbols in $O(\log k)$ space.

Theorem 2. For any $k, m \in \mathbb{Z}^+$, there exists a Simple RNN f_θ with $2mk$ hidden units that generates Dyck- (k,m) .

We state $2mk$ exactly instead of $O(mk)$ because this exactness is interesting and the constant is small; further, we find that the modeling power of the LSTM leads to a factor of 2 improvement:

Theorem 3. For any $k, m \in \mathbb{Z}^+$, there exists a LSTM f_θ with mk hidden units and $W_{\bar{c}} = \mathbf{0}$ that generates Dyck- (k,m) .

We point out the added property that $W_{\bar{c}} = \mathbf{0}$ because, informally, this matrix corresponds to the recurrent matrix W of the Simple RNN; it’s the only matrix operating on LSTM’s memory that isn’t used in computing a gate. Thus, the LSTM we provide as proof uses *only* its gates.

Using the same mechanisms as in the proofs above but using an efficient encoding of each stack element in $O(\log k)$ units, we achieve the following.

Theorem 4. For any $k, m \in \mathbb{Z}^+$, where $k > 1$, there exists a Simple RNN f_θ with $6m \lceil \log k \rceil - 2m$ hidden units that generates Dyck- (k,m) .

Likewise, for LSTMs, we achieve a more memory-efficient generator.

⁹The construction of Indyk (1995) may achieve $O(\sqrt{|Q|})$ in this case (they do not discuss how vocabulary size affects construction size), but this is still $O(k^{\frac{m}{2}})$ and thus intractable.

Theorem 5. For any $k, m \in \mathbb{Z}^+$, where $k > 1$, there exists an LSTM f_θ with $3m \lceil \log k \rceil - m$ hidden units and $W_{\bar{c}} = \mathbf{0}$ that generates Dyck- (k,m) .

Note on memory. While we have emphasized expressive power under memory constraints—what functions can be expressed, not what is learned in practice—neural networks are frequently intentionally overparameterized to aid learning (Zhang et al., 2017; Schwartz-Ziv and Tishby, 2017). Even so, known constructions for Dyck- (k,m) would require a number of hidden units far beyond practicality. Consider if we were to use a vocabulary size of 100,000, and a practical depth bound of 3. Then if we were using a k^{m+1} hidden unit construction to generate Dyck- (k,m) , we would need $100,000^4 = 10^{20}$ hidden units. By using our LSTM construction, however, we would need only $3 \times 3 \times \lceil \log_2(100,000) \rceil - 1 \times 3 = 150$ hidden units, suggesting that networks of the size commonly used in practice are large enough to learn these languages.

Lower bound. We also show that the bounds in Theorems 4, 5 are tight. Specifically, the following theorem formalizes the statement that *any* algorithm that uses a d -dimensional finite-precision vector memory to generate Dyck- (k,m) must use $d \in \Omega(m \log k)$ memory, implying that RNNs are optimal for doing so, up to constant factors.

Theorem 6 ($\Omega(m \log k)$ to generate Dyck- (k,m)). Let A be an arbitrary function from d -dimensional vectors and symbols $w_t \in \Sigma$ to d -dimensional vectors; $A : \mathbb{P}^d \times \Sigma \rightarrow \mathbb{P}^d$, $A(h_{t-1}, w_t) \mapsto h_t$. Let ψ be an arbitrary function from \mathbb{P}^d to probability distributions over $\Sigma \cup \{\omega\}$. Let f_ψ be a probability distribution over $\Sigma^* \omega$, with the form $f_\psi(w_{1:T}) = \prod_{t=1}^T f(w_t | w_{1:t-1})$, where $f(w_t | w_{1:t-1}) = \psi(h_{t-1})$. If f generates Dyck- (k,m) , then $d \geq \frac{m \log k}{p} = \Omega(m \log k)$.

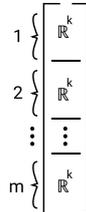
Intuitively, A is an all-powerful recurrent algorithm that represents prefixes $w_{1:t}$ as vectors, and ψ , also all-powerful, turns each vector into a probability distribution over the next token. The proof follows from a simple communication complexity argument: to generate Dyck- (k,m) , any algorithm needs to distinguish between all subsequences of unclosed open brackets, of which there are k^m . So, $2^{dp} \geq k^m$, and the dimensionality $d \geq \frac{m \log k}{p}$. Since $p = O(1)$, we have $d = \Omega(m \log k)$.

5 Stack constructions in Simple RNNs

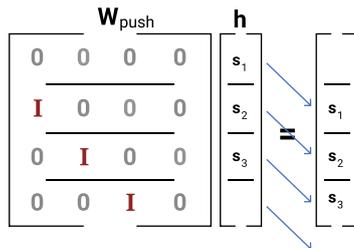
The memory needed to close all the brackets in a Dyck- (k,m) prefix $w_{1:t}$ can be represented as a stack of (yet unclosed) open brackets $[\langle i_1, \dots, \langle i_{m'}], m' \leq m$; reading each new parenthesis either pushes or pops from this stack. Informally, all of our efficient RNN constructions generate Dyck- (k,m) by writing to and reading from an implicit stack that they encode in their hidden states. In this section, we present some challenges in a naive approach, and then describe a construction to solve these challenges. We provide only the high-level intuition; rigorous proofs are provided in the Appendix.

5.1 An extended model

We start by describing what is achievable with an extended model family, *second-order RNNs* (Rabuseau et al., 2019; ?), which allow their recurrent matrix W to be chosen as a function of the input (unlike any of the RNNs we consider.) Under such a model, we show how to store a stack of up to m k -dimensional vectors in mk memory. Such a representation can be thought of as the concatenation of m k -dimensional vectors in the hidden state, like this:

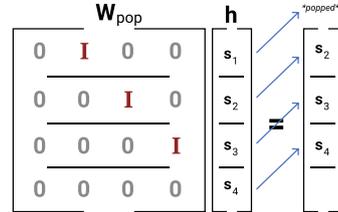


We call each k -dimensional component a “stack slot”. If we want the first stack slot to always represent the top of the stack, then there’s a natural way to implement `pop` and `push` operations. In a `push`, we want to shift all the slots toward the bottom, so there’s room at the top for a new element. We can do this with an off-diagonal matrix W_{push} .¹⁰



¹⁰Note that only needing to store m things means that when we push, there should be nothing in slot m ; otherwise, we’d be pushing element $m + 1$.

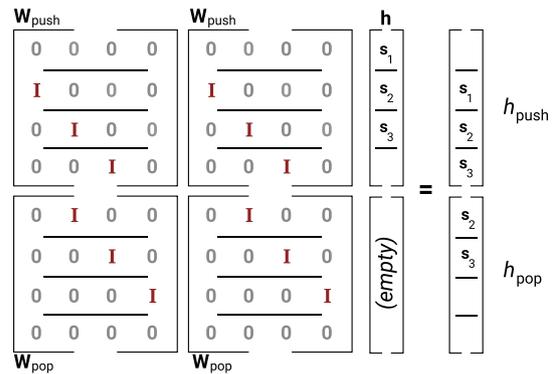
This would implement the Wh_{t-1} part of the Simple RNN equation. We can then write the new element (given by Ux_t) to the first slot. If we wanted to pop, we could do so with another off-diagonal matrix W_{pop} , shifting everything towards the top to get rid of the top element:



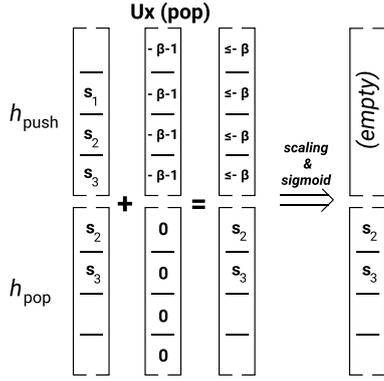
This won’t work for a Simple RNN because it only has one W .

5.2 A Simple RNN Stack in $2mk$ memory

Our construction gets around the limitation of only having a single W matrix in the Simple RNN by doubling the space to $2mk$. Splitting the space h into two mk -sized partitions, we call one h_{pop} , the place where we write the stack if we see a `pop` operation, and the other h_{push} analogously for the `push` operation. If one of h_{pop} or h_{push} is empty (equal to 0) at any time, we can try reading from both of them, as follows:



Our W matrix is actually the concatenation of two of the W_{pop} and W_{push} matrices. Now we have two candidates, both h_{push} and h_{pop} ; but we only want the one that corresponds to `push` if $w_t = \langle i$, or `pop` if $w_t = \rangle i$. We can zero out the unobserved option using the term Ux_t , adding a large negative value to every hidden unit in the stack that doesn’t correspond to `push` if x_t is an open bracket $\langle i$, or `pop` if x_t is a close bracket $\rangle i$:



So $Ux_{\langle i} = [e_i; 0, \dots, -\beta - 1, \dots]$, where e_i is a one-hot representation of $\langle i$, and $Ux_{\rangle i} = [-\beta - 1, \dots, -\beta - 1, 0, \dots]$, where β is the very large value we specified in our finite-precision arithmetic. Thus, when we apply the sigmoid to the new $Wh_{t-1} + Ux_t + b$, whichever of $\{h_{\text{pop}}, h_{\text{push}}\}$ doesn't correspond to the true new state is zeroed out.¹¹

6 Stack construction in LSTMs

We could implement our $2mk$ Simple RNN construction in an LSTM, but its gating functions suggest more flexibility in memory management, and Levy et al. (2018) claim that the LSTM's modeling power stems from its gates. With an LSTM, we achieve the mk of the oracle we described, all while *exclusively* using its gates.

6.1 An LSTM stack in mk memory

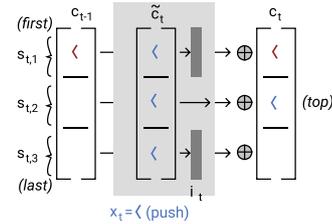
To implement a stack using the LSTM's gates, we use the same intuitive description of the stack as before: m stack slots, each of dimensionality k . However while the top of the stack is the first slot in the Simple RNN, the bottom of the stack is the first slot in the LSTM. Before we discuss mechanics, we introduce the memory dynamics of the model. Working through the example in Figure 2, when we push the first open bracket, it's assigned to slot 1; then the second and third open brackets are assigned to slots 2 and 3. Then a close bracket is seen, so slot 3 is erased. In general, the stack is represented in a contiguous sequence of slots, where the first slot represents the bottom of the stack. Thus, the top of the stack could be at *any* of the m stack slots. So to allow for ease of linearly reading out information from the stack, we store the full stack only in the cell state c_t , and let only the slot corre-

¹¹For whichever of $h_{\text{tmp}} \in \{h_{\text{pop}}, h_{\text{push}}\}$ that is not zeroed out, $\sigma(h_{\text{tmp}}) \neq h_{\text{tmp}}$. Hence, we scale all of U and W to be large, such that $\sigma(Wh_{t-1}) \in \{0, 1\}$.

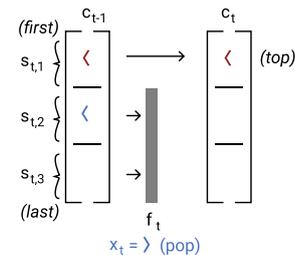
sponding to the top of the stack, which we'll refer to as the *top slot*, through the output gate to the hidden state h_t .

Recall that the LSTM's cell state c_t is specified as $c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$. The values f_t and i_t are the forget gate and input gate, while \tilde{c}_t contains information about the new input.¹² An LSTM's hidden state h is related to the cell state as $h_t = o_t \odot \tanh(c_t)$, where o_t is the output gate.

push mechanics. To implement a push operation, the **input gate** finds the first free slot (that is, equal to $\mathbf{0} \in \mathbb{R}^k$) by observing that it is directly after the top slot of h_{t-1} . The input gate is set to 1 for all hidden units in this first free stack slot, and 0 elsewhere. This is where the new element will be written. The **new cell candidate** (\tilde{c}_t) is used to *attempt* to write the input symbol $\langle i$ to *all* m stack slots. But because of the input gate, $\langle i$ is only written to the first free slot. The **forget gate** is set to 1 everywhere, so the old stack state is copied into the new cell state. This is summarized in the following diagram, where the dark grey bar indicates where the gate is set to 0:



pop mechanics. To implement a pop operation, the **forget gate** finds the top slot, which is the slot farthest from slot 1 that *isn't* empty (that is, that encodes some $\langle i$.) In practice, we do this by guaranteeing that the forget gate is equal to 1 for all stack slots before (but excluding) the last non-empty stack slot encodes the top of the stack, for it and all subsequent (empty) stack slots, the forget gate is set to 0.¹³ This erases the element at the top of the stack, summarized in the following diagram:



¹²Note that, since $W_{\tilde{c}} = \mathbf{0}$, we have $\tilde{c}_t = \tanh(U_{\tilde{c}}x_t + b_{\tilde{c}})$, so it does not depend on the history h_{t-1} .

¹³The **input gate** and the **output gate** are both set to $\mathbf{0}$.

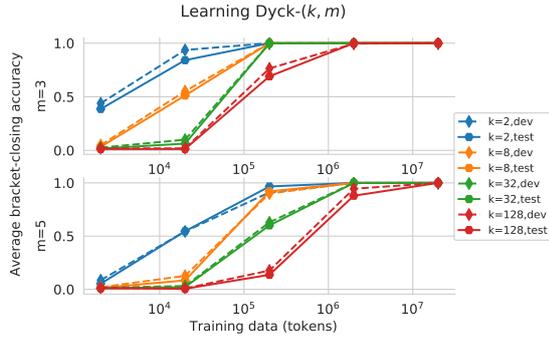


Figure 3: Learning curves for Dyck- (k,m) languages.

strict supersets of other $p^{(j)}$, so the true symbol cannot be decoded through any $V^\top h_t$. To solve this, we use a constant factor more space, to ensure each symbol is decodable by V . In the first $\log k$ units of a stack slot we use the bit configuration $p^{(i)}$. In the second, we use $(1 - p^{(i)})$ (the binary negation.) Call this encoding $\psi_i \in \mathbb{R}^{2 \log k}$. Using ψ_i for the row V_j , we have the following expression for the dot product in determining the probability distribution:

$$V_j^\top \psi_i = \sum_{\ell=1}^{\lceil \log k \rceil} p_\ell^{(i)} p_\ell^{(j)} + \sum_{\ell=1}^{\lceil \log k \rceil} (1 - p_\ell^{(i)})(1 - p_\ell^{(j)})$$

$$\begin{cases} = \lceil \log k \rceil & i = j \\ \leq \lceil \log k \rceil - 1 & i \neq j \end{cases}$$

Thus, we can always detect which symbol i is encoded by setting $b_j = \log k - 0.5$.¹⁶

8 Experiments

Our proofs have concerned constructing RNNs that generate Dyck- (k,m) ; now we present a short study connecting our theory to learning Dyck- (k,m) from finite samples. In particular, for $k \in \{2, 8, 32, 128\}$ and $m \in \{3, 5\}$, we use our theory to set the hidden dimensionality of LSTMs to $3m \lceil \log k \rceil - m$, and train them as LMs on samples from a distribution¹⁷ over Dyck- (k,m) . For space, we provide an overview of the experiments, with details in the Appendix. We evaluate the models' abilities to extrapolate to unseen lengths by setting a maximum length of 84 for $m = 3$, and 180 for $m = 5$, and

¹⁶In actuality, we use a slightly less compact encoding, spending $\log k - 1$ more hidden units set to 1, to incrementally subtract $\log k - 1$ from all logits. Then the bias terms b_j are set to 0.5, avoiding possible precision issues with representing the float $\lceil \log k \rceil - 0.5$.

¹⁷Defined in Appendix H

testing on sequences longer than those seen at training time.¹⁸ For our evaluation metric, let p_j be the probability that the model predicts the correct closing bracket given that j tokens separate it from its open bracket. We report $\text{mean}_j p_j$, to evaluate the model's bracket-closing memory.

For all configurations, we find that the LSTMs using our memory limit achieve error less than 10^{-4} when trained on 20 million tokens. Strikingly, this is despite the fact that for large m and k , a small fraction of the possible stack states is seen at training time;¹⁹ this shows that the LSTMs are not simply learning k^m structureless DFA states. Learning curves are provided in Figure 3.

9 Discussion and conclusion

We proved that finite-precision RNNs can generate Dyck- (k,m) , a canonical family of bounded-depth hierarchical languages, in $O(m \log k)$ memory, a result we also prove is tight. Our constructions provide insight into the mechanisms that RNNs and LSTMs can implement.

The Chomsky hierarchy puts all finite memory languages in the single category of regular languages. But humans generating natural language have finite memory, and context-free languages are known to be both too expressive and not expressive enough (Chomsky, 1959; Joshi et al., 1990). We thus suggest the further study of what *structure* networks can encode in their memory (here, stack-like) as opposed to (just) their position in the Chomsky hierarchy. While we have settled the representation question for Dyck- (k,m) , many open questions still remain: What broader class of bounded hierarchical languages can RNNs efficiently generate? Our experiments point towards learnability; what class of memory-bounded languages are efficiently learnable? We hope that answers to these questions will not just demystify the empirical success of RNNs but ultimately drive new methodological improvements as well.

Reproducibility Code for running our experiments is available at <https://github.com/john-hewitt/dyckkm-learning>. An executable version of the experiments in this paper is on CodaLab at <https://worksheets.codalab.org/worksheets/0xd668cf62e9e0499089626e45affee864>.

¹⁸Up to twice as long as the training maximum.

¹⁹See Table 3.

Acknowledgements The authors would like to thank Nelson Liu, Amita Kamath, Robin Jia, Sidd Karamcheti, and Ben Newman. JH was supported by an NSF Graduate Research Fellowship, under grant number DGE-1656518. Other funding was provided by a PECASE award.

References

- Noam Chomsky. 1956. Three models for the description of language. *IRE Transactions on information theory*, 2(3):113–124.
- Noam Chomsky. 1959. [On certain formal properties of grammars](#). *Information and Control*, 2(2):137 – 167.
- Noam Chomsky and Marcel P Schützenberger. 1959. The algebraic theory of context-free languages. In *Studies in Logic and the Foundations of Mathematics*, volume 26, pages 118–161. Elsevier.
- Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.
- C. Lee Giles, Guo-Zheng Sun, Hsing-Hen Chen, Yee-Chun Lee, and Dong Chen. 1990. [Higher order recurrent networks and grammatical inference](#). In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems 2*, pages 380–387. Morgan-Kaufmann.
- Edward Grefenstette, Karl Moritz Hermann, Mustafa Suleyman, and Phil Blunsom. 2015. Learning to transduce with unbounded memory. In *Advances in neural information processing systems*, pages 1828–1836.
- Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 1195–1205.
- Yiding Hao, William Merrill, Dana Angluin, Robert Frank, Noah Amsel, Andrew Benz, and Simon Mendelsohn. 2018. [Context-free transductions with neural stacks](#). In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 306–315, Brussels, Belgium. Association for Computational Linguistics.
- John Hewitt and Christopher D. Manning. 2019. [A structural probe for finding syntax in word representations](#). In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*. Association for Computational Linguistics.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Bill G Horne and Don R Hush. 1994. Bounds on the complexity of recurrent neural network implementations of finite state machines. In *Advances in neural information processing systems*, pages 359–366.
- Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. 2017. Quantized neural networks: Training neural networks with low precision weights and activations. *The Journal of Machine Learning Research*, 18(1):6869–6898.
- Piotr Indyk. 1995. Optimal simulation of automata by neural nets. In *Annual Symposium on Theoretical Aspects of Computer Science*, pages 337–348. Springer.
- Lifeng Jin, Finale Doshi-Velez, Timothy Miller, William Schuler, and Lane Schwartz. 2018. Depth-bounding is effective: Improvements and evaluation of unsupervised PCFG induction. *arXiv preprint arXiv:1809.03112*.
- Aravind K Joshi, K Vijay Shanker, and David Weir. 1990. The convergence of mildly context-sensitive grammar formalisms. *Technical Reports (CIS)*, page 539.
- Armand Joulin and Tomas Mikolov. 2015. Inferring algorithmic patterns with stack-augmented recurrent nets. In *Advances in neural information processing systems*, pages 190–198.
- Fred Karlsson. 2007. Constraints on multiple center-embedding of clauses. *Journal of Linguistics*, 43(2):365–392.
- Andrej Karpathy, Justin Johnson, and Li Fei-Fei. 2015. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*.
- Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. 2018. [Sharp nearby, fuzzy far away: How neural language models use context](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 284–294, Melbourne, Australia. Association for Computational Linguistics.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Samuel A Korsky and Robert C Berwick. 2019. On the computational power of rnns. *arXiv preprint arXiv:1906.06349*.
- Adhiguna Kuncoro, Chris Dyer, John Hale, Dani Yogatama, Stephen Clark, and Phil Blunsom. 2018. LSTMs can learn syntax-sensitive dependencies well, but modeling structure makes them better. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1426–1436.

- Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. [The emergence of number and syntax units in LSTM language models](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 11–20, Minneapolis, Minnesota. Association for Computational Linguistics.
- Y C Lee, G Doolen, H H Chen, G Z Sun, T Maxwell, H Y Lee, and C L Giles. 1986. Machine learning using a higher order correlation network. *Phys. D*, 2(1–3):276–306.
- Omer Levy, Kenton Lee, Nicholas FitzGerald, and Luke Zettlemoyer. 2018. [Long short-term memory as a dynamically computed element-wise weighted sum](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Melbourne, Australia. Association for Computational Linguistics.
- Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. [Assessing the ability of LSTMs to learn syntax-sensitive dependencies](#). *Transactions of the Association for Computational Linguistics*, 4:521–535.
- Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.
- William Merrill. 2019. [Sequential neural networks as automata](#). In *Proceedings of the Workshop on Deep Learning and Formal Languages: Building Bridges*, pages 1–13, Florence. Association for Computational Linguistics.
- William Merrill. 2020. On the linguistic capacity of real-time counter automata. *ArXiv*, abs/2004.06866.
- William Merrill, Gail Weiss, Yoav Goldberg, Roy Schwartz, Noah A Smith, and Eran Yahav. 2020. [A formal hierarchy of rnn architectures](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Seattle, USA. Association for Computational Linguistics.
- George A. Miller and Noam Chomsky. 1963. Finitary models of language users. In R. Duncan Luce, Robert R. Bush, and Eugene Galanter, editors, *Handbook of mathematical psychology*, pages 419–491. John Wiley and Sons, Inc., New York and London.
- Christian W Omlin and C Lee Giles. 1996. Constructing deterministic finite-state automata in recurrent neural networks. *Journal of the ACM (JACM)*, 43(6):937–972.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035.
- Guillaume Rabusseau, Tianyu Li, and Doina Precup. 2019. Connecting weighted automata and recurrent neural networks through spectral learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1630–1639.
- Philip Resnik. 1992. Left-corner parsing and psychological plausibility. In *Proceedings of the 14th conference on Computational linguistics-Volume 1*, pages 191–197. Association for Computational Linguistics.
- Marten van Schijndel and Tal Linzen. 2018. Modeling garden path effects without explicit hierarchical syntax. In Tim Rogers, Marina Rau, Jerry Zhu, and Chuck Kalish, editors, *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, pages 2600–2605. Cognitive Science Society, Austin, TX.
- Luzi Sennhauser and Robert Berwick. 2018. [Evaluating the ability of LSTMs to learn context-free grammars](#). In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 115–124, Brussels, Belgium. Association for Computational Linguistics.
- Ravid Shwartz-Ziv and Naftali Tishby. 2017. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*.
- Hava T. Siegelmann and Eduardo D. Sontag. 1992. [On the computational power of neural nets](#). In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory, COLT '92*, page 440–449, New York, NY, USA. Association for Computing Machinery.
- Mirac Suzgun, Yonatan Belinkov, Stuart Shieber, and Sebastian Gehrmann. 2019. [LSTM networks can perform dynamic counting](#). In *Proceedings of the Workshop on Deep Learning and Formal Languages: Building Bridges*, pages 44–54, Florence. Association for Computational Linguistics.
- Mirac Suzgun, Yonatan Belinkov, and Stuart M Shieber. 2018. On evaluating the generalization of lstm models in formal languages. *arXiv preprint arXiv:1811.01001*.
- Gail Weiss, Yoav Goldberg, and Eran Yahav. 2018. [On the practical computational power of finite precision rnns for language recognition](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 740–745. Association for Computational Linguistics.
- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus

Macherey, et al. 2016. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Xiang Yu, Ngoc Thang Vu, and Jonas Kuhn. 2019. [Learning the Dyck language with attention-based Seq2Seq models](#). In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 138–146, Florence, Italy. Association for Computational Linguistics.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. [Understanding deep learning requires rethinking generalization](#). In *International Conference on Learning Representations*.

A Appendix outline

This Appendix has the following order. In (§B), we provide a definition of Dyck- (k,m) equivalent to that in the main text but more useful for our proofs. In (§C), we state preliminary definitions and assumptions, and prove the lower-bound of $\Omega(m \log k)$ hidden units to generate Dyck- (k,m) . In (§D), we formally introduce our Simple RNN stack construction, and prove its correctness in a lemma. In (§E), we formally introduce our LSTM stack construction, and prove its correctness in a lemma. In (§F), we prove that a linear (+softmax) decoder on the Simple RNN and LSTM hidden states can be used to generate Dyck- (k,m) in $O(mk)$ hidden units using a 1-hot encoding of stack elements. In this section we also prove that a general RNN construction of DFAs allows for generation of Dyck- (k,m) in $O(k^{m+1})$ hidden units. In (§G), we provide an alternative encoding of elements in our stack constructions for the Simple RNN and LSTM that uses $O(\log k)$ space per element, and prove that our stack constructions still hold using this encoding. We provide a linear (+softmax) decoder on the hidden states of the Simple RNN and LSTM (when using the $O(\log k)$ stack element encoding) that can be used as a drop-in replacement for the decoder from the 1-hot representations, thus generating Dyck- (k,m) . This proving that both the Simple RNN and LSTM generate Dyck- (k,m) in $O(m \log k)$ hidden units.

B The Dyck- (k,m) languages

To better understand the success of neural networks on natural language syntax, we aim for a formal language that models the unbounded recursiveness of natural language while also reflecting its bounded memory requirements. We thus introduce the Dyck- (k,m) languages, corresponding to sequences of balanced brackets of k types with a maximal number of m unclosed brackets at any point in the sequences (yielding a bound stack depth of m to parse such sentences).

Though in the main text we defined Dyck- (k,m) by intersecting Dyck- k with a language that simply bounds the difference between the number of open brackets and the number of close brackets, here we provide an equivalent definition that will aid in our proofs. Each Dyck- (k,m) language, specified by fixing a value of m and k , is defined using a deterministic finite automaton (DFA). Here, we provide a general description of any Dyck- (k,m)

Definition/Proof	Main paper	Appendix
Dyck- (k,m)	Definition 1	Definition 6
Simple RNN generator	Definition 2	Definition 2
LSTM generator	Definition 3	Definition 3
Locally ϵ -truncated support	Definition 4	Definition 4
Fixed-precision setting		Definition 8
Stack correspondence lemma for Simple RNNs		Lemma 3
Stack correspondence lemma for LSTMs		Lemma 4
Simple RNN generates Dyck- (k,m) using $O(k^m)$	Theorem 1	Theorem 1
Simple RNN generates Dyck- (k,m) using $2mk$	Theorem 2	Theorem 2
LSTM generates Dyck- (k,m) using mk	Theorem 3	Theorem 3
Simple RNN generates Dyck- (k,m) in $6m\lceil\log k\rceil - 2m$	Theorem 4	Theorem 4
LSTM generates Dyck- (k,m) in $3m\lceil\log k\rceil - m$	Theorem 5	Theorem 5
Lower bound of $\Omega(m \log k)$	Theorem 6	Theorem 6

Table 1: Correspondence between and hyperlinks for definitions and theorems in the main paper and the same objects in the appendix.

DFA.

Formally, we define each language by the deterministic finite automaton $\mathcal{D}_{m,k} = (Q, \Sigma, \delta, F, q_0)$. The vocabulary Σ consists of k types of open brackets: $\{\langle_i\}_{i=1,\dots,k}$ corresponding closing brackets $\{\rangle_i\}_{i=1,\dots,k}$. Strings over the vocabulary are $w_{1:T} \in \Sigma^*\omega$, where $\omega \notin \Sigma$ is a special symbol representing the end of the sequence. $\Sigma \cup \{\omega\}$ are collectively referred to as symbols.

This slightly nonstandard requirement allows for a natural connection with language models, which must estimate the probability that a string ends at any given token.²⁰ Overloading notation, we'll also use $\mathcal{D}_{m,k}$ to refer to the language (the set of strings) itself, defined as the strings accepted by the DFA.

B.1 DFA States, Q

We now define the states $q \in Q$. First, we define reject state r , and accept state $[\omega]$. Each other state is uniquely identified by a list of open bracket symbols of length up to m ; thus the full set of states is provided by:

$$Q = \{[\omega], r\} \cup \left\{ [\langle_{i_1} \langle_{i_2}, \dots, \langle_{i_{m'}}] \right\}_{m' \in 1 \dots m, i_j \in 1 \dots k}$$

²⁰But crucially, since $\omega \notin \Sigma$, the language model need not define a distribution *after* ω is seen; it can only be the last token.

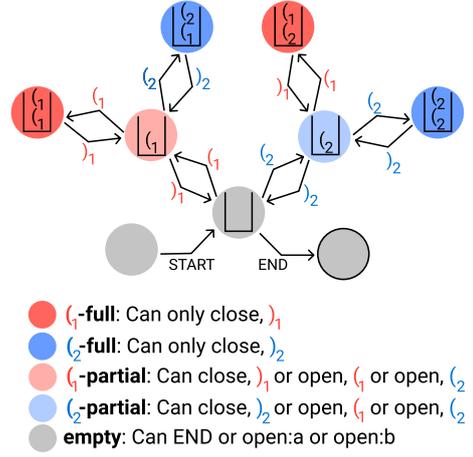


Figure 4: A deterministic finite automaton describing the transitions of 2-bounded Dyck-2 that do not lead to a reject state (omitted for space) and a qualitative description of the classes of states.

The number of states is thus $k^{m+1} + 1$, where all but two states reflect a list of open brackets. We will denote each list $[\langle_{i_1} \langle_{i_2}, \dots, \langle_{i_{m'}}]$, $0 \leq m' \leq m$ as a *stack state* with m' *elements*, and the value of $\langle_{i_{m'}}$ as the top element of the stack. We let $q_0 = [] = [\langle_{i_1} \langle_{i_2}, \dots, \langle_{i_{m'}}]$, $m' = 0$, the empty stack state.

B.2 Transition function, δ

We now define the transition function, δ .

Empty stack state The state $[]$ can transition either to the accept state or to another stack state:

$$\delta([], \omega) = [\omega] \quad (\text{B.1})$$

$$\delta([], \langle_i) = [\langle_i] \quad (\text{B.2})$$

while any other symbol transitions to the reject state.

Partial list states For any state of the form $[\langle_{i_1} \langle_{i_2}, \dots, \langle_{i_{m'}}]$, where $m' < m$, an open bracket can be pushed to the list (since $m' < m$), or the last open bracket can be removed, by observing its corresponding close bracket.

$$\delta([\langle_{i_1} \langle_{i_2}, \dots, \langle_{i_{m'}}], \langle_i) = [\langle_{i_1} \langle_{i_2}, \dots, \langle_{i_{m'}}, \langle_i] \quad (\text{B.3})$$

$$\delta([\langle_{i_1} \langle_{i_2}, \dots, \langle_{i_{m'}}], \rangle_{i_{m'}}) = [\langle_{i_1} \langle_{i_2}, \dots, \langle_{i_{m'-1}}] \quad (\text{B.4})$$

All other symbols transition to the reject state.

Full list states For any state of the form $[\langle i_1 \langle i_2, \dots, \langle i_m \rangle \rangle \rangle]$, that is, the list is of length m and thus full, the close bracket of the top of the stack removes that bracket,

$$\delta([\langle i_1 \langle i_2, \dots, \langle i_m \rangle \rangle \rangle, i_m) = [\langle i_1 \langle i_2, \dots, \langle i_{m-1} \rangle \rangle \rangle] \quad (\text{B.5})$$

All other symbols transition to the reject state.

r The reject state r transitions to itself for all symbols.

$[\omega]$ No transitions from the accept state need be defined, since only $\delta([\], \omega) = [\omega]$, and ω must be the last symbol of any string in the universe, and can only occur once.

This accounts for the transition function from all states in Q , and completes our definition of the DFA $D_{m,k}$. We are now prepared to formally define the language:

Definition 6 (Dyck- (k,m)). For any $k, m \in \mathbb{Z}^+$, The language Dyck- (k,m) is the set of strings accepted by the DFA $D_{m,k}$; overloading notation: $D_{m,k} \subseteq \Sigma^* \omega$.²¹

Overloading notation, for any $w_{1:T} \in \Sigma^* \omega$, for any $t \in 1, \dots, T$, we'll say $q_t = D_{m,k}(w_{1:t})$, denoting the state that $D_{m,k}$ is in after consuming $w_{1:t}$.

C Preliminaries

Definition 4 (locally ϵ -truncated support). Let f_θ be a probability distribution over $\Sigma^* \omega$, with conditional probabilities $f_\theta(w_t | w_{1:t-1})$. Then the locally ϵ -truncated support of the distribution is the set

$$\{w_{1:T} \in \Sigma^* \omega : \forall t \in 1 \dots T, f_\theta(w_t | w_{1:t-1}) \geq \epsilon\}.$$

Through the notion of ϵ -truncated support, a language model specifies which tokens are allowable continuations of each string prefix by assigning them greater than ϵ probability. With this, we're ready to connect RNNs and formal languages:

Definition 7 (generating a language). A probability distribution f_θ generates a formal language \mathcal{L} if there exists $\epsilon > 0$ such that the ϵ -truncated support \mathcal{L}_{f_θ} is equal to \mathcal{L} .

²¹Where \mathbb{Z}^+ denotes the positive integers.

C.1 Technical considerations

Definition 8 (fixed-precision setting). For language parameters m, k of $D_{m,k}$, and input sequence lengths T , we assume that each floating-point value can be specified in $p \in O(1)$ bits (that is, not scaling in any parameter.) These bits choose elements of the set $\mathbb{P} \subset \mathbb{Q}$; we do not specify how \mathbb{P} must be set. The \mathbb{P} chosen for our constructions is described in Appendix G.3, after the relevant constants have been defined.

Under our finite-precision setting, a reasonable assumption about the properties of the sigmoid and hyperbolic tangent functions make our claim considerably simpler. The sigmoid function, $\sigma(x) = \frac{1}{1+e^{-x}}$, has range $(0, 1)$, excluding its boundaries $\{0, 1\}$. However, in a finite-precision arithmetic, $\sigma(x)$ cannot become arbitrarily close to 0 or 1. In fact, in popular deep learning library PyTorch²², $\sigma(x)$ is exactly equal to 1 for all $x > 6$. We define the floating point sigmoid function to equal 0 or 1 if the absolute value of its input is greater or equal in absolute value to some threshold β :

$$\sigma_{\text{fp}}(x) = \begin{cases} \sigma(x) & -\beta < x < \beta \\ 1 & x \geq \beta \\ 0 & x \leq -\beta, \end{cases} \quad (\text{C.1})$$

Similarly for the hyperbolic tangent function, we define:

$$\tanh_{\text{fp}}(x) = \begin{cases} \tanh(x) & -\beta < x < \beta \\ 1 & x \geq \beta \\ -1 & x \leq -\beta, \end{cases} \quad (\text{C.2})$$

For the rest of this paper, we will refer to σ_{fp} as σ , and \tanh_{fp} as \tanh .

C.2 Lower bound of $\Omega(m \log k)$ for generation of Dyck- (k,m)

Theorem 6 ($\Omega(m \log k)$ to generate Dyck- (k,m)). Let A be an arbitrary function from d -dimensional vectors and symbols $w_t \in \Sigma$ to d -dimensional vectors; $A : \mathbb{P}^d \times \Sigma \rightarrow \mathbb{P}^d$, $A(h_{t-1}, w_t) \mapsto h_t$. Let ψ be an arbitrary function from \mathbb{P}^d to probability distributions over $\Sigma \cup \{\omega\}$. Let f_ψ be a probability distribution over $\Sigma^* \omega$, with the form $f_\psi(w_{1:T}) = \prod_{t=1}^T f(w_t | w_{1:t-1})$, where $f(w_t | w_{1:t-1}) = \psi(h_{t-1})$. If f generates Dyck- (k,m) , then $d \geq \frac{m \log k}{p} = \Omega(m \log k)$.

²²Under the `float` datatype, PyTorch v1.3.1 (Paszke et al., 2019).

Proof. We provide a communication complexity argument. Assume for contradiction that there exists such a machine with $d < \frac{m \log k}{p}$. Consider any string $w_{1:m} = \langle i_1, \dots, i_m \rangle$ of m open brackets, where each is one of k types. There are k^m such strings, but because $d < \frac{m \log k}{p}$, we have that the model has $2^{dp} < k^m$ possible representations, so at least two such strings must share the same representation. Let such a pair be $w \neq w'$, where likewise w' is defined as $\langle i'_1, \dots, i'_m \rangle$. Consider the sequence of open bracket indices that defines w : i_1, \dots, i_m . Let s be the string $\rangle_{i_m}, \dots, \rangle_{i_1}$ of close brackets in the reverse order of the indices of w , and likewise s' for w' .

The string $w :: s$, where $::$ denotes concatenation, is in Dyck- (k, m) . However, $w' :: s$ is not in Dyck- (k, m) as it breaks the well-nested brackets condition wherever $\langle i_j \neq i'_j$. This must occur at least once, since $w \neq w'$. Since the model assigns them the same representation even though they must be distinguished to generate Dyck- (k, m) , the model does not generate Dyck- (k, m) .

D Proving Simple RNN stack correspondence in $2mk$ units

In this section, we provide a formal description of our Simple RNN stack construction, and introduce and prove the *stack correspondence lemma*, to guarantee its correctness.

Definition 2 (Simple RNN (generator)). A Simple RNN (generator) with d hidden units is a probability distribution f_θ of the following form:

$$\begin{aligned} h_0 &= \mathbf{0} \\ h_t &= \sigma(W h_{t-1} + U x_t + b) \\ w_t | w_{1:t-1} &\sim \text{softmax}(g_\theta(h_{t-1})) \end{aligned}$$

where $h_t \in \mathbb{R}^d$ and the function g has the form $g_\theta(h_{t-1}) = V h_{t-1} + b_v$. The input $x_t = E w_t$; overloading notation, w_t is the one-hot (indicator) vector representing the respective word. θ is the set of trainable parameters, W, U, b, V, b_v, E .

Encoding a Stack in the Hidden State We define mappings between Q —the DFA state space—and the RNN’s state space \mathbb{R}^{2km} . First, we encode a stack into a km -dimensional vector as follows:

$$\begin{aligned} R([\langle i_1, \langle i_2, \dots, \langle i_{m'} \rangle]) \\ = [e_{i_1}, e_{i_2}, \dots, e_{i_{m'}}, \mathbf{0}_{\in \mathbb{R}^{(m-m')k}}] \end{aligned} \quad (\text{D.1})$$

where $[\dots]$ denotes concatenation of vectors into a km -dimensional vector. We note that R is injective, that is, the inverse map R^{-1} is well-defined on all vectors in the image of R .

Later, when constructing the efficient $O(m \log k)$ memory encoding in Section G.1, we will swap the one-hot vectors e_i for other vectors that will also have entries in $\{0, 1\}$.²³

Based on this, \mathcal{S} maps a pair of a stack operation ($\sigma \in \{\text{push}, \text{pop}\}$) and a DFA state q (assuming $q \notin \{r, [\omega]\}$) to a hidden state in \mathbb{R}^{2km} :

$$\mathcal{S}(\sigma, q) = \begin{cases} [R(q), \mathbf{0}_{\in \mathbb{R}^{km}}] & \sigma = \text{push} \\ [\mathbf{0}_{\in \mathbb{R}^{km}}, R(q)] & \sigma = \text{pop} \end{cases} \quad (\text{D.2})$$

Let $\mathcal{R} := \text{Im}(\mathcal{S}) \subset \mathbb{R}^{2km}$. By writing a hidden state $h \in \mathcal{R}$ as $h = [h_{\text{push}}, h_{\text{pop}}]$, we can obtain the corresponding DFA state by

$$\mathcal{Q}(h) = R^{-1}(h_{\text{push}} + h_{\text{pop}}) \quad (\text{D.3})$$

If $h = \mathcal{S}(\sigma, q)$, then the definition guarantees that $\mathcal{Q}(h) = q$.

Defining Transition Matrices To define the RNN (apart from V, b_v), we have to specify the matrices W, U, E , and the vector b .

We start by defining $W \in \mathbb{R}^{2km \times 2km}$. Let $M_{\text{pop}}, M_{\text{push}} \in \mathbb{R}^{km \times km}$ be the matrices where

$$(M_{\text{pop}})_{i,j} = \begin{cases} 2\beta & i = j + k \\ 0 & \text{o.w.} \end{cases} \quad (\text{D.4})$$

$$(M_{\text{push}})_{i,j} = \begin{cases} 2\beta & i = j - k \\ 0 & \text{o.w.} \end{cases} \quad (\text{D.5})$$

Let I_{km} be the $km \times km$ identity matrix. Take $W \in \mathbb{R}^{2km \times 2km}$ defined as

$$W = \begin{pmatrix} M_{\text{push}} \\ M_{\text{pop}} \end{pmatrix} \cdot (I_{km} \quad I_{km}) \quad (\text{D.6})$$

To specify the matrices E and U , we need to fix an assignment to the integers $1, \dots, 2k$ to the symbols in Σ . We will assign the integers $1, \dots, k$ to the opening brackets (i.e., $\langle_1 \mapsto 1, \dots, \langle_k \mapsto k$), and the integers $1, \dots, k$ to the closing brackets (i.e., $\rangle_1 \mapsto k + 1, \dots, \rangle_k \mapsto 2k$).

²³Looking ahead to the more complex LSTM construction, we’ll introduce notation to denote these vectors $e_i = \psi^{-1}(\langle_i)$, and look to replace ψ when we develop the $O(m \log k)$ construction.

Define $E := I_{2k \times 2k}$, and define $U \in \mathbb{R}^{2km \times 2k}$ as:

$$\begin{pmatrix} 2\beta I_{k \times k} & -2\beta \cdot \mathbf{1}_{k \times k} \\ \mathbf{0}_{(m-1)k \times k} & -2\beta \cdot \mathbf{1}_{(m-1)k \times k} \\ -2\beta \cdot \mathbf{1}_{k \times k} & \mathbf{0}_{k \times k} \\ -2\beta \cdot \mathbf{1}_{(m-1)k \times k} & \mathbf{0}_{(m-1)k \times k} \end{pmatrix} \quad (\text{D.7})$$

Here, we write $I_{k \times k}$ for the identity matrix, and $\mathbf{1}_{k \times k}$ for a matrix filled entirely with ones.

Finally, we define

$$b := -\beta \cdot \mathbf{1}_{\mathbb{R}^{2km}} \quad (\text{D.8})$$

Proof of Correctness To prove correctness of the construction, the first step is to show that the transition dynamics of the Simple RNN correctly simulates the dynamics of the stack when consuming one symbol.

Let $\phi(\langle i) = \text{push}$, $\phi(\rangle i) = \text{pop}$.

Lemma 1 (One-Step Lemma). *Assume that the hidden state h_t encodes a stack q :*

$$\mathcal{Q}(h_t) = q \quad (\text{D.9})$$

Let w_{t+1} be the new input symbol, and assume that $w_{1:t+1}$ is a valid prefix of a word in Dyck- (k, m) . Let h_{t+1} be the next hidden state, after reading w_{t+1} . Then

$$\mathcal{S}(\phi(w_{t+1}), \delta(q, w_{t+1})) = h_{t+1} \quad (\text{D.10})$$

Before showing this lemma, we note the following property of the activation function σ , under the finite precision assumption:

Lemma 2. *If $u \in \{0, 1\}$, then*

$$\sigma(\beta \cdot (2u - 1)) = u \quad (\text{D.11})$$

$$\sigma(\beta \cdot (2u - 3)) = 0 \quad (\text{D.12})$$

Proof. By calculation. \square

Proof of the Lemma. We can write $h_t = [h_{\text{push}}, h_{\text{pop}}]$. Set $h' = h_{\text{push}} + h_{\text{pop}}$. By construction of \mathcal{S} ,

$$h' = R(q) \quad (\text{D.13})$$

We will write $h' = [h'_1, \dots, h'_m]$, where $h'_i \in \mathbb{R}^k$.

At this point, we note that the only relevant property of the encodings h'_i for this proof is that their entries are in $\{0, 1\}$, not that they are one-hot vectors. This will make it possible to plug in a more efficient $O(\log k)$ encoding later in Section G.1.

With this, we can write

$$Wh_t = \begin{pmatrix} M_{\text{push}} h' \\ M_{\text{pop}} h' \end{pmatrix} = \begin{pmatrix} 2\beta[\mathbf{0}_{\in \mathbb{R}^{km}}, h'_1, \dots, h'_{m-1}] \\ 2\beta[h'_2, \dots, h'_m, \mathbf{0}_{\in \mathbb{R}^{km}}] \end{pmatrix} \quad (\text{D.14})$$

Case 1: Assume $w_{t+1} = \langle i$. We have

$$Ux_t = [2\beta e_i, \mathbf{0}_{(m-1)k}, -2\beta \mathbf{1}_{mk}] \quad (\text{D.15})$$

so that $Wh_{t-1} + Ux_t + b$ equals

$$\begin{pmatrix} 2\beta[e_i, h'_1, \dots, h'_{m-1}] - \beta \mathbf{1} \\ 2\beta[h'_2, \dots, h'_{m-1}, h'_m, \mathbf{0}] - 3\beta \mathbf{1} \end{pmatrix} \quad (\text{D.16})$$

Then, by Lemma 2, h_{t+1} equals

$$\begin{pmatrix} [e_i, h'_1, \dots, h'_{m-1}] \\ \mathbf{0}_{km} \end{pmatrix} \quad (\text{D.17})$$

which is equal to $\mathcal{S}(\text{push}, \delta(q, \langle i))$.

Case 2: Assume $w_{t+1} = \rangle i$. We have

$$Ux_t = [-2\beta \mathbf{1}_{mk}, \mathbf{0}_{mk}] \quad (\text{D.18})$$

so that $Wh_{t-1} + Ux_t + b$ equals

$$\begin{pmatrix} 2\beta[\mathbf{0}, h'_1, \dots, h'_{m-1}] - 3\beta \mathbf{1} \\ 2\beta[h'_2, \dots, h'_{m-1}, h'_m, \mathbf{0}] - \beta \mathbf{1} \end{pmatrix} \quad (\text{D.19})$$

Then, h_{t+1} equals

$$\begin{pmatrix} \mathbf{0}_{km} \\ [h'_2, \dots, h'_m, \mathbf{0}] \end{pmatrix} \quad (\text{D.20})$$

which is equal to $\mathcal{S}(\text{push}, \delta(q, \rangle i))$. \square

From the previous lemma, we can derive the following Stack Correspondence lemma, which asserts that the Simple RNN correctly reflects stack dynamics over an entire input string:

Lemma 3 (Stack Correspondence). *For all strings $w_{1:T}$ in Dyck- (k, m) , for all $t = 1, \dots, T$, let q_t be the state of $D_{m,k}$ after consuming prefix $w_{1:t}$. Let h_t be the RNN hidden state after consuming $w_{1:t}$. Then if $q_t \notin \{[\omega]\}$,*

$$\mathcal{Q}(h_t) = q_t, \quad (\text{D.21})$$

Proof. The claim is shown by induction over t , applying the previous lemma in each step. We will show the claim for all $t = 0, 1, \dots, T$, setting h_0 to be the zero vector $\mathbf{0}_{2km}$.

As $\mathcal{Q}(h_0) = q_0$, this proves the claim for $t = 0$. To prove the inductive step, we assume $\mathcal{Q}(h_t) = q_t$ has already been shown.

Then, by the preceding One-Step Lemma,

$$\mathcal{S}(\phi(w_{t+1}), \delta(q_t, w_{t+1})) = h_{t+1} \quad (\text{D.22})$$

Noting $q_{t+1} = \delta(q_t, w_{t+1})$, we obtain

$$\mathcal{S}(\phi(w_{t+1}), q_{t+1}) = h_{t+1} \quad (\text{D.23})$$

As noted in the definition of \mathcal{Q} above (D.3), this entails

$$\mathcal{Q}(h_{t+1}) = q_{t+1} \quad (\text{D.24})$$

concluding the inductive step. \square

E Proving LSTM stack correspondence in mk units

In this section, we prove our lemma describing the stack construction implemented by an LSTM. To do so, we first define the LSTM:

Definition 3 (LSTM (generator)). An LSTM (generator) with d hidden units is a probability distribution f_θ of the following form:

$$\begin{aligned} h_0, c_0 &= \mathbf{0} \\ f_t &= \sigma(W_f h_{t-1} + U_f x_t + b_f) \\ i_t &= \sigma(W_i h_{t-1} + U_i x_t + b_i) \\ o_t &= \sigma(W_o h_{t-1} + U_o x_t + b_o) \\ \tilde{c}_t &= \tanh(W_{\tilde{c}} h_{t-1} + U_{\tilde{c}} x_t + b_{\tilde{c}}) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\ h_t &= o_t \odot \tanh(c_t) \\ w_t | w_{1:t-1} &\sim \text{softmax}(g_\theta(h_{t-1})) \end{aligned}$$

where $h_t, c_t \in \mathbb{R}^d$, the function g has the form $g_\theta(h_{t-1}) = V h_{t-1} + b$, and $x_t = E w_t$, where w_t is overloaded as above. θ is the set of trainable parameters: all W, U, b , as well as V, E .

First, we provide notation for describing the structure of the memory cell (c_t) that f_θ will maintain. Next, we proceed to describe precisely, but without referring to the LSTM equations, how this memory changes over time. We then describe how these high-level dynamics are implemented by the LSTM equations, assuming that we're able to set the values of the intermediate values (gates and new cell candidate) as desired. We then explicitly construct LSTM weight matrices that provide the desired intermediate values. This completes the proof of the memory dynamics, which we formalize in a *stack correspondence lemma* once we've introduced the notation.

E.1 Description of stack-structured memory

Our LSTM construction operates by constructing, in its cell state, a representation of the list that defines the DFA stack state. To make this idea formal, we define a mapping from \mathbb{R}^d , the space of the LSTM cell state, to Q , the DFA state space.

To do so, we need some language for discussing the LSTM cell states. For a cell state $c_t \in \mathbb{R}^{mk}$, we define a list of m stack slots, $s_{t,1}, \dots, s_{t,m}$, where $s_{t,j} \in \mathbb{R}^k$ is given by dimensions

$$s_{t,j} = c_t[k(j-1) + 1 : kj + 1] \quad (\text{E.1})$$

Intuitively, each stack slot will correspond to one item in the list defining a stack state in the DFA $D_{m,k}$, or be empty. We formalize this by defining a function from stack slots $s_{t,j} \in \mathbb{R}^d$ to elements:

$$\psi(s_{t,j}) = \begin{cases} \langle i & s_{t,j} = e_i \\ \emptyset & s_{t,j} = \mathbf{0} \end{cases} \quad (\text{E.2})$$

where $e_i \in \mathbb{R}^k$ is the i^{th} standard basis vector (one-hot vector) of \mathbb{R}^k . We'll also use the inverse map, ψ^{-1} , to map from elements to vectors in \mathbb{R}^k .

Applying this function to each stack slot allows us to map from $s_{t,1}, \dots, s_{t,m}$ to stack states:

$$\mathcal{Q}(s_{t,1}, \dots, s_{t,m}) = [\psi(s_{t,1}), \dots, \psi(s_{t,m'})] \quad (\text{E.3})$$

where $m' \leq m$ is the maximal integer such that $\psi(s_{t,m'}) \neq \emptyset$. Intuitively, this means we filter out any empty stack slots at the end of the sequence (those slots $s_{t,j}$ where $j > m'$.) We'll also make use of the inverse, \mathcal{Q}^{-1} , mapping DFA stack states to stack slot lists,

$$\begin{aligned} \mathcal{Q}^{-1}([\langle i_1, \dots, \langle i_{m'}] & \\ &= \mathcal{Q}^{-1}([\langle i_1, \dots, \langle i_{m'}, \emptyset, \dots, \emptyset]) \\ &= \psi^{-1}(\langle i_1), \dots, \psi^{-1}(\langle i_{m'}), \mathbf{0}, \dots, \mathbf{0} \\ &= e_{i_1}, \dots, e_{i_{m'}}, \mathbf{0}, \dots, \mathbf{0} \end{aligned}$$

where slots $s_{m'+1}, \dots, s_m$ are the zero vector $\mathbf{0}$ because only the first m' slots encode a symbol $\langle i$, and $\psi(\mathbf{0}) = \emptyset$.

On the ψ function These one-hot encodings (e_i) of symbols ($\langle i$) allow for clear exposition, but are not fundamental to the construction; in (§ G), we replace these k -dimensional one-hot encodings given by ψ with $O(\log k)$ -dimensional encodings. Throughout our description of the stack construction, we'll explicitly rely on the following property, so we can easily replace this particular ψ in (§ G) by providing another ψ' with the same property.

Definition 9 (encoding-validity property). Let ψ' be a function $\psi' : \{\langle i\}_{i \in [k]} \rightarrow \mathbb{R}^n$ mapping open brackets to vectors in \mathbb{R}^n . Then ψ' obeys the encoding-validity property if for all $i = 1, \dots, k$,

$$\sum_{\ell=1}^n \psi'^{-1}(\langle i)_\ell = 1 \quad (\text{E.4})$$

that is, the sum of all dimensions of the representation is equal to 1, and

$$\forall_{i \in [k]} \psi'^{-1}(\langle i) \in \{0, 1\}^n \quad (\text{E.5})$$

that is, the representation takes on values only in 0 and 1, and

$$\psi(\mathbf{0}) = \emptyset, \quad (\text{E.6})$$

that is, the empty stack slot is encoded by the zero vector, and

$$\forall_{i,j \in [k], i \neq j} \implies \psi^{-1}(\langle i \rangle) \neq \psi^{-1}(\langle j \rangle) \quad (\text{E.7})$$

$$\forall_{i \in [k], \psi^{-1}(\langle i \rangle) \neq \mathbf{0}} \quad (\text{E.8})$$

that is, encodings of symbols $\langle i \rangle$ are unique, and none are equal to the encoding of the empty symbol \emptyset .

The encoding we've so far provided, $\psi^{-1}(e_i) = \langle i \rangle$, obeys these properties: the sum of a single one-hot vector is 1, no two such vectors are the same, one-hot vectors only take on values in $\{0, 1\}$, and we let $\psi(\mathbf{0}) = \emptyset$.

E.2 Description of stack state dynamics

Before we describe the memory dynamics, we introduce a useful property of the stack slots.

Definition 10 (*j*-top). A cell state c_t composed of stack slots $s_{t,1}, \dots, s_{t,m}$ is *j*-top if there exists $i \in \{1, \dots, k\}$ such that $s_{t,j} = \psi^{-1}(\langle i \rangle)$, and for all $j' \in \{j+1, \dots, m\}$, $s_{t,j'} = \mathbf{0}$.

Intuitively, we'll enforce the constraint that a cell state that is *j*-top **encodes the element $\langle m' \rangle$ at the top of the stack $[\langle i, \dots, \langle m' \rangle]$ in stack slot $j = m'$** .

We're now ready to describe the memory dynamics of f_θ . The memory c_t is initialized to $\mathbf{0}$. So for all $j \in \{1, \dots, m\}$,

$$s_{0,j} = \mathbf{0} \quad (\text{E.9})$$

The recurrent equation of stack slots is then

$$s_{j,t} = \begin{cases} \mathbf{0} & c_{t-1} \text{ is } j\text{-top, and } w_t = \rangle_i \\ \psi^{-1}(\langle i \rangle) & j = 1, c_{t-1} = \mathbf{0}, \text{ and } w_t = \langle i \\ \psi^{-1}(\langle i \rangle) & c_{t-1} \text{ is } (j-1)\text{-top and } w_t = \langle i \\ s_{t-1,j} & o.w. \end{cases} \quad (\text{E.10})$$

Intuitively, the first case of Equation E.10 implements pop, the second case implements push (if pushing to an otherwise empty stack) and the third case implement push (if pushing to a non-empty stack). The fourth case specifies that slots that are not pushed to or popped are maintained from timestep $t-1$ to timestep t .

So far, our definitions of the LSTM memory have concerned the cell state, c_t . However, the values of the gates f_t, i_t, o_t , as well as the new cell candidate \tilde{c}_t , are functions not of the previous cell state but the previous hidden state, $h_{t-1} = o_{t-1} \odot \tanh(c_{t-1})$. The dynamics of h_t under f_θ we specify as follows, using the same notation $h_{t,j}$ to refer to the dimensions of h_t that correspond to stack slot $s_{t,j}$ of c_t :

$$h_{t,j} = \begin{cases} \tanh(\psi^{-1}(\langle i \rangle)) & c_t \text{ is } j\text{-top} \\ & \text{and } s_{t,j} = \psi^{-1}(\langle i \rangle) \\ \mathbf{0} \in \mathbb{R}^k & o.w., \end{cases} \quad (\text{E.11})$$

Intuitively, this means **only the top of the stack is stored in the hidden state**, and it's stored in whichever units correspond to the *j*-top slot in c_t . Since $\psi^{-1}(\langle i \rangle)$ takes on values in $\{0, 1\}$, $\tanh(\psi^{-1}(\langle i \rangle))$ takes on values in $\{0, \tanh(1)\}$.

We're now ready to formalize the first part of our proof in a lemma that encompasses how f_θ structures and manipulates its memory.

Lemma 4 (stack correspondence). *For all strings $w_{1:T}$ in Dyck- (k, m) , for all $t = 1, \dots, T$, let q_t be the state of $D_{m,k}$ after consuming prefix $w_{1:t}$. Let c_t be the hidden state of the LSTM after consuming $w_{1:t}$. Then if $q_t = \notin \{[\omega]\}$,*

$$\mathcal{Q}(s_{t,1}, \dots, s_{t,m}) = q_t, \quad (\text{E.12})$$

and letting $q_t = [\langle i_1, \dots, \langle i_{m'} \rangle]$ for $m' \leq m$ without loss of generality,

$$h_{t,j} = \begin{cases} \tanh(\psi^{-1}(\langle i_{m'} \rangle)) & j = m' \\ \mathbf{0} & o.w. \end{cases} \quad (\text{E.13})$$

where if $q_t = []$, then $m' = 0$ and $h_{t,j} = \mathbf{0}$ for all j .

Note again that the value of ψ^{-1} is not specified by the lemma; whereas so far we've used $\psi^{-1}(\langle i \rangle) = e_i$, a one-hot encoding, in (§ G) we'll replace ψ^{-1} while maintaining the lemma.

Intuitively, Equation E.12 states that f_θ keeps an exact representation of the stack of unclosed open brackets so far, a statement solely concerning the cell state c_t . However, all intermediate values in the LSTM equations are functions of the hidden state h_t , not the cell state directly; Equation E.13 states how the top of the stack is represented in the hidden state, h_t . We state these as one lemma since it is convenient to prove them by induction together.

E.3 Proof of the stack correspondence lemma assuming stack slot dynamics

We first prove Lemma 4 under the assumption that the stack slot dynamics given in Equation E.10 hold. We proceed by induction on the prefix length.

When the prefix length is 0, the DFA is in state \square , and $s_{0,j} = \mathbf{0}$ for all j , so $\mathcal{Q}(s_{0,1}, \dots, s_{0,m}) = \square$, as required. Further, we have $h_0 = \mathbf{0}$, as required. This completes the base case.

Assume for the sake of induction that Lemma 4 holds for strings up to length t . Now consider a prefix $w_{1:t+1}$. From it, we can take the prefix $w_{1:t}$. Running f_θ , we have $s_{t,1}, \dots, s_{t,m}$. By the induction hypothesis, we have that $\mathcal{Q}(s_{t,1}, \dots, s_{t,m}) = q_t = [\langle i_1, \dots, \langle i_{m'} \rangle]$, (Equation E.12) as well as that

$$h_{t,j} = \begin{cases} \tanh(\psi^{-1}(\langle i_{m'} \rangle)) & j = m' \\ \mathbf{0} & o.w. \end{cases}$$

by Equation E.13. Now, the symbol w_{t+1} can be one of $2k + 1$ symbols: any of the k open brackets $\langle i$, the k close brackets $\rangle i$, or ω . The proof now proceeds by each of these three cases.

Case $w_{t+1} = \langle i$: Without loss of generality, we let $q_t = [\langle i_1, \dots, \langle i_{m'} \rangle]$ for $m' < m$. Strict inequality is guaranteed since if $m' = m$, then $\delta(q_t, \langle i) = r$, meaning $w_{1:t+1}$ cannot be a prefix of any string in Dyck- (k, m) , contradicting a premise of the lemma. Through the inverse of the stack mapping Equation E.3, we have

$$\begin{aligned} \mathcal{Q}^{-1}(q_t) &= s_{t,1}, \dots, s_{t,m} \\ &= \psi^{-1}(\langle i_1 \rangle), \dots, \psi^{-1}(\langle i_{m'} \rangle), \mathbf{0}_{m'+1}, \dots, \mathbf{0}_m \\ &= e_{i_1}, \dots, e_{i_{m'}}, \mathbf{0}_{m'+1}, \dots, \mathbf{0}_m, \end{aligned}$$

the stack slots of f_θ at timestep t . Since $s_{t,m'} = \psi^{-1}(\langle i_j \rangle)$ and $s_{t,m'+1} = \mathbf{0}$, we have that c_t is m' -top. Thus, since $w_{t+1} = \langle i$, the second or third condition in Equation E.10 (depending on whether $m' = 0$) dictate that $s_{t+1,m'+1} = \psi^{-1}(\langle i \rangle)$ (where the i is because $w_{t+1} = \langle i$). All other stack slots fall under the condition $s_{t+1,j} = s_{t,j}$.

With this, we can reason about the DFA state encoded by c_{t+1} at timestep $w_{1:t+1}$:

$$\begin{aligned} \mathcal{Q}(s_{t+1,1}, \dots, s_{t+1,m}) &= \mathcal{Q}(\psi^{-1}(\langle i_1 \rangle), \dots, \psi^{-1}(\langle i_{m'} \rangle), \psi^{-1}(\langle i \rangle), \mathbf{0}, \dots, \mathbf{0}) \\ &= [\langle i_1, \dots, \langle i_{m'} \rangle, \langle i \rangle] \\ &= \delta(q_t, \langle i) \end{aligned}$$

which is the DFA's state at timestep $t + 1$, as required.

Finally, we reason about $h_{t+1,j}$. We've just shown that that c_{t+1} is $(m' + 1)$ -top, and that $s_{t+1,m'+1} = \psi^{-1}(\langle i \rangle)$. By Equation E.11, we have that $h_{t+1,m'+1} = \tanh(\psi^{-1}(\langle i \rangle))$, and $h_{t+1,j} = \mathbf{0}$ for all $j \neq m' + 1$ as required, completing this case.

Case $w_{t+1} = \rangle i$: We have $q_t = [\langle i_1, \dots, \langle i_{m'} \rangle]$ for $0 < m'$. this inequality is strict since if $m' = 0$, then $q_t = \square$, and $\delta(q_t, \rangle i) = r$, meaning $w_{1:t+1}$ is not a prefix of any string in $D_{m,k}$. Thus, we have that c_t is m' -top. Since $w_{t+1} = \rangle i$, we have that $s_{t+1,m'} = \mathbf{0}$, and $s_{t+1,j} = s_{t,j}$ for all $j \neq m'$. Thus, we have

$$s_{t+1,1}, \dots, s_{t+1,m} = \psi^{-1}(\langle i_1 \rangle), \dots, \psi^{-1}(\langle i_{m'-1} \rangle), \mathbf{0}, \dots, \mathbf{0}$$

and thus the DFA state corresponding to the stack slots is:

$$\begin{aligned} \mathcal{Q}(s_{t+1,1}, \dots, s_{t+1,m}) &= [\psi^{-1}(s_{t+1,1}), \dots, \psi^{-1}(s_{t+1,m'-1}), \mathbf{0}, \dots, \mathbf{0}] \\ &= [\langle i_1, \dots, \langle i_{m'-1} \rangle] \end{aligned}$$

Which is $\delta(q_t, \rangle i)$, as required. By the same reasoning as for case $w_{t+1} = \langle i$, since c_{t+1} is $(m' - 1)$ -top and that $s_{t+1,m'-1} = \psi^{-1}(\langle i_{m'-1} \rangle)$, we have that $h_{t+1,m'-1} = \tanh(\psi^{-1}(\langle i \rangle))$ and $h_{t+1,j} = \mathbf{0}$ for all $j \neq m' - 1$, as required. This completes the case.

Case $w_{t+1} = \omega$: If $w_{t+1} = \omega$, then by the definition of δ , $\delta(q_t, \omega) \in \{[\omega], r\}$; hence the premises of Lemma 4 don't hold, so the lemma vacuously holds.

Summary In this section we've proved the stack correspondence lemma for our LSTM construction, assuming we can implement the dynamics of the model as specified in Equations E.10 E.11; we have yet to show that the LSTM update can implement the equations as promised.

E.4 Implementation of stack state dynamics in LSTM equations assuming intermediate values

In this subsection, we define how the dynamics defined in Equations E.10, E.11 are implemented in the LSTM equations, referring to intermediate gate values, assuming such values can be reached with some setting of LSTM parameters.

As we discussed stack slots $s_{t,1}, \dots, s_{t,m}$ of cell state c_t , where each $s_{t,j} \in \mathbb{R}^k$, we similarly discuss the gates in terms of slots, where e.g., $f_{t,j} \in [0, 1]^k$ are the k elementwise gates that apply to the values $s_{t,j}$ (and similarly for $i_{t,j}$, $o_{t,j}$, and the new cell candidate values $\tilde{c}_{t,j} \in [-1, 1]^k$.) With this, we re-write Equation E.10 using the definition of c_t :

$$s_{t,j} = f_{t,j} \odot s_{t-1,j} + i_{t,j} \odot \psi^{-1}(w_t), \quad (\text{E.14})$$

as well as Equation E.11:

$$h_{t,j} = o_{t,j} \odot \tanh(s_{t,j}) \quad (\text{E.15})$$

How these fulfill Equations E.10, E.11? We can set the gate values to do so as follows. First, the forget gate is used to detect the condition under which slot j should be erased:

$$f_{t,j} = \begin{cases} \mathbf{0} \in \mathbb{R}^k & c_{t-1} \text{ is } j\text{-top and } w_t = \rangle_i \\ \mathbf{1} \in \mathbb{R}^k & o.w. \end{cases} \quad (\text{E.16})$$

Second, the new cell candidate expression is used to give the *option* of writing a new open bracket, $\psi^{-1}(w_t)$ to any stack slot:

$$\tilde{c}_{t,j} = \begin{cases} \psi^{-1}(\langle_i) & w_t = \langle_i \\ \mathbf{0} \in \mathbb{R}^k & o.w. \end{cases} \quad (\text{E.17})$$

Third, the input gate determines which stack slot is the first non-empty slot in the sequence, and only allows the information of the new open bracket to be written to that slot:

$$i_{t,j} = \begin{cases} \mathbf{1} & c_{t-1} = \mathbf{0}, j = 1, \text{ and } w_t = \langle_i \\ \mathbf{1} & c_{t-1} \text{ is } (j-1)\text{-top, and } w_t = \langle_i \\ \mathbf{0} \in \mathbb{R}^k & o.w. \end{cases} \quad (\text{E.18})$$

Finally, the output gate is set to identify the top of the stack and not let any other non- $\mathbf{0}$ (i.e., all stack slots below the top) through:

$$o_{t,j} = \begin{cases} \mathbf{0} & c_t \text{ is } j'\text{-top for } j' > j \\ \mathbf{1} \in \mathbb{R}^k & o.w. \end{cases} \quad (\text{E.19})$$

Many of these conditions refer to values, like c_t and c_{t-1} , not available when computing the gates (as only h_{t-1} and x_t are available); it is convenient to refer to these conditions and later show how they are implemented with the available values.

E.5 Proof of stack state dynamics given gate and new cell candidate values

In this section, we prove that Equations E.10, E.11 are implemented by the LSTM f_θ given the gate and new cell candidate values defined in the previous subsection.

We start with Equation E.10, the definition of the stack slot dynamics, proceeding by each of the four cases for defining $s_{t,j}$ in Equation E.10.

Case 1 (popping $s_{t-1,j}$) In this case, the condition is that c_{t-1} is j -top, and $w_t = \rangle_i$. By Equation E.16, we have the forget gate value $f_{j,t} = \mathbf{0}$. By Equation E.18, we have the input gate value $i_{j,t} = \mathbf{0}$. Finally, by Equation E.17, we have $\tilde{c}_{t,j} = \mathbf{0}$. Plugging these values into the LSTM cell expression in Equation E.14, we get

$$\begin{aligned} s_{t,j} &= \mathbf{0} \cdot s_{t-1,j} + \mathbf{0} \cdot \mathbf{0} \\ &= \mathbf{0}, \end{aligned}$$

as required.

Case 2 (pushing to $s_{t,1}$) In this case, the condition is that $c_{t-1} = \mathbf{0}$, that is, the stack is empty, $j = 1$, that is, we'll write to the first stack slot, and $w_t = \langle_i$. By Equation E.18, we have $i_{t,1} = \mathbf{1}$. By Equation E.17, we have $\tilde{c}_{t,j} = \psi^{-1}(\langle_i)$. Further, we have $s_{t-1,j} = \mathbf{0}$ for all j , since $c_{t-1} = \mathbf{0}$. Plugging these values into the LSTM cell expression in Equation E.14, we get:

$$\begin{aligned} s_{t,1} &= f_{1,j} \cdot \mathbf{0} + \mathbf{1} \cdot \psi^{-1}(\langle_i) \\ &= \psi^{-1}(\langle_i), \end{aligned}$$

as required.

Case 3 (pushing to $s_{t,j>1}$) In this case, the condition is that c_{t-1} is $(j-1)$ -top or $c_{t-1} = \mathbf{0}$, and $w_t = \langle_i$. By Equation E.16, we have the forget gate value $f_{j,t} = \mathbf{1}$, since Case 1 does not hold. By Equation E.18, we have the input gate value $i_{j,t} = \mathbf{1}$. Finally, by Equation E.17, we have $\tilde{c}_{t,j} = \psi^{-1}(\langle_i)$. Plugging these values into the LSTM cell expression in Equation E.14, we get

$$\begin{aligned} s_{t,j} &= \mathbf{1} \cdot s_{t-1,j} + \mathbf{1} \cdot \psi^{-1}(\langle_i) \\ &= s_{t-1,j} + \psi^{-1}(\langle_i) \\ &= \psi^{-1}(\langle_i), \end{aligned}$$

where the last equality hold because of the condition that c_{t-1} is $(j-1)$ -top, which implies $s_{t-1,j} = \mathbf{0}$.

Case 3 (maintaining $s_{t-1,j}$) This case catches all conditions under which Cases 1 and 2 do not hold. By Equation E.16, we have the forget gate value $f_{j,t} = 1$, since Case 1 does not hold. By Equation E.18, we have the input gate value $i_{j,t} = 0$, since Case 2 does not hold. Finally, by Equation E.17, we have $\tilde{c}_{t,j} = 0$. Plugging these values into the LSTM cell expression in Equation E.14, we get

$$\begin{aligned} s_{t,j} &= \mathbf{1} \cdot s_{t-1,j} + \mathbf{0} \cdot \mathbf{0} \\ &= s_{t-1,j}, \end{aligned}$$

as required.

Proving Equation E.11 is implemented Next, we prove that Equation E.11 is implemented by the LSTM f_θ , given that Equation E.10 is. First, we must show $h_{t,j}$ is equal to $\tanh(e_i)$ if c_t is j -top. If c_t is j -top, it must be that $s_{t,j} = e_i$ for some i . Since c_t is j -top, we have $o_{t,j} = 1$, and $s_{t,j} = e_i$. Plugging these values into the hidden state expression in Equation E.15, we get:

$$\begin{aligned} h_{t,j} &= o_{t,j} \odot \tanh(s_{t,j}) \\ &= \mathbf{1} \odot \tanh(e_i) \\ &= \tanh(e_i), \end{aligned}$$

as required.

If that condition does not hold, we must have that c_t not j -top. thus, c_t is either j -top for some $j' < j$ or $j' > j$, or $c_t = 0$.

If c_t is j' -top for $j' < j$, then we have that $s_{t,j} = 0$, by the definition of j -top. In this case, we have

$$\begin{aligned} h_{t,j} &= o_{t,j} \odot \tanh(s_{t,j}) \\ &= \mathbf{1} \odot \tanh(\mathbf{0}) \\ &= \mathbf{0}, \end{aligned}$$

as required.

If c_t is j' -top for $j' > j$, then we have that $o_{t,j} = 0$, by Equation E.19. In this case, we have

$$\begin{aligned} h_{t,j} &= o_{t,j} \odot \tanh(s_{t,j}) \\ &= \mathbf{0} \odot s_{t,j} \\ &= \mathbf{0}, \end{aligned}$$

as required.

Finally, if $c_t = 0$, meaning it is not j -top for any j , then

$$\begin{aligned} h_{t,j} &= o_{t,j} \odot \tanh(\mathbf{0}) \\ &= \mathbf{0}, \end{aligned}$$

as required.

Summary So far, we've proved the stack correspondence lemma's induction step for LSTMs assuming that we can provide parameters of the network such that the values assumed in Equations E.16, E.17, E.18, E.19, that is, the gates and new cell candidate, are achieved. We have yet to provide the settings of parameters to do so.

E.6 Construction and proof of LSTM parameters providing gate and new cell candidate values

We now come to the final portion of the proof of the inductive step in Lemma 4. So far, we've proven the induction step conditioned on the existence of parameters that implement the intermediate gate values $f_{t,j}, i_{t,j}, \tilde{c}_{t,j}, o_{t,j}$. Now we construct said parameters, and prove their correctness assuming the induction hypothesis. We'll describe all matrices W , that is, $(W_i, W_f, W_0, W_{\tilde{c}})$, as block matrices, with the block structure specified by the stack slot structure. Precisely, we'll refer to $W_{j,j'}$ as rows $(j-1)k+1$ through $(j-1)k+k$, intersected with columns $(j'-1)k+1$ through $(j'-1)k+k$. This can be interpreted as the block that specifies how the values of slot j' affect the intermediate values constructed for slot j . We'll describe all matrices U as block-matrices through rows alone, that is, U_j referring to rows $(j-1)k'+1$ through $(j-1)k+k$. We'll describe biases b through contiguous chunks b_j , similarly.

Embedding parameters By definition, $x_t = Ew_t$, where each $w_t \in \Sigma$ is overloaded to be a one-hot vector picking out a row of embedding matrix $E \in \mathbb{R}^{d \times |\Sigma|}$. We define $E = I^{2k \times 2k}$, the identity matrix. Note that $\{Ew\}_{w \in \Sigma}$ is a set of mutually orthogonal vectors, with $\|Ew\|_2 = 1$.

Parameters implementing $o_{t,j}$ as Equation E.19

The equation defining $o_{t,j}$ is as follows:

$$o_{t,j} = [\sigma(W_o h_{t-1} + U_o x_t + b_o)]_j \quad (\text{E.20})$$

$$= \sigma([W_o h_{t-1}]_j + [U_o x_t]_j + [b_o]_j) \quad (\text{E.21})$$

where the subscript j on the RHS indicates that we're picking the units corresponding to stack slot j . Let $\gamma = \tanh(1) \approx 0.762$ for ease of notation. We define $W_{o,j,j'}$ as follows:

$$W_{o,j,j'} = \begin{cases} -\lambda \in \mathbb{R}^{k \times k} & j' \in \{j, j+1\} \\ -2\lambda \in \mathbb{R}^{k \times k} & j' > j+1 \\ \mathbf{0} \in \mathbb{R}^{k \times k} & j' < j \end{cases}, \quad (\text{E.22})$$

where $\lambda \in \mathbb{R}$ is a scaling factor to be defined. This gives the intermediate value:

$$[W_o h_{t-1}]_j = \begin{cases} -\lambda\gamma & \exists j' \in \{j, j+1\}, i : \\ & h_{t-1, j'} = \tanh(\psi^{-1}(\langle i \rangle)) \\ -2\lambda\gamma & \exists j' > j+1, i : \\ & h_{t-1, j'} = \tanh(\psi^{-1}(\langle i \rangle)) \\ \mathbf{0} \in \mathbb{R}^k & o.w. \end{cases} \quad (\text{E.23})$$

Note that Equation E.23 relies on the encoding-validity property, in particular that $\psi^{-1}(\langle i \rangle) \in \{0, 1\}^k$ and $\sum_{\ell=1}^k \psi^{-1}(\langle i \rangle)_\ell = 1$ to ensure that, in the first condition,

$$\begin{aligned} -\lambda \mathbf{1}^\top h_{t-1, j} &= -\lambda \mathbf{1}^\top \tanh(s_{t-1, j}) \\ &= -\lambda \mathbf{1}^\top \tanh(\psi^{-1}(\langle i \rangle)) \\ &= -\lambda \sum_{\ell=1}^k \tanh(\psi^{-1})_\ell \\ &= -\lambda \tanh(1) \\ &= -\lambda\gamma. \end{aligned}$$

and likewise for the second condition. The third condition relies on the fact that $\psi^{-1}(\emptyset) = \mathbf{0}$.

We define $U_{o, j}$ as follows:

$$U_{o, j} = \begin{bmatrix} \lambda\gamma \sum_i x_{\langle i \rangle}^\top \\ \dots \end{bmatrix} \quad (\text{E.24})$$

that is, each row of U_f is equal to the sum, over all close brackets, of the embedding of that close bracket (transposed.) Since all $x_{\langle i \rangle}$ are orthogonal and unit norm (by the definition of E), this gives the following intermediate value:

$$[U_o x_t]_j = \begin{cases} \lambda\gamma \in \mathbb{R}^k & \exists i : w_t = \langle i \rangle \\ \mathbf{0} & o.w. \end{cases} \quad (\text{E.25})$$

Finally, we specify the bias term as

$$b_{o, j} = \begin{bmatrix} 0.5\lambda\gamma \\ \dots \end{bmatrix}, \quad (\text{E.26})$$

that is, all units of the bias are equal to the same value.

As we stated earlier, Equation E.19 refers to conditions on c_t , which do not show up in the computation of o_t ; instead, o_t is a function of h_{t-1} , and x_t . As such, we re-write Equation E.19 in terms of these values, in particular splitting the condition

on c_t being j' -top for $j' > j$ into two separate conditions,

$$o_{t, j} = \begin{cases} \mathbf{0} & \exists j' > j+1, i : h_{t-1, j'} = \tanh(\psi^{-1}(\langle i \rangle)) \\ \mathbf{0} & \exists j' \in \{j, j+1\}, i : h_{t-1, j'} = \tanh(\psi^{-1}(\langle i \rangle)) \\ & \text{and } w_t \neq \rangle_i \\ \mathbf{1} & o.w. \end{cases} \quad (\text{E.27})$$

Recall that $h_{t, j'} = \tanh(\psi^{-1}(\langle i \rangle))$ indicates that c_{t-1} is j' -top, by the induction hypothesis. With our equation above we outline two different conditions that we'll show together are necessary and sufficient to guarantee that the top slot is located at some $j' > j$. Under each condition, we set the output gate to 0 because of that (since if the top slot is at $j' > j$, then slot j is not equal to 0, and must not be let through the gate.)

While we cannot condition directly on the value of c_t , we know that c_t is j' -top for $j' > j$ under two conditions on h_{t-1} , and thus on c_{t-1} . First, c_{t-1} can be $(j+2)$ -top or greater; since only one element can be popped at once, c_t can be no less than $(j+1)$ -top. Second, c_{t-1} can be j -top or $(j+1)$ -top, and the input w_t is not some \rangle_i ; that is, it does not pop the top element off of the stack. Because w_t is not \rangle_i , it must be an open bracket, $\langle i$, pushing to the stack. Hence, under these conditions, c_t must be at least $(j+1)$ -top. If neither of these conditions hold, then c_t cannot be j' -top for $j' > j$, since if c_{t-1} is j -top or $(j-1)$ -top and $w_t = \rangle_i$, then c_t must be at most j -top. And if c_{t-1} is j' -top for $j' < j$, then it is impossible since only one bracket can be pushed at a time. Thus Equation E.27 is equivalent to Equation E.19, as required.

We now prove that parameters (W_o, U_o, b_o) implement Equation E.19, by implementing Equation E.27. In the first condition, $h_{t-1, j'} = \psi^{-1}(\langle i \rangle)$ for some $j' > j+1$ and i ; we want to show that $o_{t, j} = \mathbf{0}$. Based on our construction of parameters, we have that $o_{t, j}$ is upper-bounded by the following, when $w_t = \rangle_i$:

$$\begin{aligned} o_{t, j} &= \sigma([W_o h_{t-1}]_j + [U_o x_t]_j + [b_o]_j) \\ &\leq \sigma(-2\lambda\gamma + \lambda\gamma + .5\lambda\gamma) \\ &= \sigma(-0.5\lambda\gamma) \\ &= \mathbf{0} \end{aligned}$$

where the last equality holds under our finite precision σ by setting the scaling factor λ such that $-0.5\lambda\gamma < -\beta$, that is $\lambda > \frac{2\beta}{\gamma}$.

In the next condition, we have that $h_{t-1,j'} = e_i$ for $j' \in \{j, j+1\}$, and i , and $x_t \neq \rangle_i$. In this case, we can write out the value of $o_{t,j}$ as follows:

$$\begin{aligned} o_{t,j} &= \sigma(-\lambda\gamma + 0 + .5\lambda\gamma) \\ &= \sigma(-0.5\lambda\gamma) \\ &= \mathbf{0}, \end{aligned}$$

as required.

Finally, if neither of those conditions hold, we have that $h_{t-1,j'} \neq \psi^{-1}(\langle_i)$ for any $j' \geq j$, and so $h_{t-1,j'} = \tanh(\mathbf{0}) = \mathbf{0}$. In this case, we can lower-bound the value of $o_{t,j}$:

$$\begin{aligned} o_{t,j} &\leq \sigma(0 + 0 + 0.5\lambda\gamma) \\ &= \mathbf{1}, \end{aligned}$$

as required. This completes the proof of the output gate.

Parameters implementing $f_{t,j}$ as Equation E.16

The equation defining $f_{t,j}$ is as follows:

$$f_{t,j} = [\sigma(W_f h_{t-1} + U_f x_t + b_f)]_j \quad (\text{E.28})$$

$$= \sigma([W_f h_{t-1}]_j + [U_f x_t]_j + [b_f]_j) \quad (\text{E.29})$$

We define $W_{f,j,j'}$ as follows:

$$W_{f,j,j'} = \begin{cases} -\lambda \in \mathbb{R}^{k \times k} & j = j' \\ \mathbf{0} \in \mathbb{R}^{k \times k} & o.w. \end{cases} \quad (\text{E.30})$$

which gives the intermediate value:

$$[W_f h_{t-1}]_j = \begin{cases} -\lambda\gamma \in \mathbb{R}^k & \exists_i : h_{t-1,j} \\ & = \tanh(\psi^{-1}(\langle_i)) \\ \mathbf{0} \in \mathbb{R}^k & o.w., \end{cases} \quad (\text{E.31})$$

again relying on the embedding-validity property.

We define $U_{f,j}$ as follows:

$$U_{f,j} = \begin{bmatrix} -\lambda\gamma \sum_i x_{\rangle_i}^\top \\ \dots \end{bmatrix} \quad (\text{E.32})$$

that is, each row of U_f is equal to the negative of the sum, over all close brackets, of the embedding of that close bracket (transposed.) Since all x_{\rangle_i} are orthogonal and unit norm, this gives the following intermediate value:

$$[U_f x_t]_j = \begin{cases} -\lambda\gamma \in \mathbb{R} & \exists_i : w_t = \rangle_i \\ \mathbf{0} & o.w. \end{cases} \quad (\text{E.33})$$

Finally, we specify the bias term as

$$b_{f,j} = \begin{bmatrix} 1.5\lambda\gamma \\ \dots \end{bmatrix}, \quad (\text{E.34})$$

that is, all units of the bias are equal to the same value.

We now prove that, as given, the parameters (W_f, U_f, b_f) implement Equation E.16 assuming the induction hypothesis of Lemma 4.

Equation E.16 has two cases. In the first case, c_{t-1} is j -top and $w_t = \rangle_i$, and we need to prove $f_{t,j} = \mathbf{0}$. From the induction hypothesis, this means that $h_{t-1,j} = \tanh(\psi^{-1}(\langle_i))$, and all other slots $h_{t-1,j'} = \mathbf{0}$. The value of the gate is as follows:

$$\begin{aligned} f_{t,j} &= \sigma([W_f h_{t-1}]_j + [U_f x_t]_j + [b_f]_j) \\ &= \sigma(-\lambda\gamma - \lambda\gamma + 1.5\lambda\gamma) \\ &= \sigma(-0.5\lambda\gamma) \\ &= \mathbf{0} \end{aligned}$$

where the last equality holds by setting λ as already stated: $\lambda > \frac{2\beta}{\gamma}$.

The second case is whenever the conditions of the first case don't hold, in which case we must prove $f_{t,j} = \mathbf{1}$. Thus, we either have c_{t-1} not j -top, in which case,

$$\begin{aligned} f_{t,j} &\geq \sigma(0 - \lambda\gamma + 1.5\lambda\gamma) \\ &= \sigma(0.5\lambda\gamma) \\ &= \mathbf{1}, \end{aligned}$$

or we have $w_t \neq \rangle_i$, in which case,

$$\begin{aligned} f_{t,j} &\geq \sigma(-\lambda\gamma - 0 + 1.5\lambda\gamma) \\ &= \sigma(0.5\lambda\gamma) \\ &= \mathbf{1}, \end{aligned}$$

as required.

Parameters implementing $i_{t,j}$ as Equation E.18

The equation defining $i_{t,j}$ is as follows:

$$i_{t,j} = [\sigma(W_i h_{t-1} + U_i x_t + b_i)]_j \quad (\text{E.35})$$

$$= \sigma([W_i h_{t-1}]_j + [U_i x_t]_j + [b_i]_j) \quad (\text{E.36})$$

We define $W_{i,j,j'}$ as follows:

$$W_{i,j,j'} = \begin{cases} -\lambda \in \mathbb{R}^{k \times k} & j = 1 \\ \lambda \in \mathbb{R}^{k \times k} & j = j' + 1 \\ \mathbf{0} \in \mathbb{R}^{k \times k} & o.w. \end{cases} \quad (\text{E.37})$$

which gives the intermediate value:

$$[W_i h_{t-1}]_j = \begin{cases} -\lambda\gamma \in \mathbb{R}^k & j = 1, \exists_{i,j'} : \\ & h_{t-1,j'} = \tanh(e_i) \\ \lambda\gamma \in \mathbb{R}^k & \exists_i : h_{t-1,j-1} \\ & = \tanh(e_i) \\ \mathbf{0} \in \mathbb{R}^k & o.w. \end{cases} \quad (\text{E.38})$$

once again relying only on the encoding-validity property.

We define $U_{i,j}$ as follows:

$$U_{i,j} = \begin{bmatrix} -\lambda\gamma \sum_i x_{\langle i}^\top \\ \dots \end{bmatrix} \quad (\text{E.39})$$

Since all $x_{\langle i}$ are orthogonal and unit norm, this gives the following intermediate value:

$$[U_i x_t]_j = \begin{cases} -\lambda\gamma \in \mathbb{R}^k & \exists_i : w_t = \langle i \\ \mathbf{0} & o.w. \end{cases} \quad (\text{E.40})$$

Finally, we specify the bias term as

$$b_{i,j} = \begin{cases} -.5\lambda\gamma \in \mathbb{R}^k & j = 1 \\ -1.5\lambda\gamma \in \mathbb{R}^k & o.w. \end{cases} \quad (\text{E.41})$$

We now prove that the parameters (W_i, U_i, b_i) , when plugged into f_θ , implement Equation E.18 assuming the induction hypothesis of Lemma 4. Equation E.18 has three cases.

The condition of the first case is that $c_{t-1} = \mathbf{0}$, $j = 1$, and $w_t = \langle i$; we need to show that $i_{t,1} = \mathbf{1}$. Since $c_{t-1} = \mathbf{0}$, we have that $h_{t-1} = \mathbf{0}$. We can compute $i_{t,1}$ as follows:

$$\begin{aligned} i_{t,1} &= \sigma([W_i h_{t-1}]_j + [U_i x_t]_j + [b_i]_j) \\ &= \sigma(0 + \lambda\gamma - 0.5\lambda\gamma) \\ &= \sigma(0.5\lambda\gamma) \\ &= \mathbf{1}. \end{aligned}$$

note the bias term $-0.5\lambda\gamma$ is only set that way for $j = 1$.

The condition of the second case is that c_{t-1} is $(j-1)$ -top, and $w_t = \langle i$; we need to show that $i_{t,j} = \mathbf{1}$. From the induction hypothesis, we have that $h_{t-1,j-1} = \psi^{-1}(\langle i')$ for some i' , and h_{t-1} is 0 elsewhere. We calculate $i_{t,j}$ as follows,

$$\begin{aligned} i_{t,j>1} &= \sigma([W_i h_{t-1}]_j + [U_i x_t]_j + [b_i]_j) \\ &= \sigma(\lambda\gamma + \lambda\gamma - 1.5\lambda\gamma) \\ &= \sigma(0.5\lambda\gamma) \\ &= \mathbf{1}, \end{aligned}$$

where we do not have to consider the case $i_{t,1}$ since c_{t-1} cannot be 0-top, so this case cannot hold.

Finally, in the third case, none of the above conditions hold, and we need to prove $i_{t,j} = \mathbf{0}$. There are a number of possibilities here to enumerate. First, let $c_{t-1} = \mathbf{0}$ and $j = 1$, but $w_t \neq \langle i$. Then,

$$\begin{aligned} i_{t,1} &= \sigma(0 + 0 - 0.5\lambda\gamma) \\ &= \sigma(-0.5\lambda\gamma) \\ &= \mathbf{0}, \end{aligned}$$

Next, we let $c_{t-1} = \mathbf{0}$ and $w_t = \langle i$, but $j > 1$. Then,

$$\begin{aligned} i_{t,1} &= \sigma(0 + \lambda\gamma - 1.5\lambda\gamma) \\ &= \sigma(-0.5\lambda\gamma) \\ &= \mathbf{0}, \end{aligned}$$

Next, we let $j = 1$, $w_t = \langle i$, but $c_{t-1} \neq \mathbf{0}$. Thus c_{t-1} is j' -top for some j' .

If the second case of Equation E.18 doesn't hold, then either c_{t-1} is not $(j-1)$ -top, or $w_t \neq \langle i$; we need to prove $i_{t,j} = \mathbf{0}$. If c_{t-1} is not $(j-1)$ -top and $j > 1$, then we can upper-bound the value of $i_{t,j}$ as follows:

$$\begin{aligned} i_{t,j>1} &\leq \sigma(0 + \lambda\gamma - 1.5\lambda\gamma) \\ &= \sigma(-0.5\lambda\gamma) \\ &= \mathbf{0}. \end{aligned}$$

If c_{t-1} is $(j-1)$ -top but $w_t \neq \langle i$, then the value of $i_{t,j}$ is as follows,

$$\begin{aligned} i_{t,j>1} &= \sigma(\lambda\gamma + 0 - 1.5\lambda\gamma) \\ &= \sigma(-0.5\lambda\gamma) \\ &= \mathbf{0}, \end{aligned}$$

as required.

Parameters implementing $\tilde{c}_{t,j}$ as Equation E.17

The equation defining $\tilde{c}_{t,j}$ is as follows:

$$\begin{aligned} \tilde{c}_{t,j} &= [\tanh(W_{\tilde{c}} h_{t-1} + U_{\tilde{c}} x_t = b_{\tilde{c}})]_j \quad (\text{E.42}) \\ &= \tanh([W_{\tilde{c}} h_{t-1}]_j + [U_{\tilde{c}} x_t]_j = [b_{\tilde{c}}]_j) \quad (\text{E.43}) \end{aligned}$$

We define $W_{\tilde{c}} = \mathbf{0}$, which leads to the intermediate values:

$$[W_{\tilde{c}} h_{t-1}]_j = \mathbf{0} \quad (\text{E.44})$$

We define $U_{\tilde{c}}$ as follows:

$$U_{\tilde{c},j} = \begin{bmatrix} \lambda x_{\langle 1}^\top \\ \lambda x_{\langle 2}^\top \\ \vdots \\ \lambda x_{\langle k}^\top \end{bmatrix} \quad (\text{E.45})$$

which leads to the intermediate values:

$$[U_{\tilde{c}}x_t]_j = \begin{cases} \lambda e_i & w_t = \langle i \\ \mathbf{0} & o.w. \end{cases} \quad (\text{E.46})$$

We define $b_{\tilde{c}} = \mathbf{0}$.

We now prove that the parameters $W_{\tilde{c}}, U_{\tilde{c}}, b_{\tilde{c}}$ implement Equation E.17. There are two cases.

In the first case, $w_t = \langle i$, and need to show $\tilde{c}_{t,j} = \psi^{-1}(\langle i)$. In this case, we have

$$\begin{aligned} \tilde{c}_{t,j} &= \tanh([W_{\tilde{c}}h_{t-1}]_j + [U_{\tilde{c}}x_t]_j = [b_{\tilde{c}}]_j) \\ &= \tanh(\mathbf{0} + \lambda e_i + \mathbf{0}) \\ &= e_i, \\ &= \psi^{-1}(\langle i) \end{aligned}$$

where the last equality holds by setting $\lambda > \beta$.²⁴

In the second case, we have $w_t \neq \langle i$, and need to show $\tilde{c}_{t,j} = \mathbf{0}$. In this case, we have

$$\begin{aligned} \tilde{c}_{t,j} &= \tanh([W_{\tilde{c}}h_{t-1}]_j + [U_{\tilde{c}}x_t]_j = [b_{\tilde{c}}]_j) \\ &= \tanh(\mathbf{0} + \mathbf{0} + \mathbf{0}) \\ &= \mathbf{0} \end{aligned}$$

as required.

Summary We've now specified all parameters of the LSTM and completed the induction step in our proof of the stack correspondence lemma, Lemma 4; this completes the proof of the lemma.

F Proving generation in $O(mk)$ hidden units

For both the Simple RNN and the LSTM, we've proven *stack correspondence lemmas*, Lemmas 3, 4. These guarantee that for all prefixes of strings in Dyck- (k,m) , we can rely on properties of the hidden states of each model to be perfectly informative about the state of the DFA stack.

Given those lemmas, this section describes the proof of the following results, for the Simple RNN:

Theorem 2. *For any $k, m \in \mathbb{Z}^+$, there exists a Simple RNN f_θ with $2mk$ hidden units that generates Dyck- (k,m) .*

²⁴Which we've already required earlier by letting $\lambda > \frac{2\beta}{\gamma}$.

and for the LSTM, noting that the condition that $W_{\tilde{c}} = \mathbf{0}$ holds from our proof of the stack correspondence lemma.

Theorem 3. *For any $k, m \in \mathbb{Z}^+$, there exists a LSTM f_θ with mk hidden units and $W_{\tilde{c}} = \mathbf{0}$ that generates Dyck- (k,m) .*

The two constructions differ in where they make information accessible; we describe those differences here and then provide a general proof that is agnostic to which construction is used.

F.1 Softmax distribution parameters

Recall that we must show that the ϵ -truncated support of f_θ (for both Simple RNN and LSTM) is the same set as Dyck- (k,m) . The token-level probability distribution conditioned on the history is given as follows:

$$w_t \sim \text{softmax}(Vh_{t-1} + b), \quad (\text{F.1})$$

and we denote this distribution p_{f_θ} . We first specify V by row; a single row exists for each of the $2k + 1$ words in the vocabulary. We let $v_{w,j}$ refer to the row of word w , and the k columns that will participate in the dot product with the rows of stack slot j in h_{t-1} .

We start with the only difference in the softmax matrix between the Simple RNN and the LSTM. For the Simple RNN:

$$v_{\langle i,1} = \zeta e_i \quad (\text{F.2})$$

$$v_{\langle i,j>1} = \mathbf{0}, \quad (\text{F.3})$$

because the top of the stack is guaranteed to be in the first stack slot, and where ζ is a positive scaling constant we'll define later. And for the LSTM,

$$v_{\langle i,j} = \zeta e_i \quad (\text{F.4})$$

for all j , since exactly one slot is guaranteed to be non-empty (if the stack is non-empty) and it could be at any of the m slots. The rest of the softmax construction is common between the Simple RNN and the LSTM:

$$v_{\langle i,j<m} = \mathbf{0} \in \mathbb{R}^k \quad (\text{F.5})$$

$$v_{\langle i,m} = -\mathbf{1}\zeta \in \mathbb{R}^k \quad (\text{F.6})$$

$$v_{\omega,j,1} = -\mathbf{1}\zeta \in \mathbb{R}^k \quad (\text{F.7})$$

Likewise, the bias terms:

$$b_{v,\rangle i} = -0.5\zeta\gamma \quad (\text{F.8})$$

$$b_{v,\langle i} = 0.5\zeta\gamma \quad (\text{F.9})$$

$$b_{v,\omega} = 0.5\zeta\gamma \quad (\text{F.10})$$

So that we can swap out V when we swap out ψ for a more efficient encoding, we state here a properties of V we rely on (once ψ is fixed):

Definition 11 (softmax validity property). Given encoding $\psi : \{\langle i \rangle\}_{i \in [k]} \rightarrow \mathbb{R}^n$, a softmax matrix V obeys the softmax validity property relative to ψ if

$$\forall_{i \in [k], u} v_{\langle i, u \rangle}^\top \psi^{-1}(\langle j \rangle) \begin{cases} = \zeta & i = j \\ \leq 0 & i \neq j, \end{cases} \quad (\text{F.11})$$

where u specifies the stack slot where the construction (Simple RNN or LSTM) stores the element at the top of the stack; for the Simple RNN, $u = 1$, for the LSTM, u is such that c_t is u -top, as defined. This ensures that the softmax matrix correctly distinguishes between the symbol, $\langle i \rangle$, that is encoded in the top of the stack, from any other symbol. Further, we rely on

$$\forall_{i, j \in [k], m} v_{\langle i, m \rangle}^\top \psi^{-1}(\langle j \rangle) = -\zeta \quad (\text{F.12})$$

$$\forall_{i, j \in [k], m' < m} v_{\langle i, m' \rangle}^\top \psi^{-1}(\langle j \rangle) = 0 \quad (\text{F.13})$$

to detect whether the stack is full, and

$$\forall_{i, j \in [k], z \in [m], \omega} v_{\langle \omega, z \rangle}^\top \psi^{-1}(\langle j \rangle) = -\zeta \quad (\text{F.14})$$

to detect if the stack is not empty.

The softmax matrix we've provided obeys the softmax validity property with respect to ψ since $\zeta e_i^\top e_i = \zeta$, and $\zeta e_i^\top e_j = 0$, $j \neq i$. Further, $-\zeta \mathbf{1}^\top e_i = -\zeta$ for all i .

We're now prepared to state the final lemma in our proof of Theorems 2, 3.

Lemma 5 (probability correctness). *Let $w_{1:T} \in D_{m,k}$. For $t = 1, \dots, t$, let $q_t = D_{m,k}(w_{1:T})$. Then for all $w \in \Sigma$,*

$$\delta(q_t, w) \neq r \leftrightarrow p_{f_\theta}(w|w_{1:t}) \geq \epsilon \quad (\text{F.15})$$

Intuitively, Lemma 5 states that f_θ assigns greater than ϵ probability mass in context to tokens w_t such that the prefix $w_{1:t+1}$ is the prefix of some member of $D_{m,k}$.

Proof of Lemma 5 We proceed in cases by the state, q . We'll show that a lower-bound on the probabilities of allowed symbols is greater than an upper-bound on the probabilities of disallowed symbols. We should note, however, that this is effectively a technicality to ensure our construction is fully constructive – that is, we provide concrete values for each parameter in the model; else, we could simply indicate that as ζ grows, the probability mass on w such that $\delta(q, w) = r$ converges to 0, while the mass on all other symbols converges 1 over the number of such symbols.

Case $q = \square$ First, consider the case that $q_{t-1} = \square$. For all $\langle i \rangle$, $\delta(q, \langle i \rangle) = r$, and for all other symbols w , $\delta(q, w) \neq r$. By the stack correspondence lemmas, Lemma 4 3, we have that $\mathcal{Q}(h_{t-1}) = q_{t-1}$ (Simple RNN) or $\mathcal{Q}(c_{t-1}) = q_{t-1}$ (LSTM), and $h_{t-1} = \mathbf{0}$.

For all $\langle i \rangle$, we have $\delta(q, \langle i \rangle) = r$, and we have logits:

$$\begin{aligned} p_{f_\theta}(w_t = \langle i \rangle | h_{t-1}) \\ &\propto v_{\langle i \rangle}^\top h_{t-1} + b_{v, \langle i \rangle} \\ &\leq 0 - 0.5\zeta\gamma \\ &= -0.5\zeta\gamma, \end{aligned}$$

as required. For all $w_t = \langle i \rangle$, we have $\delta(q, \langle i \rangle) \neq r$, and we have logits:

$$\begin{aligned} p_{f_\theta}(w_t = \langle i \rangle | h_{t-1}) \\ &\propto v_{\langle i \rangle}^\top h_{t-1} + b_{v, \langle i \rangle} \\ &= 0 + 0.5\zeta\gamma \\ &= 0.5\zeta\gamma \end{aligned}$$

Finally, for ω , $\delta(q, \omega) \neq r$, and we have logits:

$$\begin{aligned} p_{f_\theta}(w_t = \omega | h_{t-1}) \\ &\propto v_\omega^\top h_{t-1} + b_{v, \omega} \\ &= 0 + 0.5\zeta\gamma \\ &= 0.5\zeta\gamma, \end{aligned}$$

as required. With the logits specified for our whole vocabulary, we can compute the partition function of the softmax function by summing over all k open brackets, all k close brackets, and the END bracket, to determine probabilities for each token,

$$\begin{aligned} p_{f_\theta}(w_t = \langle i \rangle | h_{t-1}) \\ &= \frac{e^{-0.5\zeta}}{Z_\square} \\ Z_\square &\leq (k+1)e^{0.5\zeta\gamma} + ke^{-0.5\zeta\gamma} \end{aligned}$$

while the probability for any $\langle i \rangle$ and ω is $p_{f_\theta}(w_t = \langle i \rangle | h_{t-1}) = e^{0.5\zeta\gamma}/Z_\square$. The partition function is computed as $e^{0.5\zeta\gamma}$ multiplied by the number of words w such that $\delta(q, w) \neq r$ (in this case, k open brackets and the end bracket), plus a quantity upper bounded by $e^{-0.5\zeta\gamma}$ multiplied by the number of words w such that $\delta(q, w) = r$ (in this case, k close brackets.) All probabilities under this model are of this form.

Case $q = [\langle i_1, \dots, \langle i_{m'} \rangle]$, $m' < m$ Next, consider the case that $q = [\langle i_1, \dots, \langle i_{m'} \rangle]$ for $0 < m' < m$. For the Simple RNN, we have by Lemma 3 that $h_{t-1,1} = \psi^{-1}(\langle i_{m'} \rangle)$ (the top of the stack.) Likewise for the LSTM, we have by Lemma 3, that $h_{t-1,m'} = \tanh(e_{i_{m'}})$, and $h_{t-1,j \neq m'} = \mathbf{0}$. Either way, we have $v_{\rangle_i}^\top h_{t-1} \geq \zeta\gamma$ by the softmax validity property.²⁵

As with the first case, we construct the logits for each of the symbols. First, we have $\delta(q, \langle i_{m'} \rangle) \neq r$; this is the top element of the stack, which can be popped with a close bracket. We have logits:

$$\begin{aligned} p_{f_\theta}(w_t = \rangle_i | h_{t-1}) & \propto v_{\rangle_i}^\top h_{t-1} + b_{v,\rangle_i} \\ & = \zeta\gamma - 0.5\zeta\gamma \\ & = 0.5\zeta\gamma \end{aligned}$$

as required. For $w_t = \rangle_i$ where $i \neq i_{m'}$, we have $\delta(q, \langle i \rangle) = r$ (since $\langle i \rangle$ is not the top of the stack, seeing close bracket \rangle_i would break the well-balancing requirement), and the logits:

$$\begin{aligned} p_{f_\theta}(w_t = \rangle_i | h_{t-1}) & \propto v_{\rangle_i}^\top h_{t-1} + b_{v,\rangle_i} \\ & \leq 0 - 0.5\zeta\gamma \\ & = -0.5\zeta\gamma, \end{aligned}$$

again by the softmax validity property. For $w_t = \langle i$, we have $\delta(q, \langle i \rangle) \neq r$, and the logits:

$$\begin{aligned} p_{f_\theta}(w_t = \langle i | h_{t-1}) & \propto v_{\langle i}^\top h_{t-1} + b_{v,\langle i} \\ & = 0 + 0.5\zeta\gamma \\ & = 0.5\zeta\gamma \end{aligned}$$

as required, since $m' < m$, and by the softmax validity property, $v_{\langle i}^\top h_{t,m' < m} = \mathbf{0}$. Finally, for ω , we have $\delta(q, \omega) = r$, and the logits

$$\begin{aligned} p_{f_\theta}(w_t = \omega_i | h_{t-1}) & \propto v_{\omega_i}^\top h_{t-1} + b_{v,\omega_i} \\ & = -\zeta\gamma + 0.5\zeta\gamma \\ & = -0.5\zeta\gamma, \end{aligned}$$

²⁵In particular, for the Simple RNN, this expression is equal to ζ due to the softmax validity property; because of the tanh in $h_t = o_t \odot \tanh(c_t)$ in the LSTM expression, it's equal to $\zeta\gamma < \zeta$ where $\gamma = \tanh(1)$.

by the softmax validity property, as required. To reason about the probabilities, we again need to construct the partition function of the softmax

$$Z_p \leq (k+1)e^{0.5\zeta\gamma} + ke^{-0.5\zeta\gamma}$$

where Z_p stands for Z -“partial”, for a partially full stack. (k open brackets are allowed, plus 1 close bracket; $k-1$ close brackets are disallowed, as well as the ω word.)

Case $q = [\langle i_1, \dots, \langle i_m \rangle]$ Next, consider the case that $q = [\langle i_1, \dots, \langle i_m \rangle]$, that is, a full stack with m elements. For the Simple RNN, by Lemma 3, we have that $h_{t-1,1} = \psi^{-1}(\langle i \rangle)$, and $h_{t-1,m} = \psi^{-1}(\langle j \rangle)$ for some j . (The top of the stack is $\langle i$, and the stack is full.) And for the LSTM, by Lemma 4, we have that $h_{t-1,m} = \tanh(\psi^{-1}(\langle i_m \rangle))$, and $h_{t-1,j \neq m} = \mathbf{0}$. The logits for each symbol are as follows. Identical to the last case, we have $\delta(q, \rangle_{i_m}) \neq r$, and $p_{f_\theta}(\rangle_{i_m} | h_{t-1}) \propto e^{0.5\zeta\gamma}$ as required; this is the element at the top of the stack. Also identically, all other brackets have $\delta(q, \rangle_{i \neq i_m}) = r$, and the values for those symbols are $\leq -0.5\zeta\gamma$, as required. For $w_t = \langle i$, we have $\delta(q, \langle i \rangle) = r$; this is because with m elements, the m -bound means no more elements can be pushed. The logits are:

$$\begin{aligned} p_{f_\theta}(w_t = \langle i | h_{t-1}) & \propto v_{\langle i}^\top h_{t-1} + b_{v,\langle i} \\ & = -\zeta\gamma + 0.5\zeta\gamma \\ & = -0.5\zeta\gamma, \end{aligned}$$

by the softmax validity property, as required. Finally, for ω , identically to the previous case, we have $\delta(q, \omega) = r$, and logits $e^{-0.5\zeta\gamma}$, as required. The partition function is as follows:

$$Z_f \leq ke^{0.5\zeta\gamma} + (k+1)e^{-0.5\zeta\gamma}$$

where Z_f stands for Z -“full”, for a full stack.

Case $q = [\omega]$ Last, consider the case that $q = [\omega]$. In this case, if at timestep t , then symbol $w_t = \omega$, since the only transition to state $[\omega]$ is $\delta([\], \omega) = [\omega]$. Because of this and the definition of the universe of strings as $\Sigma^*\omega$, f_θ need not be defined on strings progressing from $q = [\omega]$. Intuitively, this is because ω indicates that the string has terminated. Hence this case vacuously holds.

We thus have three partition functions, Z_p , Z_\square , and Z_f , whose values we only have bounds for.

However, we can see that as we scale ζ large, they converge to $Z_{\square} = (k+1)e^{0.5\zeta\gamma}$, $Z_f = ke^{0.5\zeta\gamma}$, $Z_p = (k+1)e^{0.5\zeta\gamma}$ because the contributions from the disallowed symbols converges to zero. Thus, the probability assigned to the lowest-probability allowed symbol converges to $\frac{1}{k+1}$ as ζ grows large, while the probability assigned to the highest-probability disallowed symbol converges to 0. So, we choose $\epsilon = \frac{1}{2(k+1)}$, and let $\zeta > \frac{2.4}{\gamma}$, so that $e^{0.5\zeta\gamma} > 10e^{-0.5\zeta\gamma}$. Under this, the smallest probability assigned to any allowed symbol is lower-bounded by $\frac{e^{0.5\zeta\gamma}}{(k+1)e^{0.5\zeta\gamma} + ke^{-0.5\zeta\gamma}} > \frac{1}{(k+1)+0.1k} > \epsilon$, and the largest probability assigned to any disallowed symbol is upper-bounded by $\frac{e^{-0.5\zeta\gamma}}{ke^{0.5\zeta\gamma}} \leq \frac{0.1}{k} = \frac{1}{10k} < \epsilon$. This completes the proof of Lemma 5.

F.2 Completing proofs of generation (Theorems 2, 3)

Now that we've proved the probability correctness lemma, we're ready to complete our proof that our Simple RNN and LSTM constructions generate Dyck- (k, m) .

Recall that we've overloaded notation, calling $D_{m,k} \subset \sigma^*$ the set of strings defining the language Dyck- (k, m) . We must show that the ϵ -truncated support of f_{θ} , which we'll call $\mathcal{L}_{f_{\theta}}$ is equal to $D_{m,k}$. We show both inclusions.

Pf. ($D_{m,k} \subseteq \mathcal{L}_{f_{\theta}}$) Let $w \in D_{m,k}$. We'll show $w \in \mathcal{L}_{f_{\theta}}$. For all prefixes $w_{1:t}$, $t = 1 \dots T$, let $q_{t-1} = D_{m,k}(w_{1:t-1})$, the state of the DFA after consuming all tokens of the prefix except the last. For the Simple RNN, by Lemma 3, we have that $\mathcal{Q}(h_{t-1}) = q_{t-1}$. For the LSTM, by Lemma 4, we have that $\mathcal{Q}(s_{t-1,1}, \dots, s_{t-1,m}) = q_{t-1}$. Given this, for either construction, by Lemma 5, we have that $p_{f_{\theta}}(w_t | w_{1:t-1}) > \epsilon$. Since this is true for all $t = 1, \dots, T$, we have that w is in the ϵ -truncated support of f_{θ} , and so $w \in \mathcal{L}_{f_{\theta}}$.

Pf. ($\mathcal{L}_{f_{\theta}} \subseteq D_{m,k}$) Let $w_{1:T} \in \mathcal{L}_{f_{\theta}}$. We'll show $w_{1:T} \in D_{m,k}$ by proving the contrapositive.

Let $w_{1:T} \notin D_{m,k}$. We have that $w_T = \omega$ by definition. From the transition function δ , we know that $\delta(q, \omega) \in \{[\omega], r\}$, that is, ω transitions from any state either to the accept state or the reject state. Because $w_{1:T} \notin D_{m,k}$, it must be that $\delta(q_{T-1}, \omega) = r$, that is, $q_T = r$. We have no guarantee about $f_{\theta}(w_T | w_{1:T-1})$, however, since we don't know whether w_{T-1} is a prefix of some string in Dyck- (k, m) . However, we do know that

there must be some t' such that the first time $q_t = r$ is for $t = t'$, that is, the first timestep in which a disallowed symbol is seen and $D_{m,k}$ transitions to the reject state r (after which it self-loops in r by definition.)

Consider then the prefix $w_{1:t'-1}$. We know that $q_{t'-1} \neq r$. So without loss of generality, let $q_{t'-1} = [\langle i_1, \dots, \langle i_{m'} \rangle]$. We then construct a string:

$$w_{1:T'} = w_1, \dots, w_{1:t'-1}, \rangle i_1, \dots, \rangle i_{m'}, \omega$$

which simply closes all the brackets on the stack represented by $q_{t'}$. By recursive application of δ , we have that $q_{T'-1} = \square$, and thus $q_{T'} = [\omega]$, meaning $w_{1:T'} \in D_{m,k}$.

Coming back to our prefix $w_{1:t'-1}$, we now know that it is a prefix of a string $w_{1:T'} \in D_{m,k}$. Thus, for our Simple RNN, by Lemma 3, we have that $\mathcal{Q}(h_{t'-1}) = q_{t'-1}$. Likewise for our LSTM, by Lemma 4, we have that $\mathcal{Q}(c_{t'-1}) = q_{t'-1}$. And by Lemma 5, since $\delta(q_{t'-1}, w_{t'}) = r$, we have $p_{f_{\theta}}(w_{t'} | w_{1:t'-1}) < \epsilon$. And so, $w_{1:T} \notin \mathcal{L}_{f_{\theta}}$. This completes the proof of Theorems 2, 3. \square

F.3 Proving the general construction in $O(k^{m+1})$

As a corollary of the above, we now formalize our proof that a general DFA construction in RNNs, using $O(k^{m+1})$ hidden units to simulate the DFA $D_{k,m}$ of Dyck- (k, m) , permits an RNN construction that generates Dyck- (k, m) . Formally,

Theorem 1 (Naive generation of Dyck- (k, m)). *For any $k, m \in \mathbb{Z}^+$, there exists a Simple RNN f_{θ} with $O(k^{m+1})$ hidden units that generates Dyck- (k, m) .*

Proof. A general construction of any DFA in an RNN has $|Q||\Sigma|$ states, (Merrill, 2019; Giles et al., 1990). For Dyck- (k, m) , $|Q||\Sigma| \in O(k^{m+1})$. Each state of the DFA $q \in Q$ is represented Σ times, once for each word in the vocabulary. If $q_t = \delta(q_{t-1}, w)$, then the hidden state is $h_{q_t, w}$, a 1-hot vector, equal to one at an index specified by and unique to (q_{t-1}, w) . By defining the mapping $\mathcal{Q}(h_{q,w}) = q$, this construction obeys a stack correspondence lemma. Finally, since the state of the RNN specifies the DFA state as a 1-hot encoding, we can define the softmax matrix V as follows. For each state q , we can simply set all rows of V corresponding to state q to explicitly encode the log-probabilities of probability distributions we just proved in (§ F). \square

G Extending to generation in $O(m \log k)$ hidden units

In this section, we prove an $O(m \log k)$ upper bound on the number of hidden units necessary to capture Dyck- (k, m) with an RNN, matching the $\Omega(m \log k)$ lower-bound. This is accomplished by defining a new mapping ψ from slots s_t to open brackets \langle_i such that we can encode k open brackets using just $3 \log k - 1$ hidden units while maintaining the stack correspondence lemmas and Dyck- (k, m) generation properties. Formally, for the Simple RNN:

Theorem 4. *For any $k, m \in \mathbb{Z}^+$, where $k > 1$, there exists a Simple RNN f_θ with $6m \lceil \log k \rceil - 2m$ hidden units that generates Dyck- (k, m) .*

Next, for the LSTM:

Theorem 5. *For any $k, m \in \mathbb{Z}^+$, where $k > 1$, there exists an LSTM f_θ with $3m \lceil \log k \rceil - m$ hidden units and $W_{\bar{c}} = \mathbf{0}$ that generates Dyck- (k, m) .*

G.1 $O(\log k)$ vocabulary encoding in the Simple RNN.

Intuitively, a simple way to encode k elements in $\log k$ space without making use of floating-point precision is to assign each element one of the $2^{\log k}$ binary configurations of $\{0, 1\}^{\log k}$. Our construction will build off of this.

Recall that there are k open brackets that need encodings in $O(\log k)$ space. Let $p^{(i)}$, for $i \in \{1, \dots, k\}$, be the i^{th} member of an arbitrary ordering of the set $\{0, 1\}^{\log k}$, that is, the set of $\log k$ binary variables. Let $\mathbf{1} \in \mathbb{R}^{\log k - 1}$ be the vector of all 1s.

Then the encoding of symbol i is:

$$\psi^{-1}(\langle_i)_* = \zeta[p^{(i)}; (1 - p^{(i)}); \mathbf{1}] \in \{0, 1\}^{3 \lceil \log k \rceil - 1} \quad (\text{G.1})$$

Where the semicolon $(;)$ denotes concatenation. This can be efficiently implemented in our simple RNN construction by modifying the U matrix to encode each ψ^{-1} . Each row of the softmax matrix V is as follows:

$$v_{\langle_i, 1} = \zeta[p^{(i)}; (1 - p^{(i)}); -\mathbf{1}; \mathbf{0}] \quad (\text{G.2})$$

$$v_{\langle_i, m} = -\zeta[\mathbf{1} \in \mathbb{R}^{2 \lceil \log k \rceil}; -\mathbf{1} \in \mathbb{R}^{\lceil \log k \rceil - 1}] \quad (\text{G.3})$$

$$v_{\omega, j} = -\zeta[\mathbf{1} \in \mathbb{R}^{2 \lceil \log k \rceil}; -\mathbf{1} \in \mathbb{R}^{\lceil \log k \rceil - 1}] \quad (\text{G.4})$$

where j is for all $j \in [1, \dots, m]$, and all slots not specified are equal to $\mathbf{0}$.

Now it suffices to prove the stack correspondence lemma and probability correctness lemmas using ψ_* and V .

Proof of stack correspondence lemma, Lemma 3. As noted in (§ D), the only relevant property of the encodings e_i used for symbols \langle_i in Lemma 3 is that it take on values in $\{0, 1\}$. This is also true of ψ_*^{-1} , so the lemma still holds.

Proof of probability correctness lemma, Lemma 5. It suffices to show that V obeys the softmax validity property (Definition 11) with respect to ψ_* to prove Lemma 5 using new encoding ψ_* . First, we have that

$$\begin{aligned} v_{\langle_j, 1}^\top \psi_*^{-1}(\langle_i) &= \left(\sum_{i=1}^{\lceil \log k \rceil} p^{(i)} p^{(j)} \right. \\ &\quad + \sum_{i=1}^{\lceil \log k \rceil} (1 - p^{(i)})(1 - p^{(j)}) \\ &\quad \left. + \sum_{i=1}^{\lceil \log k \rceil - 1} 1(-1) \right) \\ &\begin{cases} = \log k - (\log k - 1) & i = j \\ \leq (\log k - 1) - (\log k - 1) & i \neq j \end{cases} \\ &\begin{cases} = 1 & i = j \\ \leq 0 & i \neq j \end{cases}, \end{aligned}$$

Where $v_{\langle_j, 1}$ is specified since the top of the stack is always in slot 1, ensuring the first requirement. Intuitively, this dot product simply counts up the number of bits that agree between the row j and the symbol i encoded; if they're the same, all bits agree; if not, at least 1 must disagree.

Further, we have

$$\begin{aligned} v_{\langle_i, m}^\top \psi_*^{-1}(\langle_i) &= -\zeta \left(\sum_{\ell=1}^{\lceil \log k \rceil} 1(p_\ell^{(i)}) \right. \\ &\quad + \sum_{\ell=1}^{\lceil \log k \rceil} 1(1 - p_\ell^{(i)}) \\ &\quad \left. + \sum_{\ell=1}^{\lceil \log k \rceil - 1} (-1 * 1) \right) \\ &= -\zeta((\lceil \log k \rceil) - (\lceil \log k \rceil - 1)) \\ &= -\zeta \end{aligned}$$

since the count of bits $p_\ell^{(i)}$ that are 1 plus the count of negated bits $(1 - p_\ell^{(i)})$ that are one must always equal exactly $\lceil \log k \rceil$. This holds identically for $v_{\omega,j}^\top \psi_*^{-1}(\langle i \rangle)$, for all $j \in [m]$. Finally, $v_{\langle i, m' \rangle} = 0$, for $m' < m$, ensuring that $v_{\langle i, m' < m \rangle}^\top \psi_*^{-1}(\langle i \rangle) = 0$, as required. This proves that ψ_* obeys the softmax validity properties, so Lemma 5 still holds for our Simple RNN construction.

Counting hidden units Since ψ_* encodes each symbol in $d = 3\lceil \log k \rceil - 1$ space, and the Simple RNN construction constructs a stack in $2md$ space, we have that Simple RNNs can generate Dyck- (k, m) in $2m(3\lceil \log k \rceil - 1) = 6m\lceil \log k \rceil - 2m$ space. This proves Theorem 4.

G.2 $O(\log k)$ vocabulary encoding in the LSTM

We'll use a slight variation of the construction we used for the Simple RNN, made possible since the LSTM can encode the value -1 due to the hyperbolic tangent. First, for the encoding, we swap the last $\lceil \log k \rceil - 1$ values 1 for -1 :

$$\psi_*^{-1}(\langle i \rangle) = [p^{(i)}; (1 - p^{(i)}); -1] \quad (\text{G.5})$$

$$(\text{G.6})$$

And negate the corresponding values of the softmax matrix, as follows:

$$v_{\langle i, m \rangle} = -\zeta[\mathbf{1} \in \mathbb{R}^{2\lceil \log k \rceil}; -\mathbf{1} \in \mathbb{R}^{\lceil \log k \rceil - 1}] \quad (\text{G.7})$$

$$v_{\omega, j} = -\zeta[\mathbf{1} \in \mathbb{R}^{2\lceil \log k \rceil}; -\mathbf{1} \in \mathbb{R}^{\lceil \log k \rceil - 1}] \quad (\text{G.8})$$

where j is for all $j \in [1, \dots, m]$, and all slots not specified are equal to $\mathbf{0}$.

Now it suffices to prove the stack correspondence lemma and probability correctness lemmas using ψ_* and V .

Proof of stack correspondence lemma, Lemma 4. It suffices to show that ψ_* obeys the encoding-validity property (Definition 9). First, for all i , we have,

$$\begin{aligned} \mathbf{1}^\top \psi_*^{-1}(\langle i \rangle) &= \mathbf{1}^\top [p^{(i)}; (1 - p^{(i)})] - (\log k - 1) \\ &= \log k - (\log k - 1) = 1, \end{aligned}$$

as required. This was possible because we could use the -1 value in the encoding ψ_* . Second, we see that $\psi_*^{-1}(\langle i \rangle) \in \{0, 1\}^{3\lceil \log k \rceil - 1}$, as required. Third, we have still that $\psi_*(\mathbf{0}) = \mathbf{0}$. Fourth and

finally, we have by construction that all encodings $\psi_*^{-1}(\langle i \rangle)$ are distinct and none are equal to zero, by construction, since each is assigned a different bit configuration (and each bit configuration is concatenated to its negation.)

So, the encoding-validity property holds on ψ_* , and the stack correspondence lemma, Lemma 4 holds.

Proof of probability correctness lemma, Lemma 5. The proof of the probability correctness lemma holds as an immediate corollary of the proof of the same lemma for the Simple RNN using ψ_* . In particular, the only difference between the ψ_* and V used for the Simple RNN and that used for the LSTM is the swapping of a factor of -1 from a span of v to that of ψ_* , so we still have

$$v_{\langle j, u \rangle}^\top \psi_*^{-1}(\langle i \rangle) = \begin{cases} = 1 & i = j \\ \leq 0 & i \neq j \end{cases}, \quad (\text{G.9})$$

where u picks out the top of the stack from the stack correspondence lemma. We identically have

$$v_{\langle i, m \rangle}^\top \psi_*^{-1}(\langle i \rangle) = -\zeta \quad (\text{G.10})$$

again from the proof for the Simple RNN, and likewise for $v_{\omega, j}^\top \psi_*^{-1}(\langle i \rangle)$ for all $j \in [m]$. Finally, we have $v_{\langle i, m' \rangle} = \mathbf{0}$, for $m' < m$, again from the proof of the simple RNN. This proves that ψ_* obeys the softmax validity properties, so Lemma 5 still holds for our LSTM construction.

Counting hidden units Since ψ_* encodes each symbol in $d = 3\lceil \log k \rceil - 1$ space, and the LSTM construction constructs a stack in md space, we have that LSTM can generate Dyck- (k, m) in $m(3\lceil \log k \rceil - 1) = 3m\lceil \log k \rceil - m$ space. This proves Theorem 5.

G.3 Accounting of finite precision.

We've used a number of constants in our proofs; we now enumerate them to make it clear that our construction can operate in a finite-precision setting. We use the symbols as defined in their respective sections. The only values written to vector memory are $\{0, 1\}$. In the recurrent equations, we use $\{\beta, -\beta, 2\beta, -2\beta, -3\beta\}$. We refer to $\tanh(1)$ as γ , and use $\{\gamma, -\gamma, -2\gamma\}$. We use λ as a constant as well in our recurrent parameters, and use the constants $\{-\gamma\lambda, -2\gamma\lambda, \frac{1}{2}\gamma\lambda, -\frac{1}{2}\gamma\lambda, \frac{3}{2}\gamma\lambda, -\frac{3}{2}\gamma\lambda\}$. None of

these constants grow or shrink with m or k . However, when defining a distribution that must assign mass to, e.g., all k open brackets, some values *must* scale with k . This is okay, since the number of *unique* values need not scale with k ; some of the values in finite set we use just become very small as k grows. We use ζ in our definitions of V and b_v , as well as the constants $\{-\zeta, \frac{1}{2}\zeta, -\frac{1}{2}\zeta, \zeta\gamma, -\zeta\gamma, \frac{1}{2}\zeta\gamma, -\frac{1}{2}\zeta\gamma\}$ in the equations in that section. Likewise our threshold ϵ for ϵ -truncated support shrinks with k , as do the probabilities assigned by the model to tokens. For those probabilities, we choose to represent ϵ ; thus allowed symbols (with probability $> \epsilon$) will be evaluated correctly as $\geq \epsilon$. For disallowed symbols, we choose to represent the value half-way between ϵ and our upper-bound on disallowed symbols’ probability, $\frac{1}{10k}$, so these symbols’ probabilities will be correctly evaluated as $< \epsilon$. Finally, we refer to $\log k$ in our intermediate equations as well (though it need not be computed explicitly.). Our finite-precision set \mathbb{P} is the union of all finitely many constants named in this section.

H Experiment Details

In this section, we provide detail on our preliminary study on LSTM LMs learning Dyck- (k,m) from samples.

H.1 Data

We run experiments on Dyck- (k,m) for $k \in \{2, 8, 32, 128\}$ and $m \in \{3, 5\}$.

Distribution over Dyck- (k,m) . As Dyck- (k,m) is an (infinite) set, in order to train language models on it, we must define a probability distribution over it. Intuitively, we sample tokens conditioned on the current DFA (stack) state. Depending on the stack state, one or two of the actions $\{\text{push } \langle_i, \text{pop}, \text{end}\}$ are possible. For example, in the empty stack state, $\text{push } \langle_i$ and end are possible. We sample uniformly at random from the possible actions. Conditioned on choosing the $\text{push } \langle_i$ action, we sample uniformly at random from the open brackets \langle_i . Conditioned on choosing the pop action, we generate with probability 1 the \rangle_i that corresponds to the \langle_i on the top of the DFA stack (to obey the well-balancing condition. Upon choosing the end action, we generate ω and terminate.

m	3	5
Train lengths	1:84	1:180
Test lengths	85:168	181:360

Table 2: Training and testing length cutoffs (min/max) for the distributions we define over Dyck- (k,m)

Formally,

$$q_0 = [] \tag{H.1}$$

$$a_t \sim \begin{cases} U(\{\text{push } \langle_i, \text{end}\}) & |q| = 0 \\ U(\{\text{push } \langle_i, \text{pop}\}) & 0 < |q| < m \\ U(\{\text{pop}\}) & |q| = m \end{cases} \tag{H.2}$$

$$w_t \sim \begin{cases} U(\{\langle_i\}_{i \in 1:k}) & a_t = \text{push } \langle_i \\ U(\{\rangle_j\}) & a_t = \text{pop}, q = [\dots, \langle_j] \\ U(\{\omega\}) & a_t = \text{end} \end{cases} \tag{H.3}$$

We choose this distribution to provide the following statistical properties. Consider the markov chain with nodes 1 to m representing $|q_t|$, the number of symbols on the DFA state stack at timestep t . (This is a Markov chain because q_t is sufficient to describe the probability distribution over all suffixes after timestep t .) From the probability distribution we’ve defined, for any timestep where the stack is neither empty nor full, that is, $0 < |q_t| < m$ there is probability $1/2$ of advancing in the markov chain towards m , that is $|q_{t+1}| = |q_t| + 1$, and probability $1/2$ of retreating towards 0, that is, $|q_{t+1}| = |q_t| - 1$. This *hitting time* of state m from state 0 is the expected number of timesteps it takes for the generation sequence to start at the empty stack state $q_t = []$ and end up at a full stack state $|q_t| = m$. Because of the $1/2$ probability of advancing or retreating along the markov chain, the hitting time is $O(m^2)$.

Length conditions. ? showed that the choice of formal language training lengths has a significant effect on the generalization properties of trained LSTMs. We choose training lengths carefully, keeping in mind the $O(m^2)$ hitting time from empty to full stack states, to empirically ensure that in expectation, the longest training sequences traverse from the empty state to a full stack state $|q_t| = m$ and back to the empty stack state at least three times. This ensures that models are not able to use simple heuristics at training time, like re-

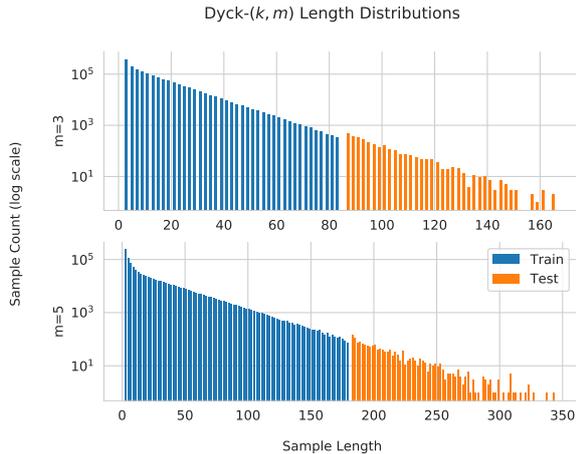


Figure 5: Sample length statistics for the twenty million token datasets. Note that the y-axis is on a log scale.

k	2	8	32	128
% of Total	100	100	15	.03
% of Test	100	100	58	31

Table 3: Percent of DFA states seen in twenty million tokens of training data, as measured with respect to both the set of all DFA states and the set of DFA states seen at test time.

remembering the first open bracket in the sequence to close the last close bracket. The exact length statistics are provided in Table 2. Length statistics for training and testing sets are shown in Figure 5..

DFA state analysis. Since $\text{Dyck-}(k, m)$ is a regular language, it is reasonable to believe it may learn equivalences between strings that result in the same DFA state, but fail to generalize to DFA states not seen during training time. In Table 3, we see that for k equal to 2 and 8, all DFA states are seen at training time. Equivalently, every possible stack configuration of 2 or 8 brackets, of stack sizes up to 3 or 5, are seen during training time. For $k = 32$, however, only 15% of all possible DFA states are seen at training time, and only 58% of DFA states seen at testing time are also seen at training time. For $k = 128$, the numbers are even more stark, where 0.3% of all possible states, and 31% of states seen at testing time are also seen at training time. Thus, the ability of models to generalize to the test set for k equal to 32 and 128 shows that the learned LSTMs are not simply memorizing

DFA states from training time.²⁶ Instead, we speculate that they’re performing stack-like operations, and leave further investigation to future work.

Sample counts To test sample efficiency of learning, we study four dataset sizes: $\{2k, 20k, 200k, 2m, 20m\}$ tokens for training for each k, m combination. In all training settings, we use identical development and test sets of size $20k$ and $300k$ tokens, respectively. The development set is sampled from the training distribution.

H.2 Models

We use LSTMs, defined as in the main text with a linear readout layer, and implemented in PyTorch (Paszke et al., 2019). We set the hidden dimensionality of the LSTM to $3m \log(k) - m$, and the input dimensionality to $2k + 10$.

H.3 Training

We use the default LSTM initialization provided by PyTorch. We train using Adam (?), using a starting learning rate of 0.01 for all training sets less than $2m$ tokens. Based on hand hyperparameter optimization, we found that for $k = 128$, at $2m$ tokens it was better to use a starting learning rate of 0.001. For training sets of size $20m$ tokens, we use a starting learning rate of 0.001 for all settings of k . We use a batch size of 10 for all experiments. We evaluate perplexity on the development set after every epoch, restarting Adam with a learning rate decayed by 0.5 if the development perplexity does not achieve a new minimum. After three consecutive epochs without a new minimum development perplexity, we stop training. We use no explicit regularization.

H.4 Evaluation

We’re interested in evaluating the behavior of the LSTM LMs in hierarchical memory management, that is, in remembering what type of bracket $\}_i$ can come next. The model cannot possibly do better than random at predicting the next open bracket. Other aspects of the language, like whether the string can end, or if the stack is full, can be solved easily with a counter, which LSTMs are known to implement (Weiss et al., 2018; Suzgun et al., 2019). We thus evaluate whether, for each observed close bracket, the LM is confident about which close

²⁶There are over 34 billion possible DFA states for $k = 128, m = 5$.

bracket (that is, i for \rangle_i) it is, since this is deterministic; it must be the close bracket corresponding to the open bracket at the top of the stack of q_{t-1} . We do this by normalizing the probability assigned by the LM to the *correct* close bracket by the sum of probabilities assigned to *any* close bracket:

$$p(\rangle_j|\rangle) = \frac{p(\rangle_j)}{\sum_i p(\rangle_i)} \quad (\text{H.4})$$

We evaluate whether the model is confident, when we define as $p(\rangle_j|\rangle) > 0.8$.

Most pairs of open and close brackets in our distribution of Dyck- (k,m) are near linearly, but we're interested in *long-distance* memory management. Thus, we let p_ℓ be the empirical probability that the model confidently predicts a close bracket, conditioned on it being separated from its open bracket by ℓ tokens. We then evaluate models by taking $\text{mean}_\ell p_\ell$; that is, to evaluate its performance at closing brackets? To get a value of 1 for this metric, models must confidently close every bracket.

We train with three independent seeds for each training setting (training set size, k , and m), and report the median across seeds of our bracket closing metric in Figure 3.