

# On Efficiently Representing Regular Languages as RNNs

Anej Svete Robin Shing Moon Chan Ryan Cotterell

{asvete, chanr, ryan.cotterell}@inf.ethz.ch

ETH zürich

## Abstract

Recent work by Hewitt et al. (2020) provides an interpretation of the empirical success of recurrent neural networks (RNNs) as language models (LMs). It shows that RNNs can efficiently represent bounded hierarchical structures that are prevalent in human language. This suggests that RNNs’ success might be linked to their ability to model hierarchy. However, a closer inspection of Hewitt et al.’s (2020) construction shows that it is not inherently limited to hierarchical structures. This poses a natural question: What other classes of LMs RNNs can efficiently represent? To this end, we generalize Hewitt et al.’s (2020) construction and show that RNNs can efficiently represent a larger class of LMs than previously claimed—specifically, those that can be represented by a pushdown automaton with a bounded stack and a specific stack update function. Altogether, the efficiency of representing this diverse class of LMs with RNN LMs suggests novel interpretations of their inductive bias.



<https://github.com/rycolab/bpdas>

## 1 Introduction

Neural LMs have demonstrated a human-level grasp of grammar and linguistic nuance. Yet, a considerable gap remains between our empirical observations and our theoretical understanding of their capabilities. One approach to bridge this gap is to study what classes of formal languages neural LMs can efficiently represent. The rationale behind this object of study is that a neural LM’s ability to represent a language efficiently suggests the presence of an inductive bias in the model architecture that prefers that language over others, and may make it specifically easier to learn, e.g., due to Occam’s razor.

Exploring LMs’ ability to model formal languages has garnered significant interest in recent years; see, e.g., the surveys by Merrill (2023) and Strobl et al. (2023). Investigation into this area has a particularly long history in the context of RNNs (McCulloch and Pitts, 1943; Minsky, 1954; Siegel-

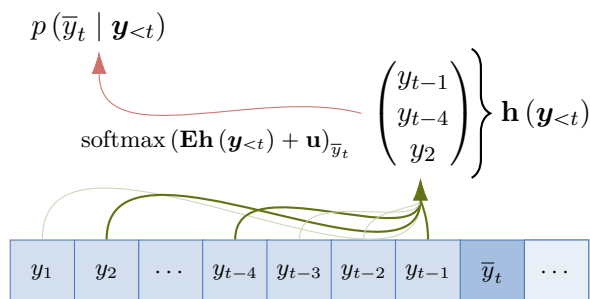


Figure 1: An illustration of how an RNN can store information about a fixed number of symbols (in this case, three) that have appeared in the string  $\mathbf{y}_{<t}$ . Using some mechanism, the symbols  $y_2, y_{t-4}, y_{t-1}$  have been selected for determining the continuation of the string and are stored in  $\mathbf{h}$ . These symbols are used to compute the conditional probability of the next symbol  $\bar{y}_t$ .

mann and Sontag, 1992).<sup>1</sup> For example, a classic result (Minsky, 1954) states that RNNs are equivalent to finite-state automata (FSAs). How *efficiently* an RNN can encode an FSA was studied by Indyk (1995), who showed that an FSA with states  $Q$  over an alphabet  $\Sigma$  can be simulated by an RNN with  $\mathcal{O}(|\Sigma|\sqrt{|Q|})$  neurons. This construction is optimal in the sense that there exist FSAs that *require* this many neurons to be emulated by an RNN.

Indyk’s (1995) bound presents a *worst-case* analysis: It reflects the number of neurons needed for an adversarially selected FSA. However, it is easy to construct FSAs that can be encoded with exponentially fewer neurons than the number of states. Fig. 2 exhibits an  $n$ -gram LM that, when encoded as an FSA, has  $|\Sigma|^{n-1}$  states, but can still be represented with  $\mathcal{O}(n \log |\Sigma|)$  neurons. Building on this insight, Hewitt et al. (2020) show that an entire class of languages— $D(b, n)$ , i.e., the Dyck language over  $b$  parentheses types with nesting up to depth  $n$ —can be encoded in logarithmic space. Specifically, even though an FSA that accepts the

<sup>1</sup>In terms of empirical performance on statistical language modeling, RNNs constituted the empirical state of the art until recently (Qiu et al., 2020; Orvieto et al., 2023), and have, despite the prominence of transformer-based LMs, seen a resurgence of late (Peng et al., 2023; Orvieto et al., 2023; Zhou et al., 2023).

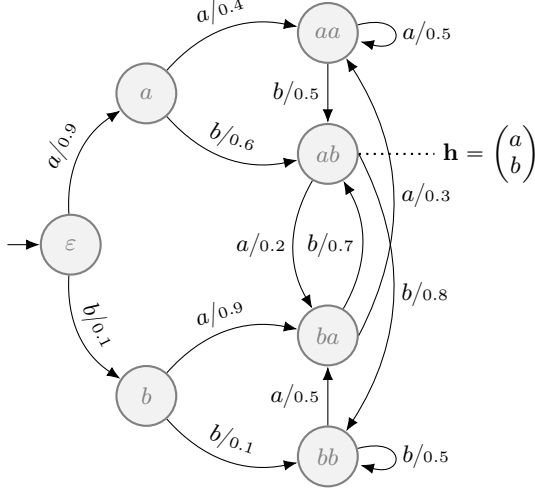


Figure 2: A simplified 3-gram LM over  $\Sigma = \{a, b\}$ . Even though the number of states is exponential in  $n$ , the hidden state of the RNN only has to keep the  $n - 1 = 2$  symbols of interest, each of which is represented by  $\lceil \log_2 |\Sigma| \rceil$  bits. This is illustrated by the state  $ab$  being represented as  $\mathbf{h} = \begin{pmatrix} a \\ b \end{pmatrix}$ .<sup>3</sup>

$D(b, n)$  language requires  $\mathcal{O}(b^n)$  states, it can be encoded by an RNN with  $\mathcal{O}(n \log b)$  neurons, which is memory-optimal. Attractively,  $D(b, n)$  languages capture, to some extent, the bounded nested nature of human language.<sup>2</sup>

Hewitt et al.’s (2020) result poses an interesting question: Are all languages that RNNs can implement efficiently hierarchical in nature? We show that this is *not* the case. We revisit Hewitt et al.’s (2020) construction and demonstrate that the same optimal compression can be achieved for a more general class of languages, which do not necessarily exhibit hierarchical structure. Stated differently, RNNs do not necessarily encode hierarchical languages any more compactly than non-hierarchical ones. To show this, we introduce probabilistic **bounded pushdown automata** (BPDAs), probabilistic pushdown automata (Abney et al., 1999) whose stack is bounded. The probabilistic nature of BPDAs furthermore allows us to go beyond the binary recognition and reason about classes of **language models**—probability distributions over strings—that RNN LMs can efficiently represent. This is particularly interesting

<sup>2</sup>Bounded Dyck languages offer a useful framework for modeling the hierarchical and recursive aspects of human language syntax with limited memory. However, they fall short of capturing the full complexity of human language, which often includes context sensitivity and ambiguity that may go beyond the capabilities of a deterministic model of computation.

<sup>3</sup>For simplicity, the final weights and the binary representations in the hidden state are omitted.

because it requires the RNN to not only efficiently encode the transition dynamics of the automaton but also the string probabilities.<sup>4</sup> Thus, we generalize the result by Hewitt et al. (2020) and give sufficient conditions for a probabilistic BPDA to be optimally compressible into an RNN.

The non-hierarchical nature of BPDAs leads us to contend that the reason some LMs lend themselves to exponential compression has little to do with their hierarchical structure. This offers two intriguing insights. First, it shows that the inductive biases of RNN LMs might be difficult to link to a hierarchical structure. Second, it provides a new perspective on which languages RNNs efficiently encode, i.e., those representable by sequential machines with a bounded stack. Such a machine is illustrated in Fig. 1.

## 2 Preliminaries

We begin by introducing some core concepts. An **alphabet**  $\Sigma$  is a finite, non-empty set of **symbols**. Its **Kleene closure**  $\Sigma^*$  is the set of all strings of symbols in  $\Sigma$ . The **length** of the string  $\mathbf{y} = y_1 \dots y_T \in \Sigma^*$ , denoted by  $|\mathbf{y}| = T$ , is the number of symbols it contains. A **language model**  $p$  is a probability distribution over  $\Sigma^*$ . Two LMs  $p$  and  $q$  are **weakly equivalent** if  $p(\mathbf{y}) = q(\mathbf{y})$  for all  $\mathbf{y} \in \Sigma^*$  and two families of LMs  $\mathcal{P}$  and  $\mathcal{Q}$  are weakly equivalent if, for any  $p \in \mathcal{P}$ , there exists a weakly equivalent  $q \in \mathcal{Q}$  and vice versa.

Most modern LMs define  $p(\mathbf{y})$  as a product of conditional probability distributions:

$$p(\mathbf{y}) \stackrel{\text{def}}{=} p(\text{EOS} \mid \mathbf{y}) \prod_{t=1}^{|\mathbf{y}|} p(y_t \mid \mathbf{y}_{<t}), \quad (1)$$

where  $\text{EOS} \notin \Sigma$  is a distinguished end-of-string symbol. We denote  $\bar{\Sigma} \stackrel{\text{def}}{=} \Sigma \cup \{\text{EOS}\}$  and use the notation  $\bar{y} \in \bar{\Sigma}$  whenever  $\bar{y}$  can also be EOS. Such a definition is without loss of generality; any LM can be factorized in this form (Cotterell et al., 2024). However, not all models expressible as Eq. (1) constitute LMs, i.e., a probabilistic model of the form given in Eq. (1) may leak probability mass to infinite sequences (Du et al., 2023). In this paper, we assume all autoregressive models are tight, i.e., they place probability 1 on  $\Sigma^*$ .

A historically important class of LM are those that obey the  $n$ -gram assumption.

<sup>4</sup>Note the same lower bounds for FSAs do not apply to the probabilistic case (Svete and Cotterell, 2023b).

**Assumption 2.1.** The *n*-gram assumption states that the probability  $p(\bar{y}_t \mid \mathbf{y}_{<t})$  only depends on  $n - 1$  previous symbols  $y_{t-1}, \dots, y_{t-n+1}$ :

$$p(\bar{y}_t \mid \mathbf{y}_{<t}) = p(\bar{y}_t \mid y_{t-n+1} \cdots y_{t-1}). \quad (2)$$

Fig. 2 shows an example of an 2-gram LM. The weights on the transitions denote the conditional probabilities of the next symbol given the previous  $n - 1 = 2$  symbols encoded by the current state.

## 2.1 Recurrent Neural Language Models

The conditional distributions of recurrent neural LMs are given by a recurrent neural network. Hewitt et al. (2020) present results for both Elman RNNs (Elman, 1990) as well as those derived from the LSTM architecture (Hochreiter and Schmidhuber, 1997). Our paper focuses on Elman RNNs, as they are easier to analyze and suffice to present the main ideas, which also easily generalize to the LSTM case.

**Definition 2.1.** An *Elman RNN*  $\mathcal{R} = (\Sigma, \sigma, D, \mathbf{U}, \mathbf{V}, \mathbf{b}, \boldsymbol{\eta})$  is an RNN with the hidden state recurrence

$$\mathbf{h}_0 = \boldsymbol{\eta} \quad (t = 0) \quad (3a)$$

$$\mathbf{h}_t = \sigma(\mathbf{U}\mathbf{h}_{t-1} + \mathbf{V}\mathbf{r}(y_t) + \mathbf{b}) \quad (t > 0), \quad (3b)$$

where  $\mathbf{h}_t \in \mathbb{R}^D$  is the hidden state at time step  $t$ ,  $\boldsymbol{\eta} \in \mathbb{R}^D$  is an initialization parameter,  $y_t \in \Sigma$  is the input symbol at time step  $t$ ,  $\mathbf{r}: \Sigma \rightarrow \mathbb{R}^R$  is a symbol representation function,  $\mathbf{U} \in \mathbb{R}^{D \times D}$ ,  $\mathbf{V} \in \mathbb{R}^{D \times R}$  are parameter matrices,  $\mathbf{b} \in \mathbb{R}^D$  is a bias vector, and  $\sigma: \mathbb{R}^D \rightarrow \mathbb{R}^D$  is an element-wise non-linear activation function. We refer to the dimensionality of the hidden state,  $D$ , as the *size* of the RNN, and to each entry of the hidden state as a *neuron*.

Because  $\mathbf{h}_t$  represents the string consumed by the RNN, we also use the evocative notation  $\mathbf{h}(\mathbf{y})$  to denote the result of the application of Eq. (3b) over the string  $\mathbf{y} = y_1 \cdots y_t$ . An RNN can be used to specify an LM by using the hidden states to define the conditional distributions over  $\bar{y}$  given  $\mathbf{y}_{<t}$ .

**Definition 2.2.** Let  $\mathcal{R}$  be an Elman RNN,  $\mathbf{E} \in \mathbb{R}^{|\bar{\Sigma}| \times D}$  and  $\mathbf{u} \in \mathbb{R}^{|\bar{\Sigma}|}$ . An *RNN LM* is an LM whose conditional distributions are defined as

$$p(\bar{y} \mid \mathbf{y}_{<t}) \stackrel{\text{def}}{=} \text{softmax}(\mathbf{E}\mathbf{h}(\mathbf{y}_{<t}) + \mathbf{u})_{\bar{y}} \quad (4)$$

for  $\bar{y} \in \bar{\Sigma}$ ,  $\mathbf{y}_{<t} \in \Sigma^*$ .<sup>5</sup> We term  $\mathbf{E}$  the *output matrix* and  $\mathbf{u}$  the *bias vector*.

<sup>5</sup>Throughout the paper, we index vectors and matrices directly with symbols from  $\bar{\Sigma}$ . This is possible because of a trivial bijective relationship between  $\bar{\Sigma}$  and  $[\bar{\Sigma}]$ .

### 2.1.1 Activation Functions and Precision

An important consideration when analyzing RNN LM's ability to compute the probability of a string  $\mathbf{y} \in \Sigma^*$  is the number of bits required to represent the entries in  $\mathbf{h}_t$  and how the number of bits scales with the length of the string,  $|\mathbf{y}|$ . This depends both on the dynamics of the RNN and the activation function used (Merrill, 2019) and motivates the following definition of precision.

**Definition 2.3.** The *precision* of an RNN is the number of bits required to represent  $\mathbf{h}(\mathbf{y})$ :

$$\psi_{\mathcal{R}}(\mathbf{y}) \stackrel{\text{def}}{=} \max_{d \in [D]} \min_{\substack{p, q \in \mathbb{N}, \\ \frac{p}{q} = \mathbf{h}(\mathbf{y})_d}} \lceil \log_2 p \rceil + \lceil \log_2 q \rceil. \quad (5)$$

We say that an Elman RNN is of *constant precision* if  $\psi_{\mathcal{R}}(\mathbf{y}) = \mathcal{O}(1)$ , i.e., if  $\psi_{\mathcal{R}}(\mathbf{y}) \leq C$  for all  $\mathbf{y} \in \Sigma^*$  and some  $C \in \mathbb{R}$ . It is of *unbounded precision* if  $\psi_{\mathcal{R}}(\mathbf{y})$  cannot be bounded by a function of  $|\mathbf{y}|$ .

Common choices for the nonlinear function  $\sigma$  in Eq. (3b) are the Heaviside function  $\mathbf{H}(x) \stackrel{\text{def}}{=} \mathbb{1}\{x > 0\}$ , its continuous approximation, the sigmoid function  $\sigma(x) \stackrel{\text{def}}{=} \frac{1}{1 + \exp(-x)}$ , and ReLU  $\stackrel{\text{def}}{=} \max(0, x)$ . Hewitt et al. (2020) focus on sigmoid activations as originally presented by Elman (1990). However, to simplify the analysis, they assume that  $|x| \gg 0$ , such that  $\sigma(x) \approx 0$  or  $\sigma(x) \approx 1$  since  $\lim_{x \rightarrow -\infty} \sigma(x) = 0$  and  $\lim_{x \rightarrow \infty} \sigma(x) = 1$ . Concretely, they define a parameter  $\beta \in \mathbb{R}_+$  and assume  $\sigma(x) = 0$  for  $x < -\beta$  and  $\sigma(x) = 1$  for  $x > \beta$ . Restricting to values of  $x$  with  $|x| > \beta$  results in a *constant-precision* RNN, equivalent to one with the Heaviside activation function. Its hidden states live in  $\{0, 1\}^D$ , meaning that they can be interpreted as binary vectors. The update rule Eq. (3b) can then be interpreted as a logical operation on the hidden state and the input symbol (Svete and Cotterell, 2023b). We simplify the exposition by working directly with H-activated RNNs, which additionally allows for easy analysis of the RNN's precision.<sup>6</sup>

## 3 Bounded Pushdown Automata

In this section, we present a broad class of LMs that provide a convenient framework for analyzing the sufficient conditions for efficient representation by RNNs: bounded pushdown automata. They maintain a stack, albeit one that contains at most a fixed number of symbols.

<sup>6</sup>Since the Heaviside function can be implemented as a difference of two ReLU functions (over the *integers*), all our constructions work with the ReLU activation function as well, albeit with networks of twice the size.

**Definition 3.1.** An  $m$ -bounded pushdown automaton (BPDA) is a tuple  $(\Sigma, \Gamma, m, \mu, \lambda, \rho)$  where  $m \in \mathbb{N}$ ,  $\Sigma$  is an alphabet of input symbols,  $\Gamma$  is an alphabet of stack symbols,  $\mu: \Gamma^{\leq m} \times \Sigma \times \Gamma^{\leq m} \rightarrow [0, 1]$  is a weighted transition function,  $\lambda: \Gamma^{\leq m} \rightarrow [0, 1]$  is the initial weight function, and  $\rho: \Gamma^{\leq m} \rightarrow [0, 1]$  is the final weight function.

The string  $\gamma$  currently stored in the stack is the BPDA's **configuration**. We read  $\gamma$  bottom to top, e.g., in the stack  $\gamma = \gamma_1 \gamma_2 \dots \gamma_\ell$ ,  $\gamma_1$  is at the bottom of the stack, while  $\gamma_\ell$  is at the top. We say that the  $m$ -bounded stack is **empty** if  $|\gamma| = 0$  (equivalently,  $\gamma = \varepsilon$ ) and **full** if  $|\gamma| = m$ . Sometimes, it will be useful to think of the bounded stack as always being full—in that case, if there are  $\ell < m$  elements on the stack, we assume that the other  $m - \ell$  elements are occupied by a special placeholder symbol  $\iota \notin \Gamma$ ; see Fig. 3 for an illustration. We denote  $\Gamma_\iota \stackrel{\text{def}}{=} \Gamma \cup \{\iota\}$ .

We call a BPDA **probabilistic** if the following two equations hold

$$\sum_{\gamma \in \Gamma^{\leq m}} \lambda(\gamma) = 1 \quad (6a)$$

$$\sum_{\substack{y \in \Sigma \\ \gamma' \in \Gamma^{\leq m}}} \mu(\gamma, y, \gamma') + \rho(\gamma) = 1, \forall \gamma \in \Gamma^{\leq m}. \quad (6b)$$

We call a BPDA  $\mathcal{P} = (\Sigma, \Gamma, m, \mu, \lambda, \rho)$  **deterministic** if there exists exactly one  $\gamma \in \Gamma^{\leq m}$  with  $\lambda(\gamma) > 0$  and, for any  $\gamma \in \Gamma^{\leq m}$  and  $y \in \Sigma$ , there is at most one  $\gamma' \in \Gamma^{\leq m}$  with  $\mu(\gamma, y, \gamma') > 0$ . For deterministic BPDAs, we also define:

- $\phi(\gamma, y) \stackrel{\text{def}}{=} \gamma'$  for  $\mu(\gamma, y, \gamma') > 0$  as the function returning the (deterministic) next configuration of the BPDA,
- $\omega(\gamma, y) \stackrel{\text{def}}{=} w$  for  $w = \mu(\gamma, y, \phi(\gamma, y))$  as the weight of the only  $y$ -labeled transition from  $\gamma$ , and
- $\varphi(y)$  as the (unique) stack configuration reached upon reading  $y$ .

**Runs of a BPDA.** A BPDA processes a string  $y = y_1 \dots y_T \in \Sigma^*$  left to right by reading its symbols and changing its configurations accordingly. It starts with an initial configuration  $\gamma_0$  according to  $\lambda$ . Then, it updates the stack according to  $\mu$  for each symbol in  $y$ , resulting in a sequence of stacks  $\pi = \gamma_0, \gamma_1, \dots, \gamma_T$  where  $\mu(\gamma_{t-1}, y_t, \gamma_t) > 0$ . Each such sequence of stacks  $\pi$  is called a **run** and

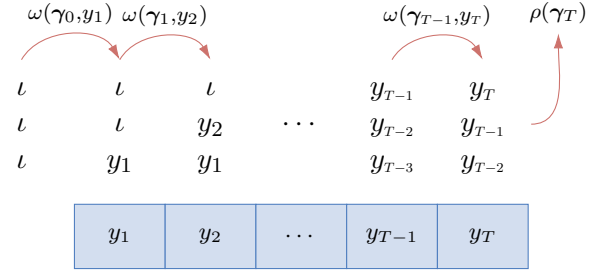


Figure 3: An illustration of how a BPDA can compute the probability of a string under an  $n$ -gram LM.

we will denote all runs of  $\mathcal{P}$  on  $y$  as  $\Pi(y; \mathcal{P})$ . A probabilistic BPDA  $\mathcal{P}$  assigns  $y$  the probability

$$\mathcal{P}(y) \stackrel{\text{def}}{=} \sum_{\substack{\pi \in \Pi(y; \mathcal{P}), \\ \pi = \gamma_0, \gamma_1, \dots, \gamma_T}} \lambda(\gamma_0) \left[ \prod_{t=1}^T \mu(\gamma_{t-1}, y_t, \gamma_t) \right] \rho(\gamma_T). \quad (7)$$

In this sense, *probabilistic* BPDAs induce distributions over strings and generalize Hewitt et al.'s (2020) thresholded string acceptance to the probabilistic setting.<sup>7</sup> This is illustrated in Fig. 3.

**Pushing and popping.** We define the transition function  $\mu$  as a general function of the stack configuration and the input symbol. As important special cases,  $\mu$  can define the standard POP and PUSH operations. Popping the top of stack  $\tau$  is performed by transitions of the form  $(\gamma\tau, y, \gamma)$  with  $\mu(\gamma\tau, y, \gamma) > 0$ ;

$$\text{POP}(\gamma\tau, y, \gamma) \stackrel{\text{def}}{=} \mu(\gamma\tau, y, \gamma). \quad (8)$$

Pushing definitionally *increases* the size of the stack, which raises the question of how to handle stack overflows. Rather than simply rejecting PUSHs that would result in a stack overflow, we allow a BPDA to *discard* the bottom of the stack when pushing; this is easy to specify with the general transition function  $\mu$ . Given the current stack configuration  $\gamma = \gamma_1 \dots \gamma_\ell$ , we define the weight of pushing the symbols  $\tau = \gamma'_1 \dots \gamma'_r \in \Gamma^{\leq m}$  as

$$\text{PUSH}(\gamma, y, \gamma\tau) \stackrel{\text{def}}{=} \begin{cases} \mu(\gamma, y, \gamma\tau) & \text{if } \ell + r \leq m \\ \mu(\gamma, y, \gamma_{\ell+r-m+1} \dots \gamma_\ell \tau) & \text{otherwise.} \end{cases} \quad (9)$$

<sup>7</sup>Hewitt et al.'s (2020) construction focuses on binary recognition of languages through their notion of truncated language recognition (cf. Def. A.1 in App. A.1). Specifically, they show that RNN LMs can assign high enough probabilities to correct continuations of strings while assigning low probabilities to incorrect continuations.



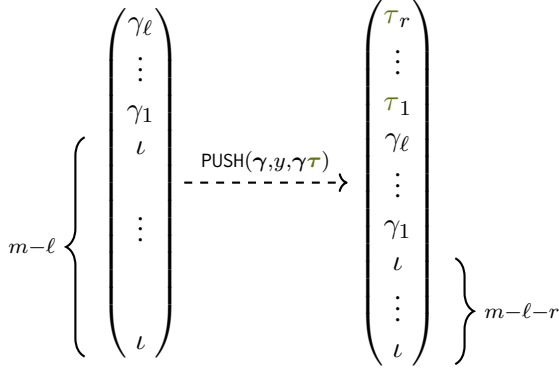


Figure 4: PUSH moves the stack down, discards the bottom-most elements, and inserts a new top.

This definition is without loss of generality; we can always define BPDAs that assign transitions resulting in a stack overflow probability 0. PUSH is illustrated in Fig. 4.

### 3.1 Efficiently Representable BPDAs

It is easy to see that there are  $\mathcal{O}(|\Gamma_\iota|^m)$  possible  $m$ -bounded stacks over  $\Gamma$ , defining the number of configurations a BPDA can be in. Interpreting the configurations as states of a (probabilistic) FSA, general constructions of RNN simulating FSAs mentioned in §1 would require  $\Omega\left(|\Sigma||\Gamma_\iota|^{\frac{m}{2}}\right)$  neurons, exponentially many in the size of the stack (Minsky, 1954; Dewdney, 1977; Indyk, 1995; Svete and Cotterell, 2023a,b). One might, however, hope to exploit the special structure of the state space induced by the stack configurations to come up with more efficient representations. This motivates the following definition.

**Definition 3.2.** A BPDA  $\mathcal{P} = (\Sigma, \Gamma, m, \mu, \lambda, \rho)$  is called *C-efficiently representable* if there exists a weakly equivalent RNN LM of size

$$D \leq C \lceil \log_2 |\Gamma_\iota|^m \rceil = Cm \lceil \log_2 |\Gamma_\iota| \rceil. \quad (10)$$

Intuitively, Def. 3.2 characterizes efficiently representable BPDAs as those that can be defined by an RNN LM with logarithmically many neurons in the number of configurations. This is a natural generalization of (efficient) *thresholded* acceptance defined by Hewitt et al. (2020) to the probabilistic (language modeling) setting.<sup>8</sup>

Def. 3.2 paves the way to a formalization of LMs efficiently representable by RNN LMs with

<sup>8</sup>Thresholded acceptance defines string recognition based on their conditional probabilities. See App. A.1 for details.

BPDAs. However, BPDAs are too rich to be efficiently representable in general. The following theorem shows that BPDA LMs are weakly equivalent to LMs induced by probabilistic finite-state automata, a classic family of computational models with a well-described relationship to RNNs.

**Theorem 3.1.** *The family of LMs induced by PF-SAs is weakly equivalent to the family of LMs induced by BPDAs.*

*Proof.* See App. B. ■

Because BPDAs define the same class of LMs as probabilistic FSAs, they cannot, in general, be represented more efficiently than with  $\Omega(|\Sigma||\Gamma_\iota|^m)$  neurons—a more efficient simulation of general BPDAs would contradict the known lower bounds on simulating FSAs (Indyk, 1995; Svete and Cotterell, 2023b). This implies that we inevitably must *restrict* the class of BPDAs to ensure their efficient representability. The following definitions will help us characterize efficiently representable BPDAs.

**Vectorial representation of a stack.** When connecting BPDAs to RNNs, it is useful to think of the stack as a vector. Setting  $G \stackrel{\text{def}}{=} \lceil \log_2 |\Gamma_\iota| \rceil$ , we define the **stack vector** function  $\chi: \Gamma^{\leq m} \rightarrow \{0, 1\}^{mG}$  as

$$\chi(\gamma_1 \cdots \gamma_m) \stackrel{\text{def}}{=} \begin{pmatrix} \text{Bin}(\gamma_m) \\ \vdots \\ \text{Bin}(\gamma_1) \end{pmatrix} \in \{0, 1\}^{mG}. \quad (11)$$

Here,  $\text{Bin}: \Gamma_\iota \rightarrow \{0, 1\}^G$  is the binary encoding function, where we assume that  $\text{Bin}(\iota) = \mathbf{0}_G$ , the  $G$ -dimensional vector of zeros. More precisely, let  $c: \Gamma_\iota \rightarrow \{0, 1, \dots, |\Gamma|\}$  be a bijection between  $\Gamma_\iota$  and  $\{0, 1, \dots, |\Gamma|\}$  such that  $c(\iota) = 0$ . Then  $\text{Bin}(\gamma)$  is the binary representation of  $c(\gamma)$ .

**Definition 3.3.** A function  $f: \Gamma^{\leq m} \rightarrow \Gamma^{\leq m}$  is *stack-affine* if there exists a matrix  $\mathbf{M} \in \mathbb{R}^{mG \times mG}$  and a vector  $\mathbf{v} \in \mathbb{R}^{mG}$  such that, for all  $\gamma \in \Gamma^{\leq m}$ , it holds that  $\chi(f(\gamma)) = \mathbf{M}\chi(\gamma) + \mathbf{v}$ .

**Definition 3.4.** A function  $\zeta: \Gamma^{\leq m} \times \Sigma \rightarrow \Gamma^{\leq m}$  is *K-varied* in  $\Sigma$  if there exists a partition  $\Sigma = \Sigma_1 \sqcup \cdots \sqcup \Sigma_K$  such that, for all  $k = 1, \dots, K$ , it holds that  $\zeta(\gamma, y) = \zeta(\gamma, y') \stackrel{\text{def}}{=} \zeta_k(\gamma)$  for all  $y, y' \in \Sigma_k$  and  $\gamma \in \Gamma^{\leq m}$ .

**Definition 3.5.** A function  $\alpha: \Gamma^{\leq m} \times \Sigma \rightarrow \Gamma^{\leq m}$  is  *$\Sigma$ -determined* if there exists a function  $s: \Sigma \rightarrow \Gamma$  and a family of partitions

$([m] = \mathcal{J}_1^y \sqcup \mathcal{J}_2^y \sqcup \mathcal{J}_3^y)_{y \in \Sigma}$  such that, for all  $\gamma \in \Gamma^{\leq m}$ , it holds that

$$\alpha(\gamma, y)_j = \begin{cases} \gamma_j & \text{if } j \in \mathcal{J}_1^y \\ s(y) & \text{if } j \in \mathcal{J}_2^y \\ \iota & \text{if } j \in \mathcal{J}_3^y \end{cases}. \quad (12)$$

In words, a stack-affine function can be implemented by an affine transformation of the vectorial representation of the stack, a  $K$ -varied function can be decomposed into  $K$  different functions that are invariant to the input symbol, and a  $\Sigma$ -determined function changes the stack in a manner that only depends on the input symbol: it either keeps a symbol the same, replaces it with  $s(y)$ , or empties the slot (inserting the placeholder symbol  $\iota$ ). Importantly, a  $\Sigma$ -determined function acts independent of the stack  $\gamma$ .

Lastly, we consider the efficient representation of next-symbol probabilities. For a deterministic BPDA  $\mathcal{P} = (\Sigma, \Gamma, m, \mu, \lambda, \rho)$ , we define

$$p(y | \gamma) \stackrel{\text{def}}{=} \omega(\gamma, y) \quad (13a)$$

$$p(\text{EOS} | \gamma) \stackrel{\text{def}}{=} \rho(\gamma), \quad (13b)$$

for  $\gamma \in \Gamma^{\leq m}$  and  $y \in \Sigma$  and

$$p(\bar{y} | \mathbf{y}) \stackrel{\text{def}}{=} p(\bar{y} | \varphi(\mathbf{y})), \quad (14)$$

for  $\mathbf{y} \in \Sigma^*$  and  $\bar{y} \in \bar{\Sigma}$ .

**Definition 3.6.** A deterministic BPDA is **representation-compatible** if there exists a matrix  $\mathbf{E} \in \mathbb{R}^{|\bar{\Sigma}| \times mG}$  and a vector  $\mathbf{u} \in \mathbb{R}^{|\bar{\Sigma}|}$  such that, for every  $\gamma \in \Gamma^{\leq m}$ , it holds for all  $\bar{y} \in \bar{\Sigma}$  that

$$\log p(\bar{y} | \gamma) = \text{softmax}(\mathbf{E}\chi(\gamma) + \mathbf{u})_{\bar{y}}. \quad (15)$$

## 4 Efficiently Representing BPDAs

The introduced technical machinery allows us to present our main result.

**Theorem 4.1.** Let  $\mathcal{P} = (\Sigma, \Gamma, m, \mu, \lambda, \rho)$  be a deterministic representation-compatible (cf. Def. 3.6) BPDA where

$$\mu(\gamma, y, \gamma') = \begin{cases} \omega(\gamma, y) & \text{if } \gamma' = \alpha(\zeta(\gamma, y)) \\ 0 & \text{otherwise} \end{cases} \quad (16)$$

for a  $\Sigma$ -determined function  $\alpha$  (cf. Def. 3.5) and a  $K$ -varied function (cf. Def. 3.4)  $\zeta$  where all  $\zeta_k$  are stack-affine (cf. Def. 3.3). Then,  $\mathcal{P}$  is  $K$ -efficiently representable (cf. Def. 3.2).

This generalizes Hewitt et al.’s (2020) result by considering the more general class of BPDA LMs, which includes  $D(b, n)$  LMs as a special case, and by incorporating the probabilistic nature of LMs. The requirement for the BPDA to be representation compatible is crucial for Thm. 4.1; there exist BPDAs that fulfill all the criteria of Thm. 4.1 but the one on representation compatibility that are not efficiently representable by RNN LMs.

**Theorem 4.2.** There exist deterministic BPDAs whose transition function  $\mu$  conforms to the structure in Eq. (16), but which are not efficiently representable for any  $K$  independent of  $|\Sigma|$ .

*Proof intuition.* Consider a BPDA where  $\Gamma = \Sigma$ .<sup>9</sup> Intuitively, this holds because a BPDA LM defines exponentially many possible (independently specified) conditional distributions over the next symbol given the current stack configuration. For an RNN LM to be weakly equivalent, it would need to, after affinely transforming the hidden state (applying the output matrix and the output bias vector), match the logits of these distributions. However, as the logits can in general span the entire  $|\Sigma|$ -dimensional space, the hidden state  $\mathbf{h}$  would have to be of size  $\Omega(|\Sigma|)$  to be able to do that, rather than  $Km \log_2 |\Sigma|$ , as we have in the case of efficient simulation. See App. C for details. ■

### 4.1 Known Efficiently Representable Families

Thm. 4.1 offers a very abstract characterization of (the components of) efficiently representable BPDAs. Here, we frame this characterization in the context of two well-known LM families: bounded Dyck and  $n$ -gram LMs.

**Proposition 4.1.** Representation-compatible LMs over  $D(b, n)$  are 2-efficiently representable with  $m = n$  and  $\Gamma = \{\langle_i | i \in [b]\}$ .

*Proof.* LMs over  $D(b, n)$  define distributions over strings of well-nested parentheses of  $b$  up to depth  $n$ . They work over the alphabet  $\Sigma = \{\langle_i | i \in [b]\} \cup \{\rangle_i | i \in [b]\}$ . A BPDA modeling an LM over  $D(b, n)$  only has to define POP and PUSH operations. Specifically,  $\langle_i$  corresponds to a push operation  $\text{PUSH}(\gamma, \langle_i, \gamma_{\langle_i})$  while  $\rangle_i$  corresponds to a pop operation  $\text{POP}(\gamma_{\langle_i}, \rangle_i, \gamma)$  for all  $i \in [b]$ .<sup>10</sup> Since popping can be performed by shifting all entries

<sup>9</sup>Examples of such BPDAs are those defining  $n$ -gram LMs.

<sup>10</sup>This is true because the input-to-stack-symbol function  $s$  here is taken to be the identity function.

on the stack one position up while pushing can be performed by shifting all entries one position down (before inserting a new symbol), the BPDA modeling a  $D(b, n)$  language is stack-affine and 2-varied with the partition  $\Sigma = \Sigma_{\text{POP}} \sqcup \Sigma_{\text{PUSH}}$  where  $\Sigma_{\text{POP}} = \{\rangle_i \mid i \in [b]\}$ ,  $\Sigma_{\text{PUSH}} = \{\langle_i \mid i \in [b]\}$ . Additionally, since each input symbol modifies the stack in a deterministic way—either it is discarded (in case of popping) or is added to the top position of the stack (in case of pushing)—the BPDA is  $\Sigma$ -determined. Moreover, since the stack only ever needs to store the  $b$  opening brackets, only  $\lceil \log_2(b+1) \rceil$  bits are needed to represent each entry. Altogether, this means that  $D(b, n)$  LMs are efficiently representable by RNN LMs of size  $D = 2n \lceil \log_2(b+1) \rceil$ . ■

Note that the construction presented by Hewitt et al. (2020) results in RNNs of size  $2 \cdot 2n \lceil \log_2(b+1) \rceil$ . This is because they rely on stack representations that contain the *complements* of the symbol encodings. This is not strictly required, which is interesting since Hewitt et al. (2020) note a difference in the constant factor between Elman RNNs and LSTMs. Our construction does away with this difference; see App. C for details.

**Proposition 4.2.** *Representation-compatible  $n$ -gram LMs are 1-efficiently representable with  $m = n - 1$  and  $\Gamma = \Sigma$ .*

*Proof.* Representing an  $n$ -gram LM requires computing the probability of every symbol given the previous  $n - 1$  symbols.<sup>11</sup> This can be performed by, at step  $t$ , (1) storing the previous  $n - 1$  symbols in the bounded stack of size  $m = n - 1$  and (2) checking the probability of the symbol  $y_t$  given the  $n - 1$  stored symbols. The required updates to the bounded stack can easily be performed by the stack-affine operation of shifting all symbols one position downward. Since *all* symbols perform that action, the function is 1-varied. Furthermore, since each input symbol induces the same update to the stack—the insertion of the symbol at the top of the stack—the BPDA is  $\Sigma$ -determined. This makes  $n$ -gram LMs efficiently representable by RNN LMs of size  $D = (n - 1) \lceil \log_2(|\Sigma| + 1) \rceil$ . ■

## 5 Discussion

This work was motivated by the question of what classes of LMs beyond those over  $D(b, n)$  can be

efficiently represented by RNN LMs, a generalization of an open question posed by Hewitt et al. (2020). We address this with Thm. 4.1, whose implications we discuss next.

**Analyzing LMs with general models of computation.** This work puts the results by Hewitt et al. (2020) in a broader context of studying the representational capacity of RNN LMs not limited to human language phenomena. Restricting the analysis to isolated phenomena might not provide a holistic understanding of the model, its capabilities, and the upper bounds of its representational capacity. General models of computation provide a framework for such holistic analysis; besides being able to provide concrete lower and upper bounds on the (efficient) representational capacity, the thorough understanding of their relationship to human language also provides apt insights into the model’s linguistic capabilities. For example, the simulation of  $n$ -gram LMs shows that RNNs can efficiently represent representation-compatible *strictly local* LMs (Jäger and Rogers, 2012), a simple and well-understood class of LMs. This provides a concrete (albeit loose) lower bound on the efficient representational capacity of RNNs. Further, we directly study the efficient *probabilistic* representational capacity of RNN LMs rather than the binary acceptance of strings. This allows for a more natural and immediate connection between the inherently probabilistic neural LMs and probabilistic formal models of computation such as BPDAs.

**Inductive biases of RNN LMs.** Understanding neural LMs in terms of the classes of LMs they can efficiently represent allows us to reason about their inductive biases. When identifying the best model to explain the data, we suspect an RNN would prefer to learn simple representations that still effectively capture the underlying patterns, rather than more complex ones. We note that this is a form of inductive bias inherent to the model architecture; we do not address other defining aspects of inductive biases, such as the learning procedure itself. Understanding such inductive biases can then, as argued by Hewitt et al. (2020), lead to architectural improvements. One can, for example, design better architectures or training procedures that exploit these inductive biases, enforce them, or loosen them if they are too restrictive. In this light, Thm. 4.1 substantiates that there is nothing inherent in RNN LMs that biases them towards hierarchi-

<sup>11</sup>The prefixes at the beginning of the string have to be appropriately padded.

cal languages since non-hierarchical ones can be modeled just as efficiently. Encouraging RNNs to learn and model cognitively plausible mechanisms for modeling language might, therefore, require us to augment them with additional mechanisms that *explicitly* model hierarchical structure, such as a stack (DuSell and Chiang, 2023). Thm. 4.1 also indicates that hierarchical languages might not be the most appropriate playground for studying the inductive biases of RNNs. For example, isolating the hidden state recurrence to the POP and PUSH operations disregards that, unlike a stack, an RNN can look at and modify the *entire* hidden state when performing the update; POP and PUSH operations are limited to modifying of the top of the stack.

**Inductive biases and learnability.** A defining aspect of an inductive bias is its effect on the *learning* behavior. While we do not discuss the learnability of bounded stack LMs by RNN LMs, we underscore the importance of this factor for a full understanding of inductive biases. Both theoretical and empirical insights are required; they provide an exciting avenue for future work, one which we see as deserving of its own treatment. More broadly, the inductive biases of a particular neural model rely on various aspects beyond the architecture, including the learning objective, the training algorithm, and features of the training data such as its ordering and size. With this in mind, we note that our results in no way suggest that RNNs are any *worse* at modeling hierarchical languages than non-hierarchical ones. The results merely provide a first step towards a more thorough understanding of the inductive biases and suggest that studies striving to understand RNNs’ learning behavior should look beyond hierarchical languages.

**On the connection to human language.** Our paper was motivated by the question of whether RNNs can efficiently represent languages beyond  $D(b, n)$ , which models hierarchical structures prominent in human language. Interestingly, our exploration of a broader class of efficiently representable LMs revealed that  $n$ -gram LMs, another class of models useful for studying human language processing, are also efficiently representable (Bickel et al., 2005; Shain et al., 2020; Wang et al., 2024). This suggests a compelling interpretation: While RNNs may not naturally prefer hierarchical languages, our extended framework associates them with a more extensive range of

human-relevant languages. This opens new avenues for understanding the inductive biases of RNNs and their connection to human language, encouraging further theoretical and empirical studies on the specific facets of human language that efficiently representable BPDAs can represent and that RNNs are adept at modeling.

#### **A new interpretation of Elman RNN recurrence.**

As detailed by the construction in the proof of Thm. 4.1, the core principle behind the efficient representation lies in the utilization of different parts of the hidden state as placeholders for relevant symbols that have occurred in the context; see also Fig. 1. This suggests that a natural interpretation of the languages efficiently representable by Elman RNNs: Those recognized by an automaton that keeps a memory of a fixed number of elements that have occurred in the string so far.

#### **Relation to other classes of formal languages.**

Striving to make the results as general as possible motivates the connection of bounded stack languages to other well-understood classes of languages. Since bounded stack languages are a particular class of finite-state languages, a natural question is how they relate to the existing **sub-regular** languages. Those have in the past been connected to various aspects of human language (Jäger and Rogers, 2012) and the representational power of convolutional neural LMs (Merrill, 2019) and transformers (Yao et al., 2021). At first glance, efficiently representable BPDAs do not lend themselves to a natural characterization in terms of known classes of sub-regular languages, but we plan on investigating this further in future work.

## **6 Conclusion**

We build on Hewitt et al.’s (2020) results and show that RNNs can efficiently represent a more general class of LMs than those over bounded Dyck languages. Concretely, we introduce bounded stack LMs as LMs defined by automata that keep  $m$  selected symbols that have occurred in the string so far and update the memory using simple update mechanisms. We show that instances of such LMs can be represented by RNN LMs in optimal space. This provides a step towards a more holistic grasp of the inductive biases of RNN LMs.



## Limitations

We conclude by discussing some limitations of our theoretical investigation and point out some ways these limitations can be addressed. We first touch on the universality of the result implied by Thm. 4.1. While we provide a new, more general, class of languages efficiently representable by RNNs and discuss the implications of this result, we do not provide an *exact* characterization of RNNs’ efficient representational capacity. In other words, we do not provide tight lower and upper bounds on the complexity of the languages that can be efficiently represented by RNNs. The lower bound of strictly local languages that can be encoded using parameter sharing is relatively loose ( $D(b, n)$  languages are much more complex than strictly local languages) and the upper bound remains evasive. This is because we do not characterize the relation of bounded stack languages to other classes of languages, such as sub-regular languages. Determining precise bounds is further complicated by the unavoidable fact that the efficient representational capacity of RNNs also depends on the parameterization of the specific probability distributions formal models of computation can represent, as made clear by Thm. 4.2. Establishing more concrete bounds is therefore a challenging open problem that is left for future work.

We note that all our results are also specifically tailored to Elman RNNs with the specific update rule from Eq. (3b). However, the results naturally generalize to the LSTM architecture in the same way that Hewitt et al.’s (2020) original constructions do.

## Ethics Statement

The paper provides a way to theoretically analyze language models. To the best of the authors’ knowledge, this paper has no ethical implications.

## Acknowledgements

Ryan Cotterell acknowledges support from the Swiss National Science Foundation (SNSF) as part of the “The Forgotten Role of Inductive Bias in Interpretability” project. Anej Svete is supported by the ETH AI Center Doctoral Fellowship. We thank the reviewers for their insightful comments and suggestions.

## References

- Steven Abney, David McAllester, and Fernando Pereira. 1999. [Relating probabilistic grammars and automata](#). In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 542–549, College Park, Maryland, USA. Association for Computational Linguistics.
- Steffen Bickel, Peter Haider, and Tobias Scheffer. 2005. [Predicting sentences using n-gram language models](#). In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 193–200, Vancouver, British Columbia, Canada. Association for Computational Linguistics.
- Nadav Borenstein, Anej Svete, Robin Shing Moon Chan, Josef Valvoda, Franz Nowak, Isabelle Augenstein, Eleanor Chodroff, and Ryan Cotterell. 2024. What languages are easy to language-model? a perspective from learning probabilistic regular languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand. Association for Computational Linguistics.
- Haw-Shiuan Chang and Andrew McCallum. 2022. [Softmax bottleneck makes language models unable to represent multi-mode word distributions](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8048–8073, Dublin, Ireland. Association for Computational Linguistics.
- Ryan Cotterell, Anej Svete, Clara Meister, Tianyu Liu, and Li Du. 2024. [Formal aspects of language modeling](#). *arXiv preprint 2311.04329*.
- A. K. Dewdney. 1977. [Threshold matrices and the state assignment problem for neural nets](#). In *Proceedings of the 8th SouthEastern Conference on Combinatorics, Graph Theory and Computing*, pages 227–245, Baton Rouge, La, USA.
- Li Du, Lucas Torroba Hennigen, Tiago Pimentel, Clara Meister, Jason Eisner, and Ryan Cotterell. 2023. [A measure-theoretic characterization of tight language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9744–9770, Toronto, Canada. Association for Computational Linguistics.
- Brian DuSell and David Chiang. 2023. [The surprising computational power of nondeterministic stack RNNs](#). *arXiv preprint 2210.01343*.
- Jeffrey L. Elman. 1990. [Finding structure in time](#). *Cognitive Science*, 14(2):179–211.
- 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, Michael Hahn, Surya Ganguli, Percy Liang, and Christopher D. Manning. 2020. [RNNs can generate bounded hierarchical languages with optimal memory](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1978–2010, Online. Association for Computational Linguistics.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. [Long short-term memory](#). *Neural Computation*, 9(8):1735–1780.
- P. Indyk. 1995. [Optimal simulation of automata by neural nets](#). In *STACS 95*, pages 337–348, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Gerhard Jäger and James Rogers. 2012. [Formal language theory: Refining the Chomsky hierarchy](#). *Philos Trans R Soc Lond B Biol Sci*, 367(1598):1956–1970.
- S. C. Kleene. 1956. [Representation of events in nerve nets and finite automata](#). In C. E. Shannon and J. McCarthy, editors, *Automata Studies. (AM-34), Volume 34*, pages 3–42. Princeton University Press, Princeton.
- Samuel A. Korsky and Robert C. Berwick. 2019. [On the computational power of RNNs](#). *CoRR*, abs/1906.06349.
- Warren S. McCulloch and Walter Pitts. 1943. [A logical calculus of the ideas immanent in nervous activity](#). *The bulletin of mathematical biophysics*, 5(4):115–133.
- 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. 2023. [Formal languages and the NLP black box](#). In *Developments in Language Theory: 27th International Conference, DLT 2023, Umeå, Sweden, June 12–16, 2023, Proceedings*, page 1–8, Berlin, Heidelberg. Springer-Verlag.
- William Merrill, Ashish Sabharwal, and Noah A. Smith. 2022. [Saturated transformers are constant-depth threshold circuits](#). *Transactions of the Association for Computational Linguistics*, 10:843–856.
- William Merrill and Nikolaos Tsilivis. 2022. [Extracting finite automata from RNNs using state merging](#). *arXiv preprint arXiv:2201.12451*.
- 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*, pages 443–459, Online. Association for Computational Linguistics.
- Marvin Lee Minsky. 1954. *Neural Nets and the Brain Model Problem*. Ph.D. thesis, Princeton University.
- Antonio Orvieto, Samuel L Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. 2023. [Resurrecting recurrent neural networks for long sequences](#).
- Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, Kranthi Kiran GV, Xuzheng He, Haowen Hou, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartłomiej Koptyra, Hayden Lau, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Xiangru Tang, Bolun Wang, Johan S. Wind, Stanislaw Wozniak, Ruichong Zhang, Zhenyuan Zhang, Qihang Zhao, Peng Zhou, Jian Zhu, and Rui-Jie Zhu. 2023. [RWKV: Reinventing RNNs for the transformer era](#). *arXiv preprint arXiv:2305.13048*.
- XiPeng Qiu, TianXiang Sun, YiGe Xu, YunFan Shao, Ning Dai, and XuanJing Huang. 2020. [Pre-trained models for natural language processing: A survey](#). *Science China Technological Sciences*, 63(10):1872–1897.
- Cory Shain, Idan Asher Blank, Marten van Schijndel, William Schuler, and Evelina Fedorenko. 2020. [fmri reveals language-specific predictive coding during naturalistic sentence comprehension](#). *Neuropsychologia*, 138:107307.
- 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.
- Lena Strobl, William Merrill, Gail Weiss, David Chiang, and Dana Angluin. 2023. [Transformers as recognizers of formal languages: A survey on expressivity](#). *arXiv preprint arXiv:2311.00208*.
- Anej Svete and Ryan Cotterell. 2023a. [Efficiently representing finite-state automata with recurrent neural networks](#). *arXiv preprint arXiv:2310.05161v3*.
- Anej Svete and Ryan Cotterell. 2023b. [Recurrent neural language models as probabilistic finite-state automata](#). *arXiv preprint arXiv:2310.05161*.
- Anej Svete, Franz Nowak, Anisha Mohamed Sahabdeen, and Ryan Cotterell. 2024. [Lower bounds on the expressivity of recurrent neural language models](#).
- Shaonan Wang, Jingyuan Sun, Yunhao Zhang, Nan Lin, Marie-Francine Moens, and Chengqing Zong. 2024. [Computational models to study language processing in the human brain: A survey](#).
- Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W. Cohen. 2018. [Breaking the softmax bottleneck: A high-rank RNN language model](#). In *International Conference on Learning Representations*.
- Shunyu Yao, Binghui Peng, Christos Papadimitriou, and Karthik Narasimhan. 2021. [Self-attention networks can process bounded hierarchical languages](#).

In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3770–3785, Online. Association for Computational Linguistics.

Wangchunshu Zhou, Yuchen Eleanor Jiang, Peng Cui, Tiannan Wang, Zhenxin Xiao, Yifan Hou, Ryan Cotterell, and Mrinmaya Sachan. 2023. [RecurrentGPT: Interactive generation of \(arbitrarily\) long text](#). *arXiv preprint arXiv:2305.13304*.

## A Related Work

This work is part of the ongoing effort to better apprehend the theoretical representational capacity of LMs, particularly those implemented with RNNs (Merrill, 2023). The study of the representational capacity of RNNs has a long history (see, e.g., McCulloch and Pitts, 1943; Minsky, 1954; Kleene, 1956; Siegelmann and Sontag, 1992; Hao et al., 2018; Korsky and Berwick, 2019; Merrill, 2019; Merrill et al., 2020; Hewitt et al., 2020; Merrill et al., 2022; Merrill and Tsilivis, 2022; Svete and Cotterell, 2023b; Svete et al., 2024, *inter alia*). Minsky (1954), for example, showed that binary-activated RNNs are equivalent to (deterministic) finite-state automata: They can represent any finite-state language and can be simulated by deterministic finite-state automata. Minsky’s construction of an RNN simulating a general deterministic FSA was extended to the probabilistic—language modeling—setting by Svete and Cotterell (2023b).<sup>12</sup>

Recently, increased interest has been put on the *efficient* representational capacity of LMs and their inductive biases. To emulate a deterministic probabilistic FSA  $\mathcal{A}$  with states  $Q$  over the alphabet  $\Sigma$ ,<sup>13</sup> Minsky’s (1954) construction requires an RNN of size  $\mathcal{O}(|\Sigma||Q|)$ . This was improved by Dewdney (1977), who established that a general FSA can be simulated by an RNN of size  $\mathcal{O}\left(|\Sigma||Q|^{3/4}\right)$ , and further by Indyk (1995), who lowered this bound to  $\mathcal{O}\left(|\Sigma|\sqrt{|Q|}\right)$ . The latter was also shown to be optimal for general (adversarial) FSAs. The starting point of this work, Hewitt et al. (2020), was the first to show the possibility of exponentially compressing a specific family of finite-state languages, namely the bounded Dyck languages. This provides an exponential improvement over the general FSA simulation results. In this work, we generalize this result to a more general class of languages, aiming to gain a more thorough understanding of the mechanisms that allow exponential compression with RNNs. As shown by Thm. 4.1, a possible candidate for the mechanism enabling exponential compression might be the notion of a bounded stack.

### A.1 Relation to Hewitt et al.’s (2020) Notion of Language Recognition

A crucial difference in our approach to that of Hewitt et al. (2020) is the direct treatment of *language models* rather than binary languages. The notion of recognition of binary languages by RNNs (or any other neural LM) can be somewhat tricky, since the LM assigns a probability to each string in the language, not just a binary decision of whether the string is in the language. To reconcile the discrepancy between the discrete nature of formal languages and the probabilistic nature of (RNN) LMs, Hewitt et al. (2020) define the so-called truncated recognition of a formal language.

**Definition A.1.** Let  $\Sigma$  be an alphabet and  $\alpha \in \mathbb{R}_+$ . An LM  $p$  is said to **recognize** the language  $\mathcal{L} \subseteq \Sigma^*$  with the **threshold**  $\alpha$  if for every  $\mathbf{y} = y_1 \dots y_T \in \mathcal{L}$ , it holds for all  $t \in [T]$  that

$$p(y_t \mid \mathbf{y}_{<t}) > \alpha \quad \text{and} \quad p(\text{EOS} \mid \mathbf{y}) > \alpha. \quad (17)$$

In words, a language  $\mathcal{L}$  is recognized by  $p$  if  $p$  assigns sufficiently high probability to all allowed continuations of (sub)strings in  $\mathcal{L}$ .<sup>14</sup>

The notion of truncated recognition allows Hewitt et al. (2020) to talk about the recognition of binary languages by RNNs and design RNNs efficiently representing bounded Dyck languages. In contrast to our results, they do not consider the exact probabilities in their constructions—they simply ensure that the probabilities of valid continuations in their constructions are bounded from below by a constant and those of invalid continuations are bounded from above by a constant. While this results in interesting insights and reusable mechanisms, as we showcase here, it is not clear how to generalize this approach to languages more general than the structured Dyck languages. Taking into account the exact probabilities

<sup>12</sup>Unlike binary FSAs, where non-determinism does not add any expressive power, non-deterministic probabilistic FSAs are more expressive than their deterministic counterparts. This is why, although Minsky’s (1954) construction only considers deterministic FSAs, it shows the equivalence of RNNs to all FSAs. The same cannot be said about the probabilistic setting, where the equivalence holds only for the deterministic case. The relationship between RNN LMs and general probabilistic FSAs is addressed in Svete et al. (2024).

<sup>13</sup>See App. B.1 for a formal definition of a PFSA.

<sup>14</sup>In contrast to  $p(\mathbf{y})$ , which necessarily diminishes with the string length, the conditional probabilities of valid continuations are bounded from below.



allows us to consider a more general class of languages, for example,  $n$ -gram LMs. When applicable, thresholding the exact probabilities from our construction in the manner described by Def. A.1 allows us to reconstruct the construction of Hewitt et al. (2020), meaning that our results provide a convenient generalization of theirs.

## B Finite-state Automata and Bounded Stack Languages

The main part of the paper discussed the connection of LMs representable by BPDAs and RNNs. BPDA LMs, however, are a particular class of finite-state LMs, and we discuss this connection in more detail here.

### B.1 Probabilistic Finite-state Automata

We begin by more formally defining the notion of probabilistic finite-state automata (PFSAs). Probabilistic finite-state automata are a well-understood real-time computational model.

**Definition B.1.** A *probabilistic finite-state automaton* (PFSA) is a 5-tuple  $(\Sigma, Q, \delta, \lambda, \rho)$  where  $\Sigma$  is an alphabet,  $Q$  is a finite set of states,  $\delta \subseteq Q \times \Sigma \times \mathbb{R}_{\geq 0} \times Q$  is a finite set of weighted transitions where we write transitions  $(q, y, w, q') \in \delta$  as  $q \xrightarrow{y/w} q'$ ,<sup>15</sup> and  $\lambda, \rho: Q \rightarrow \mathbb{R}_{\geq 0}$  are functions that assign each state its initial and final weight, respectively. Moreover, for all states  $q \in Q$ ,  $\delta, \lambda$  and  $\rho$  satisfy  $\sum_{q \in Q} \lambda(q) = 1$ , and  $\sum_{q \xrightarrow{y/w} q' \in \delta} w + \rho(q) = 1$ .

We next define some basic concepts. A PFSA  $\mathcal{A} = (\Sigma, Q, \delta, \lambda, \rho)$  is **deterministic** if  $|\{q \mid \lambda(q) > 0\}| = 1$  and, for every  $q \in Q, y \in \Sigma$ , there is at most one  $q' \in Q$  such that  $q \xrightarrow{y/w} q' \in \delta$  with  $w > 0$ . Any state  $q$  where  $\lambda(q) > 0$  is called an **initial state**, and if  $\rho(q) > 0$ , it is called a **final state**. A **path**  $\pi$  of length  $N$  is a sequence of subsequent transitions in  $\mathcal{A}$ , denoted as

$$q_1 \xrightarrow{y_1/w_1} q_2 \xrightarrow{y_2/w_2} q_3 \cdots q_N \xrightarrow{y_N/w_N} q_{N+1}. \quad (18)$$

The **yield** of a path is  $\mathbf{s}(\pi) \stackrel{\text{def}}{=} y_1 \dots y_N$ . The **prefix weight**  $\tilde{w}$  of a path  $\pi$  is the product of the transition and initial weights, whereas the **weight** of a path additionally has the final weight multiplied in. In symbols, this means

$$\tilde{w}(\pi) \stackrel{\text{def}}{=} \prod_{n=0}^N w_n, \quad (19) \quad w(\pi) \stackrel{\text{def}}{=} \prod_{n=0}^{N+1} w_n, \quad (20)$$

with  $w_0 \stackrel{\text{def}}{=} \lambda(q_1)$  and  $w_{N+1} \stackrel{\text{def}}{=} \rho(q_{N+1})$ . We write  $\Pi(\mathcal{A})$  for the set of all paths in  $\mathcal{A}$  and we write  $\Pi(\mathcal{A}, \mathbf{y})$  for the set of all paths in  $\mathcal{A}$  with yield  $\mathbf{y}$ . The sum of weights of all paths that yield a certain string  $\mathbf{y} \in \Sigma^*$  is called the **stringsum**, given in the notation below

$$\mathcal{A}(\mathbf{y}) \stackrel{\text{def}}{=} \sum_{\pi \in \Pi(\mathcal{A}, \mathbf{y})} w(\pi). \quad (21)$$

The stringsum gives the probability of the string  $\mathbf{y}$ .

### B.2 Bounded-stack Probabilistic Finite-state Automata

The main text presented bounded stack LMs as LMs defined by BPDAs. While the specification with respect to bounded stacks is enough to connect them to RNN LMs, we can also investigate weakly equivalent PFSAs, establishing an explicit connection to this well-studied class of computational models. This relationship is characterized by the following theorem.

**Theorem 3.1.** *The family of LMs induced by PFSAs is weakly equivalent to the family of LMs induced by BPDAs.*

*Proof.* ( $\Leftarrow$ ). It is easy to see that any BPDA defines a PFSA: Since the set of possible bounded stack configurations is finite and  $\mu$  simply defines transitions between the finitely many configurations,

<sup>15</sup>We further assume a  $(q, y, q')$  triple appears in at most one element of  $\delta$ .

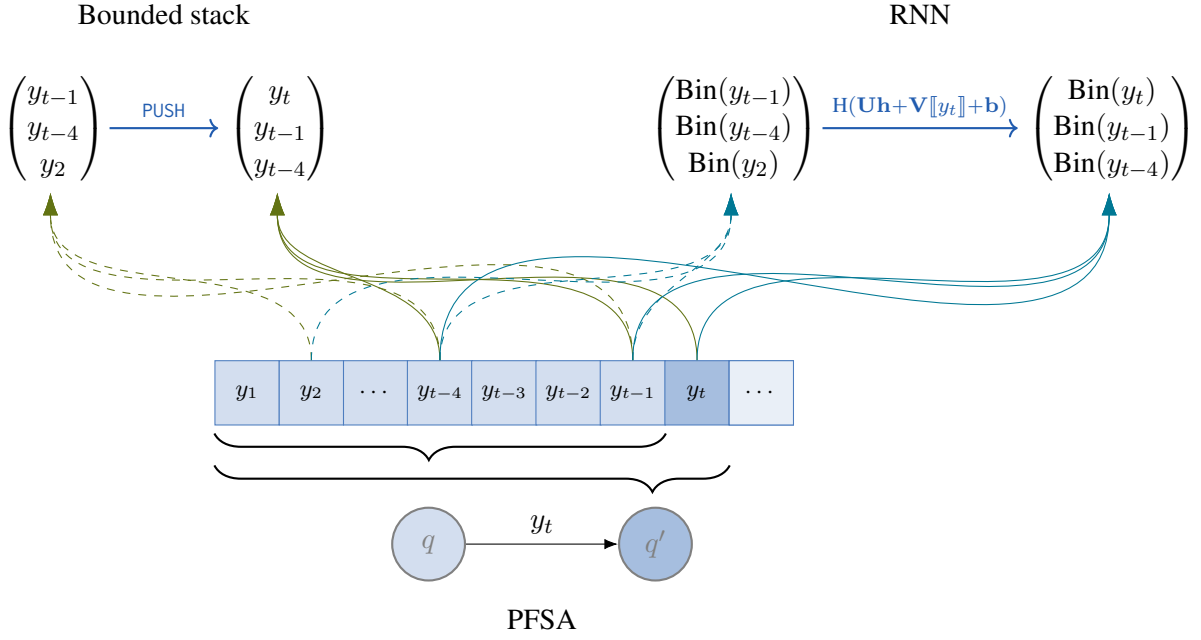


Figure 5: An illustration of how one can think of BPDA LMs as being represented by three different mechanisms: a BPDA, a black-box PFSA, and an RNN.

one can think of the BPDA as defining transitions between the finitely many states represented by the configurations.

( $\implies$ ). Let  $\mathcal{A} = (\Sigma, Q, \delta, \lambda, \rho)$  be a PFSA. We define the weakly equivalent BPDA  $\mathcal{P} = (\Sigma, \Gamma, m, \mu, \lambda_{\mathcal{P}}, \rho_{\mathcal{P}})$  with  $\Gamma = Q$ ,  $m = 1$ ,  $\lambda_{\mathcal{P}} = \lambda$ ,  $\rho_{\mathcal{P}} = \rho$ ,  $\mu(q, y, q') = \delta(q, y, q')$  for all  $q, q' \in Q$  and  $y \in \Sigma$ . By noting that there is a trivial bijection between the stack configurations and the states of the PFSA as well as between the two transition functions, it is easy to see that the BPDA  $\mathcal{P}$  is weakly equivalent<sup>16</sup> to the PFSA  $\mathcal{A}$ . ■

The interpretation of BPDA LMs in terms of three mechanisms—a stack modification function, an RNN, and a bounded-stack PFSA—is illustrated in Fig. 5, which shows the three update mechanisms in action on the same string.

## C Proofs

This section contains the proofs of the theorems stated and used in the paper.

### C.1 Emulating Stack Updates

This section contains the proof of Thm. 4.1. In many ways, it resembles the original exposition by Hewitt et al. (2020) but is presented in our notation and with the additional generality of BPDAs. Some aspects are also simplified in our framework (for example, we do not have to scale the inputs to the activation functions).

**On the use of the Heaviside activation function.** The proof below relies on the use of the Heaviside activation function. As mentioned in §2.1, due to the relationship between ReLU and H, the same results hold for ReLU-activated RNNs as well. More precisely, it is easy to show that

$$H(i) = \text{ReLU}(i) - \text{ReLU}(i - 1) \quad (22)$$

for all  $i \in \mathbb{Z}$ .<sup>17</sup> All our results therefore map to the setting of ReLU-activated RNNs, but might require hidden states of twice the size to store the results of  $\text{ReLU}(i)$  and  $\text{ReLU}(i - 1)$ .<sup>18</sup> Interestingly, the

<sup>16</sup>In fact, one could also prove strong equivalence.

<sup>17</sup>The restriction to the integers is enough since we are considering finite-precision RNNs.

<sup>18</sup>These two values can then be combined at the next step of the computation before being used as  $H(i)$ .

duplication of the size of the hidden state is *not* required for the same reason as in Hewitt et al. (2020); they use hidden states of size  $2KmG$  (in our notation) to store the logarithmic encodings and their *complements*. This is required since it allows for an easier explicit determination of next-symbol probabilities. We do not require that. Notice that the binary complement  $\bar{\mathbf{x}}$  of a vector  $\mathbf{x} \in \{0, 1\}^D$  is its affine transformation  $\bar{\mathbf{x}} = \mathbf{1}_D - \mathbf{x}$ . Thus, the transformation  $\mathbf{x} \mapsto \bar{\mathbf{x}}$  can be absorbed into the affine transformation  $\mathbf{E}\mathbf{x} + \mathbf{u}$ .<sup>19</sup>

**Theorem 4.1.** *Let  $\mathcal{P} = (\Sigma, \Gamma, m, \mu, \lambda, \rho)$  be a deterministic representation-compatible (cf. Def. 3.6) BPDA where*

$$\mu(\gamma, y, \gamma') = \begin{cases} \omega(\gamma, y) & \text{if } \gamma' = \alpha(\zeta(\gamma, y)) \\ 0 & \text{otherwise} \end{cases} \quad (16)$$

*for a  $\Sigma$ -determined function  $\alpha$  (cf. Def. 3.5) and a  $K$ -varied function (cf. Def. 3.4)  $\zeta$  where all  $\zeta_k$  are stack-affine (cf. Def. 3.3). Then,  $\mathcal{P}$  is  $K$ -efficiently representable (cf. Def. 3.2).*

*Proof.* The proof of Thm. 4.1 requires us to show that the RNN recurrence (cf. Eq. (3b)) can (i) implement the update mechanisms implementing the stack update mechanism of the BPDA and (ii) encode the same next-symbol probabilities as the BPDA. We do that by defining the appropriate recurrence and input matrices  $\mathbf{U}$  and  $\mathbf{V}$ , the bias vector  $\mathbf{b}$ , and the output parameters  $\mathbf{E}$  and  $\mathbf{u}$ . We also use one-hot encodings of input symbols  $y$ , which we denote with  $\llbracket y \rrbracket$ . More precisely, for a bijection  $h: \Sigma \rightarrow [\Sigma]$ , we define the entries of  $\llbracket y \rrbracket \in \{0, 1\}^{|\Sigma|}$  as

$$\llbracket y \rrbracket_i \stackrel{\text{def}}{=} \mathbb{1}\{i = h(y)\} \quad (23)$$

for  $i \in [\Sigma]$

The condition Eq. (16) can be equivalently expressed as the requirement that (the deterministic)  $\mathcal{P}$  defines the next-stack configuration function  $\phi$  of the form

$$\phi(\gamma, y) = \alpha(\zeta(\gamma, y), y) \quad (24)$$

for a  $\Sigma$ -determined function  $\alpha$  and a  $K$ -varied function  $\zeta$  where all  $\zeta_k$  are stack-affine. Definitionally, it thus holds for all  $\zeta_k$  that

$$\chi(\zeta_k(\gamma, y)) = \chi(\zeta_k(\gamma)) = \mathbf{M}^k \chi(\gamma) + \mathbf{v}^k \quad (25)$$

for some matrix  $\mathbf{M}^k \in \mathbb{R}^{mG \times mG}$  and  $\mathbf{v}^k \in \mathbb{R}^{mG}$ . We now define  $\kappa: \Sigma \rightarrow [K]$  as the function that maps a symbol  $y \in \Sigma$  to the index of the set  $y$  belongs to in the partition of  $\Sigma$ . That is, we define

$$\kappa(y) \stackrel{\text{def}}{=} k \quad \text{if } y \in \Sigma_k, \quad (26)$$

where  $\Sigma = \Sigma_1 \sqcup \dots \sqcup \Sigma_K$  is the partition defined by the  $K$ -varied function  $\zeta$ . Then, we have that

$$\phi(\gamma, y) = \alpha(\zeta_{\kappa(y)}(\gamma), y) = \alpha(\mathbf{M}^{\kappa(y)} \chi(\gamma) + \mathbf{v}^{\kappa(y)}, y) \quad (27)$$

This motivates the following implementation of  $\phi$  with the RNN recurrence:

1. Divide the hidden state  $\mathbf{h}$  into  $K$  copies, *one* of those containing the vectorial representation of the actual stack (cf. Eq. (11)).
2. Depending on the input symbol  $y \in \Sigma$ , perform the appropriate affine transformation  $\mathbf{h} \mapsto \mathbf{M}^{\kappa(y)} \mathbf{h} + \mathbf{v}^{\kappa(y)}$ .
3. Apply the  $\Sigma$ -determined function  $\alpha$ .

<sup>19</sup>Interestingly, the recognition of bounded Dyck languages does *not* require the duplication of the hidden state size even when we use ReLU activation function. This is because whenever the network inserts a symbol on the stack (the only place where the Heaviside function behaves differently compared to the ReLU in the proof of Thm. 4.1), the slot where the symbol is inserted is already empty. Thus, the values of the stack encoding stay  $\in \{0, 1\}$ , which means that they do not require the clipping performed by the ReLU function.

We thus define, for the (single) initial configuration of  $\mathcal{P}$ ,  $\gamma_0$ ,

$$\boldsymbol{\eta} = \mathbf{h}_0 \stackrel{\text{def}}{=} \begin{pmatrix} \chi(\gamma_0) \\ \mathbf{0}_{mG} \\ \vdots \\ \mathbf{0}_{mG} \end{pmatrix} \in \{0, 1\}^{KmG}, \quad (28)$$

which encodes the initial configuration. In general, we will write

$$\mathbf{h} = \begin{pmatrix} \mathbf{h}^1 \\ \vdots \\ \mathbf{h}^K \end{pmatrix} \in \{0, 1\}^{KmG}. \quad (29)$$

We will say that the hidden state  $\mathbf{h}$  satisfies the **single-copy invariance** if at most one of the components  $\mathbf{h}^1, \dots, \mathbf{h}^K$  is non-zero. This completes step 1.

To implement steps 2 and 3 we proceed as follows. To perform the  $K$  different stack-affine functions, we define the parameters

$$\mathbf{U} \stackrel{\text{def}}{=} \begin{pmatrix} \mathbf{M}^1 & \dots & \mathbf{M}^1 \\ \vdots & \ddots & \vdots \\ \mathbf{M}^K & \dots & \mathbf{M}^K \end{pmatrix} \in \mathbb{R}^{KmG \times KmG} \quad (30a)$$

$$\mathbf{b} \stackrel{\text{def}}{=} \begin{pmatrix} \mathbf{v}^1 \\ \vdots \\ \mathbf{v}^K \end{pmatrix} \in \mathbb{R}^{KmG}. \quad (30b)$$

We also define the input matrix  $\mathbf{V}$  as

$$\mathbf{V} \stackrel{\text{def}}{=} (\mathbf{r}(y_1) \quad \dots \quad \mathbf{r}(y_{|\Sigma|})) \in \mathbb{R}^{KmG \times |\Sigma|} \quad (31)$$

where, for  $y \in \Sigma$ , we write

$$\mathbf{r}(y) = \begin{pmatrix} \mathbf{r}(y)^1 \\ \vdots \\ \mathbf{r}(y)^K \end{pmatrix} \in \mathbb{R}^{KmG}, \quad (32)$$

and define

$$\mathbf{r}(y)^k \stackrel{\text{def}}{=} \begin{cases} \mathbf{z}(y) & \text{if } k = \kappa(y) \\ -\mathbf{1}_{mG} & \text{otherwise} \end{cases}, \quad (33)$$

where finally

$$\mathbf{z}(y)_j = \begin{cases} \mathbf{0}_G & \text{if } j \in \mathcal{J}_1^y \\ 2 \text{Bin}(s(y)) - \mathbf{1}_G & \text{if } j \in \mathcal{J}_2^y, \\ -\mathbf{1}_G & \text{if } j \in \mathcal{J}_3^y \end{cases} \quad (34)$$

for  $j \in [m]$ . Here,  $\mathbf{1}_G$  is the  $G$ -dimensional vector of ones.

We now show that the parameters defined above simulate the stack update function  $\phi$  correctly. Let  $\mathbf{y} \in \Sigma^*$ ,  $\boldsymbol{\gamma} \stackrel{\text{def}}{=} \varphi(\mathbf{y})$ , and  $y \in \Sigma$ . Furthermore, assume (by an inductive hypothesis) that  $\mathbf{h}$  satisfies the single-copy invariance and that the non-zero component of  $\mathbf{h}$  contain  $\chi(\boldsymbol{\gamma})$ . We want to show that  $\mathbf{h}' \stackrel{\text{def}}{=} \mathbf{H}(\mathbf{U}\mathbf{h} + \mathbf{V}\llbracket y \rrbracket + \mathbf{b})$  (1) satisfies the single-copy invariance, and (2) that the non-zero copy in  $\mathbf{h}'$  contains exactly the encoding  $\chi(\phi(\boldsymbol{\gamma}, y))$ . Because of the single-copy invariance of  $\mathbf{h}$  and the definition of  $\mathbf{U}$ , we see that

$$\mathbf{U}\mathbf{h} + \mathbf{b} = \begin{pmatrix} \mathbf{M}^1 \mathbf{h} + \mathbf{v}^1 \\ \vdots \\ \mathbf{M}^K \mathbf{h} + \mathbf{v}^K \end{pmatrix}. \quad (35)$$



We now note that, for  $\zeta_k$  to be a valid stack-affine function, it has to hold that  $\mathbf{M}^k \mathbf{h} + \mathbf{v}^k \in \{0, 1\}^{mG}$ . Now, let  $k \in [K]$ . We distinguish two cases:

- $k = \kappa(y)$ . Then

$$\mathbf{V}[\![y]\!] = \mathbf{r}(y)^{\kappa(y)} = \mathbf{z}(y). \quad (36)$$

This results in the entries

$$\mathbf{H} \left( \mathbf{M}^{\kappa(y)} \mathbf{h} + \mathbf{r}(y)^{\kappa(y)} + \mathbf{v}^{\kappa(y)} \right)_j = \mathbf{H} \left( \mathbf{M}^{\kappa(y)} \mathbf{h} + \mathbf{z}(y) + \mathbf{v}^{\kappa(y)} \right)_j \quad (37a)$$

$$= \mathbf{H} \left( \left( \mathbf{M}^{\kappa(y)} \mathbf{h} + \mathbf{v}^{\kappa(y)} \right)_j + \mathbf{z}(y)_j \right). \quad (37b)$$

We further consider three cases

- $j \in \mathcal{J}_1^y$ . Then

$$\mathbf{H} \left( \left( \mathbf{M}^{\kappa(y)} \mathbf{h} + \mathbf{v}^{\kappa(y)} \right)_j + \mathbf{z}(y)_j \right) = \mathbf{H} \left( \left( \mathbf{M}^{\kappa(y)} \mathbf{h} + \mathbf{v}^{\kappa(y)} \right)_j + \mathbf{0}_G \right) \quad (38a)$$

$$= \left( \mathbf{M}^{\kappa(y)} \mathbf{h} + \mathbf{v}^{\kappa(y)} \right)_j \quad (38b)$$

- $j \in \mathcal{J}_2^y$ . Then

$$\mathbf{H} \left( \left( \mathbf{M}^{\kappa(y)} \mathbf{h} + \mathbf{v}^{\kappa(y)} \right)_j + \mathbf{z}(y)_j \right) = \mathbf{H} \left( \left( \mathbf{M}^{\kappa(y)} \mathbf{h} + \mathbf{v}^{\kappa(y)} \right)_j + 2 \text{Bin}(s(y)) - \mathbf{1}_G \right) \quad (39a)$$

$$= \text{Bin}(s(y)) \quad (39b)$$

This follows from the fact that  $2 \text{Bin}(s(y)) - \mathbf{1}_G$  contains the value 1 wherever  $\text{Bin}(s(y))$  is 1 and the value  $-1$  elsewhere, masking out the entries in the vector that are not active in  $\text{Bin}(s(y))$ .

- $j \in \mathcal{J}_3^y$ . Then

$$\mathbf{H} \left( \left( \mathbf{M}^{\kappa(y)} \mathbf{h} + \mathbf{v}^{\kappa(y)} \right)_j + \mathbf{z}(y)_j \right) = \mathbf{H} \left( \left( \mathbf{M}^{\kappa(y)} \mathbf{h} + \mathbf{v}^{\kappa(y)} \right)_j - \mathbf{1}_G \right) \quad (40a)$$

$$= \mathbf{0}_G \quad (40b)$$

- $k \neq \kappa(y)$ . Then

$$\mathbf{V}[\![y]\!] = \mathbf{r}(y)^k = -\mathbf{1}_{mG}, \quad (41)$$

resulting in

$$\mathbf{M}^k \mathbf{h} + \mathbf{r}(y)^k + \mathbf{v}^k = \mathbf{M}^k \mathbf{h} - \mathbf{1}_{mG} + \mathbf{v}^k, \quad (42)$$

whose entries are  $\leq 0$ . This directly implies that

$$\mathbf{H} \left( \mathbf{M}^k \mathbf{h} + \mathbf{r}(y)^k + \mathbf{v}^k \right) = \mathbf{0}_{mG}, \quad (43)$$

masking the  $k^{\text{th}}$  component of  $\mathbf{h}'$ .

Examining Eqs. (38b), (39b) and (40b) reveals that these results cover the three conditions of  $\Sigma$ -determined function  $\alpha$ . Moreover Eq. (43) shows that the update rule preserves the single-copy invariance: All but the  $\kappa(y)^{\text{th}}$  component of the hidden state are masked to 0. Summarizing, this shows that the RNN parametrized with the parameters  $\mathbf{U}$ ,  $\mathbf{V}$ , and  $\mathbf{b}$  correctly implements the stack update function  $\phi$ .

To extend this to the probabilistic setting, we use the assumption that  $\mathcal{P}$  is representation-compatible. By definition,  $\mathcal{P}$  then defines next-symbol probabilities  $p(\bar{y} \mid \mathbf{y})$  where

$$\log p(\bar{y} \mid \varphi(\mathbf{y})) = \text{softmax}(\mathbf{E}' \chi(\varphi(\mathbf{y})) + \mathbf{u}')_{\bar{y}} \quad (44)$$

for some matrix  $\mathbf{E}' \in \mathbb{R}^{|\Sigma| \times mG}$  and  $\mathbf{u}' \in \mathbb{R}^{|\Sigma|}$ . The first part of the proof shows that  $\mathbf{h}(\mathbf{y})$  contains exactly one copy of  $\chi(\varphi(\mathbf{y}))$ . We use that fact and define

$$\mathbf{E} \stackrel{\text{def}}{=} (\mathbf{E}' \quad \dots \quad \mathbf{E}') \in \mathbb{R}^{|\Sigma| \times KmG} \quad (45a)$$

$$\mathbf{u} \stackrel{\text{def}}{=} \mathbf{u}' \in \mathbb{R}^{|\Sigma|}, \quad (45b)$$

which will result in the softmax-normalized RNN computing identical next-symbol probabilities to those computed by  $\mathcal{P}$ . This means that  $\mathcal{R}$  and  $\mathcal{P}$  are weakly equivalent. ■

## C.2 On the Impossibility of Efficiently Representing All BPDA LMs

**Theorem 4.2.** *There exist deterministic BPDAs whose transition function  $\mu$  conforms to the structure in Eq. (16), but which are not efficiently representable for any  $K$  independent of  $|\Sigma|$ .*

*Proof.* The reason behind this is intuitive—the  $|\Sigma|^m$  next-symbol probability distributions defined by a BPDA LM can be completely arbitrary and might not lend themselves to a compact parametrization with matrix multiplication (cf. Def. 3.6). This is why the proof of Thm. 4.1 relies heavily on specific families of BPDA LMs that are particularly well-suited for efficient representation by RNN LMs through parameter sharing. More formally, this is a special case of the **softmax bottleneck** (Yang et al., 2018; Chang and McCallum, 2022; Borenstein et al., 2024): The notion that the representations  $\mathbf{h}(\mathbf{y}) \in \mathbb{R}^D$  defining an LM whose conditional logits span a  $d$ -dimensional subspace of  $\mathbb{R}^{|\Sigma|}$  must be of size  $D \geq d$ . Because BPDAs can in general define full-rank distributions whose logits span  $\mathbb{R}^{|\Sigma|}$ , there exist BPDAs for which it has to hold that  $D \geq |\Sigma|$ . Any such BPDA is not efficiently representable;  $|\Sigma| \leq D \leq Cm \log_2 |\Sigma|$  would require  $C \geq \frac{|\Sigma|}{m \log_2 |\Sigma|}$ , which is not constant in  $|\Sigma|$ . ■

**Connection to the result by Hewitt et al. (2020).** The impossibility result from Thm. 4.2 of course includes distributions over bounded Dyck languages as well. That is, a general distribution over a bounded Dyck language may not be efficiently represented by an RNN LM. This does not contradict the results from Hewitt et al. (2020)—the LMs considered by Hewitt et al. (2020) form a particular family of LMs that *are* efficiently representable by Elman RNN LMs. Intuitively, this is because of two reasons:

1. In their construction, the probability of the next symbol only depends on the *top* of the stack. As such, many stack configurations (and thus RNN hidden states) result in the same next-symbol probability distribution, allowing for more efficient encoding.
2. Hewitt et al. (2020) are only interested in recognizing binary languages, which means that they only consider LMs that assign *sufficiently large* probabilities to the correct continuations of the input string. Exact probabilities under the language model are not important.