A Measure-Theoretic Characterization of Tight Language Models

Li Du[ᴅ]   **Lucas Torroba Hennigen**[�]   **Tiago Pimentel**[ð]
Clara Meister[ſ]   **Jason Eisner**[ᴅ]   **Ryan Cotterell**[ſ]
[ᴅ]Johns Hopkins University   [ᴬ]MIT
[ð]University of Cambridge   [ſ]ETH Zürich
leodu@cs.jhu.edu   lucastor@mit.edu   tp472@cam.ac.uk
clara.meister@inf.ethz.ch   jason@cs.jhu.edu   ryan.cotterell@inf.ethz.ch

Abstract

Language modeling, a central task in natural language processing, involves estimating a probability distribution over strings. In most cases, the estimated distribution sums to 1 over all finite strings. However, in some pathological cases, probability mass can "leak" onto the set of infinite sequences. In order to characterize the notion of leakage more precisely, this paper offers a measure-theoretic treatment of language modeling. We prove that many popular language model families are in fact tight, meaning that they will not leak in this sense. We also generalize characterizations of tightness proposed in previous works.

1 Introduction

Language modeling is a core task in natural language processing. As canonically defined, language modeling involves estimating a distribution over the set of strings over a given alphabet. If the alphabet is the set of words in a language,[1] then a language model can be thought of as a distribution over the language's sentences. Since Shannon (1948), language modeling has been used to estimate statistical properties of language and has become essential for computational linguistics research (Hale, 2001; Meister et al., 2021). Further, it is also central to a wide range of natural language processing applications, whether as a source model in a noisy channel architecture (Weaver, 1955; Jelinek, 1976), as a way of learning better representations of language (Peters et al., 2018), or, more recently, for prompted generation (Brown et al., 2020), where the distribution defined by a language model is employed in tasks like question-answering (Petroni et al., 2019), style transfer (Reif et al., 2022), and even sequence tagging (Liu et al., 2022).

More formally, a language model is typically defined to be a distribution over the *countably* infinite

set Σ^* of all (finite) strings (Booth and Thompson, 1973).[2] However, it has been shown that some classes of autoregressive language models have parameter settings in which the generative process terminates with probability < 1. Welleck et al. (2020) discuss this issue for recurrent neural network language models. Models whose generative process may fail to terminate are called **non-tight** (Chi, 1999, who discussed non-tight PCFGs). If an autoregressive language model is non-tight, it may generate infinite sequences and MCMC algorithms over such models will not mix to the correct distribution.

It is here that a subtlety arises: the set of infinite sequences is *uncountably* infinite. Properly treating a distribution over this sample space requires a modicum of measure theory.[3] To clarify the situation, we review the measure-theoretic treatment of distributions over infinite sequences. We then make use of a termination symbol EOS to define a random variable whose value can be a string, i.e., an element of Σ^*, or an infinite sequence. In a tight language model, this random variable has probability 1 of being a string and hence finite.

Beyond offering a measure-theoretic formalization, our paper also demonstrates how tightness relates to the Borel–Cantelli lemmata, simplifying a recent result by Meister et al. (2022). To conclude our paper, we analyze several language modeling architectures and give conditions on their tightness. We demonstrate that n-gram language models—and more generally, language models defined by stochastic finite-state automata—can be non-tight, and we give a simple necessary and sufficient condition for tightness in terms of the inverse of the automaton's transition matrix. This builds on a known

---

[1]Or perhaps alphabetic symbols or subwords; see, e.g., Bostrom and Durrett (2020).

[2]Recall that $\Sigma^* \overset{\text{def}}{=} \bigcup_n \Sigma^n$ where for $n \geq 0$, Σ^n is the set of strings of length n. The * is the Kleene closure operator.

[3]Indeed, our treatment resolves an imprecision present in the literature. For instance, Welleck et al. (2020) discusses the probability of infinite sequences despite using the canonical definition of a language model as a distribution over Σ^*.

9744

result due to Lehmann (1977). We also discuss when neural language models become non-tight. We prove that Transformer-based language models (Vaswani et al., 2017; Brown et al., 2020) are always tight and that recurrent neural language models are always tight when they employ a bounded activation function. However, we also exhibit a recurrent neural network (RNN) language model with a ReLU activation (Nair and Hinton, 2010) that is non-tight in a simpler construction than the one offered by Chen et al. (2018). As a byproduct, we also generalize and strengthen the results given by Welleck et al. (2020), who give a sufficient condition for tightness of recurrent neural language models in terms of the norm of the hidden state.

2 Motivating Examples

Let Σ be an alphabet, i.e., a finite set of symbols, and let EOS $\notin \Sigma$ be a distinguished end-of-sequence symbol. Let $\overline{\Sigma} \overset{\text{def}}{=} \Sigma \cup \{\text{EOS}\}$. A **string** of length $n \geq 0$ is a finite sequence $\boldsymbol{x} = x_1 x_2 \ldots x_n$ where each $x_t \in \Sigma$. By convention, we say that $x_{n+1} = \text{EOS}$, although x_{n+1} is not an element of the sequence $\boldsymbol{x}$. For any integer $1 \leq t \leq n+1$, we write $\boldsymbol{x}_{<t}$ for the prefix $x_1 x_2 \cdots x_{t-1}$.

We now begin to distinguish between "language models" and "sequence models." As is traditional in the NLP literature, we henceforth define a **language model** to be a probability distribution over the countable set Σ^* of all strings (see Def. 3.4). It is popular to specify such a distribution in terms of its conditional probabilities $\bar{p}(x_t \mid \boldsymbol{x}_{<t})$.

Definition 2.1. *An **autoregressive sequence model** (ASM) is a conditional probability distribution $\bar{p}(x_t \mid \boldsymbol{x}_{<t})$ where $x_t \in \overline{\Sigma}$ and $\boldsymbol{x}_{<t} \in \overline{\Sigma}^*$.*

If $\bar{p}$ is an ASM, then we define a non-negative function p over Σ^* by $p(\boldsymbol{x}) \overset{\text{def}}{=} \prod_{t=1}^{n+1} \bar{p}(x_t \mid \boldsymbol{x}_{<t}) = \bar{p}(\text{EOS} \mid \boldsymbol{x}) \prod_{t=1}^{n} \bar{p}(x_t \mid \boldsymbol{x}_{<t})$, where n denotes the length of $\boldsymbol{x}$.

But is p a language model? Not always, since as we will see below, it is possible for $p(\Sigma^*) \overset{\text{def}}{=} \sum_{\boldsymbol{x} \in \Sigma^*} p(\boldsymbol{x}) < 1$. Of course this "bad" case never happens if the ASM's conditional probabilities are derived from a known LM, in which case p simply recovers that LM.[4] In this case clearly $p(\Sigma^*) = 1$.

It follows that if $p(\Sigma^*) < 1$, then the ASM's conditional probabilities do *not* match the conditional probabilities of any language model p_0.

We now exhibit such a "bad" ASM. Although the conditional probability distributions $\bar{p}(\cdot \mid \boldsymbol{x}_{<t})$ each sum to 1 over $\overline{\Sigma}$, they fail to combine into a model p that sums to 1 over Σ^* (i.e., a language model).

Example 2.2 (non-tight bigram model). Consider the bigram model defined in Fig. 1a over the alphabet $\Sigma = \{a, b\}$. Under this model, any finite string that contains the symbol b will have probability 0, since $\bar{p}(\text{EOS} \mid b) = \bar{p}(a \mid b) = 0$. This implies $p(\Sigma^*) = \sum_{n=0}^{\infty} p(a^n) = \sum_{n=0}^{\infty} (0.7)^n \cdot 0.1 = \frac{0.1}{1-0.7} = \frac{1}{3} < 1$. ∎

Example 2.3 (tight bigram model). In contrast, in Fig. 1b, obtained from Ex. 2.2 by changing the arcs from the b state, $p(\Sigma^*) = 1$. See App. B for details of this calculation. ∎

Ex. 2.2 above confirms that the autoregressive formulation does not necessarily yield p that is a valid distribution over Σ^*.
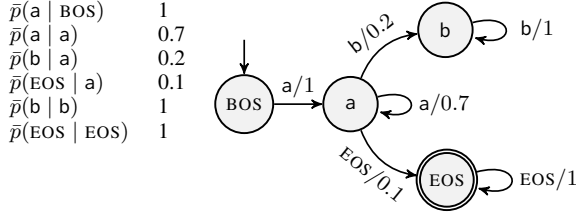
But if p is not a language model, what is it? It is intuitive to suspect that, in a model with $p(\Sigma^*) < 1$, the remainder of the probability mass "leaks" to infinite sequences, i.e., the generative process may continue forever with probability > 0. We will make this intuition formal in §3. By analogy with Chi and Geman (1998) and Cohen and Johnson (2013), we refer to such models as **non-tight**.[5]

The non-tightness of Ex. 2.2 is related to the fact that the probability of EOS is 0 at some states, in contrast to Ex. 2.3. However, requiring $\bar{p}(\text{EOS} \mid \boldsymbol{x}_{<t}) > 0$ for all prefixes $\boldsymbol{x}_{<t}$ is neither necessary nor sufficient to ensure tightness. It is not *necessary* because one can, for example, construct an ASM in which $\bar{p}(\text{EOS} \mid \boldsymbol{x}_{<t}) = 0.1$ when t is even but $= 0$ otherwise. Such a model generates only odd-length strings but is tight. It is not *sufficient* because of the following example, in which $\bar{p}(\text{EOS} \mid \boldsymbol{x}_{<t})$ is always positive but decays so rapidly toward 0 that the generative process might continue forever.
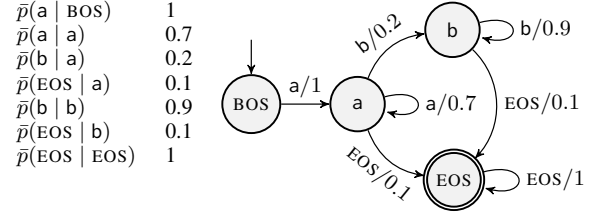
Example 2.4 (non-tight RNN). Consider an RNN over a small alphabet $\overline{\Sigma} = \{a, \text{EOS}\}$ with the fol-

---

[4] That is, suppose the ASM's conditional probabilities match the conditional probabilities of some known language model p_0: that is, $p_0(X_t = x_t \mid X_1 \ldots X_{t-1} = \boldsymbol{x}_{<t}) = \bar{p}(x_t \mid \boldsymbol{x}_{<t})$ whenever the former conditional probability is well-defined under the language model p_0, i.e., whenever $x_t \in \overline{\Sigma}$ and $\boldsymbol{x}_{<t} \in \Sigma^*$ with $p_0(X_1 \ldots X_{t-1} = \boldsymbol{x}_{<t}) > 0$.

Then by the chain rule of probability, $p(\boldsymbol{x}) = p_0(\boldsymbol{x})$ for each $\boldsymbol{x} \in \Sigma^*$. Thus $p = p_0$, so p is a language model.

[5] In Chi and Geman (1998) and Cohen and Johnson (2013), a PCFG is non-tight if its generative process may not terminate, and consequently the total probability of all finite trees is less than 1.

$\bar{p}(\text{a} \mid \text{BOS})$ 1
$\bar{p}(\text{a} \mid \text{a})$ 0.7
$\bar{p}(\text{b} \mid \text{a})$ 0.2
$\bar{p}(\text{EOS} \mid \text{a})$ 0.1
$\bar{p}(\text{b} \mid \text{b})$ 1
$\bar{p}(\text{EOS} \mid \text{EOS})$ 1

(a) Non-tight 2-gram model.

$\bar{p}(\text{a} \mid \text{BOS})$ 1
$\bar{p}(\text{a} \mid \text{a})$ 0.7
$\bar{p}(\text{b} \mid \text{a})$ 0.2
$\bar{p}(\text{EOS} \mid \text{a})$ 0.1
$\bar{p}(\text{b} \mid \text{b})$ 0.9
$\bar{p}(\text{EOS} \mid \text{b})$ 0.1
$\bar{p}(\text{EOS} \mid \text{EOS})$ 1

(b) Tight 2-gram model.

Figure 1: Tight and non-tight bigram models, expressed as Mealy machines (see §5.1). Transitions with conditional probability of 0 are omitted. The termination probability at a state is represented by an EOS arc from that state.

lowing hidden state recurrence:

$$h_0 = 0, \qquad h_t = \text{ReLU}(h_{t-1} + 1). \quad (1)$$

In this case, the hidden state admits a closed-form expression $h_t = t \in \mathbb{R}$. Setting the (1-dimensional) embedding of the alphabet to be $v_\text{a} = 1$ and $v_\text{EOS} = 0$, we arrive at

$$\bar{p}(\text{EOS} \mid \boldsymbol{x}_{<t}) = \text{softmax}(v_\text{a} h_t, v_\text{EOS} h_t)_\text{EOS}$$
$$= \frac{e^{0 \cdot t}}{e^{1 \cdot t} + e^{0 \cdot t}} = \frac{1}{e^t + 1} > 0. \quad (2)$$

The EOS probability is always strictly positive, but Thm. 4.7 shows that this sequence model is non-tight. Numerically, $p(\Sigma^*) \approx 0.702 < 1$. ∎

On the other hand, an ASM may be tight after all if the probability of EOS decays more slowly—even when it still approaches 0.

Example 2.5 (tight RNN). Consider again an RNN over the alphabet $\overline{\Sigma} = \{\text{a}, \text{EOS}\}$ with the following recurrence using softplus activation:[6]

$$h_1 = 0, \qquad h_t = \log(\exp(h_{t-1}) + 1). \quad (3)$$

Starting from $h_1 = 0 = \log 1$, a simple induction argument shows that

$$h_t = \log(\exp \log(t - 1) + 1) = \log t. \quad (4)$$

Again, setting $v_\text{a} = 1$ and $v_\text{EOS} = 0$, we arrive at

$$\bar{p}(\text{EOS} \mid \boldsymbol{x}_{<t}) = \text{softmax}(v_\text{a} h_t, v_\text{EOS} h_t)_\text{EOS} \quad (5)$$
$$= \frac{e^{0 \cdot \log t}}{e^{1 \cdot \log t} + e^{0 \cdot \log t}} = \frac{1}{t+1} > 0.$$

This decays slowly to 0: $\lim_{t \to \infty} \bar{p}(\text{EOS} \mid \boldsymbol{x}_{<t}) = 0$, but since $\sum_{t=1}^{\infty} \bar{p}(\text{EOS} \mid \boldsymbol{x}_{<t}) = \infty$, Prop. 4.3 below implies that this ASM is tight. ∎

Finally, we illustrate the peril of not treating distributions over uncountable sets carefully.

Example 2.6 (infinite coin toss). Consider the infinite independent fair coin toss model, where we aim to place a distribution over $\{\text{H}, \text{T}\}^\infty$, the uncountable set of infinite sequences of $\{\text{H}, \text{T}\}$. Intuitively, such a distribution corresponds to an ASM in which for all $\boldsymbol{x}_{<t}$, $\bar{p}(\text{H} \mid \boldsymbol{x}_{<t}) = \bar{p}(\text{T} \mid \boldsymbol{x}_{<t}) = \frac{1}{2}$ and $\bar{p}(\text{EOS} \mid \boldsymbol{x}_{<t}) = 0$. Clearly, each individual infinite sequence over $\{\text{H}, \text{T}\}$ should be assigned probability $(\frac{1}{2})^\infty = 0$. Without a formal foundation, one may arrive at the following paradox:

$$1 = p\left(\{\text{H}, \text{T}\}^\infty\right) = p\left(\bigcup_{\boldsymbol{\omega} \in \{\text{H},\text{T}\}^\infty} \{\boldsymbol{\omega}\}\right) \quad (6)$$
$$= \sum_{\boldsymbol{\omega} \in \{\text{H},\text{T}\}^\infty} p(\{\boldsymbol{\omega}\}) = \sum_{\boldsymbol{\omega} \in \{\text{H},\text{T}\}^\infty} 0 \overset{?}{=} 0. \quad \blacksquare$$

Together, these examples suggest that one must take care to characterize tightness. And, to the authors' surprise, it does not appear as if such a careful characterization yet exists in the NLP literature.

3 The Language Model Measure

In this section, we rigorously characterize the kind of distribution induced by an ASM. As mentioned earlier, an ASM can lose probability mass to the set of infinite sequences, Σ^∞. However, Σ^∞, unlike Σ^*, is uncountable, and it is due to this fact that we need to work explicitly with the measure-theoretic formulation of probability.

3.1 Measure-Theoretic Background

The goal of measure-theoretic probability is to assign probability to subsets of an **outcome space** Ω. For instance, in Ex. 2.6, $\Omega = \{\text{H}, \text{T}\}^\infty$. However, in the course of the study of measure theory, it has become clear that for many common Ω, it is impossible to assign probabilities in a way that satisfies a set of reasonable desiderata.[7] Consequently, the

---

[6] We use softplus instead of ReLU to simplify arithmetics.

[7] Measure theory texts commonly discuss such desiderata and the dilemmas that comes with them. See, e.g., Chapter 7 in Tao (2016), Chapter 3 in Royden (1988) or Chapter 3 in Billingsley (1995). We also give an example in Thm. 3.5.

standard approach to probability theory resorts to only assigning probability to certain "nice" subsets of Ω, which are referred to as **events** or **measurable subsets**, as in the theory of integration or functional analysis. The set of measurable subsets is commonly denoted as $\mathcal{F}$ (Def. 3.1), and a probability measure $\mathbb{P} : \mathcal{F} \to [0,1]$ is the function that assigns a probability to each measurable subset. As it turns out, the following simple and reasonable requirements imposed on $\mathcal{F}$ and $\mathbb{P}$ are enough to rigorously discuss probability (Kolmogorov, 1933).

Definition 3.1. *Let $\mathcal{P}(\Omega)$ be the powerset of Ω. Then $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ is called a $\boldsymbol{\sigma}$-**algebra** (or σ-field) over Ω if the following conditions hold:*
1. *$\Omega \in \mathcal{F}$,*
2. *if $E \in \mathcal{F}$, then its complement $E^{\mathsf{c}} \in \mathcal{F}$,*
3. *if $E_1, E_2, \ldots$ is a finite or infinite sequence of sets in $\mathcal{F}$, then $\bigcup_n E_n \in \mathcal{F}$.*

*If $\mathcal{F}$ is a σ-algebra over Ω, we call the tuple $(\Omega, \mathcal{F})$ a **measurable space**.*

A measurable space guarantees that operations on countably many sets are always valid, and hence permits the following definition.

Definition 3.2. *A **probability measure** $\mathbb{P}$ over a measurable space $(\Omega, \mathcal{F})$ is a function $\mathbb{P} : \mathcal{F} \to [0,1]$ such that*
1. *$\mathbb{P}(\Omega) = 1$,*
2. *if $E_1, E_2, \ldots$ is a finite or infinite sequence of disjoint sets in $\mathcal{F}$, then $\mathbb{P}(\bigcup_n E_n) = \sum_n \mathbb{P}(E_n)$.*

*In this case we call $(\Omega, \mathcal{F}, \mathbb{P})$ a **probability space**. Note that it assigns measure only to the sets in $\mathcal{F}$; other sets are said to be **non-measurable**.*

3.2 Sequence Models

As we saw in §2, sampling successive symbols from a non-tight ASM has probability > 0 of continuing forever. Hence, we hope to regard the ASM as defining a probability space over $\Omega = \Sigma^* \cup \Sigma^\infty$, where Σ^∞ denotes the set of infinite strings[8] over Σ. Note that this set Ω is uncountable whenever $|\Sigma| \geq 2$. We will first need to turn it into a measurable space by defining an appropriate σ-algebra.

This type of distribution is more general than a language model, which takes Ω to be the set Σ^* of finite strings. To distinguish the two, we refer to a distribution over $\Sigma^* \cup \Sigma^\infty$ as a sequence model.

---

[8]We will use the phrase "infinite string" in this paper when it is natural to do so, e.g., in the context of $\Sigma^* \cup \Sigma^\infty$. However, this is nonstandard terminology: in computer science, *string* generally refers to a finite object.

Definition 3.3. *A **sequence model** is a probability measure P over the set $\Sigma^* \cup \Sigma^\infty$.*

Intuitively (we will make this precise later), the event $\Sigma^\infty \subset \Sigma^* \cup \Sigma^\infty$ in Def. 3.3 represents **non-termination** of the generating process, i.e., it attempts to generate an infinitely long sequence. If this never happens, we have a language model.

Definition 3.4. *A **language model** is a probability measure P over just Σ^*. Equivalently, it is a sequence model P such that $P(\Sigma^\infty) = 0$.*

Our goal in the rest of this section is to rigorously construct a sequence model P that encodes the conditional probabilities of a given ASM. Since the ASM specifies conditional distributions over the augmented alphabet $\overline{\Sigma}$, we first use it to construct a probability measure $\mathbb{P}$ over a measurable space $(\overline{\Sigma}^\infty, \sigma(\overline{\mathcal{C}}))$. We then derive our sequence model P from $\mathbb{P}$ as the probability measure of a random variable X in a measurable space $(\Sigma^* \cup \Sigma^\infty, \sigma(\mathcal{C}))$. The σ-algebras $\sigma(\overline{\mathcal{C}})$ and $\sigma(\mathcal{C})$ will be built below.

3.3 Pre-Measure

As mentioned in §3.1, it is often impossible to measure the probability of every single subset of Ω. For example, in the infinite coin toss model in Ex. 2.6, we might begin by reasonably assigning probability 0 to each individual sequence $\boldsymbol{\omega} \in \{\mathsf{H}, \mathsf{T}\}^\infty$. Unfortunately, it is then impossible to assign probability to *every* subset of $\{\mathsf{H}, \mathsf{T}\}^\infty$; we must restrict our measurable space to a strict subset of $\mathcal{P}(\Omega)$, where $\mathcal{P}()$ is the powerset operator.

Theorem 3.5. *Assuming the Axiom of Choice and the Continuum Hypothesis, there exists no probability measure $\mathbb{P}$ over $(\{\mathsf{H}, \mathsf{T}\}^\infty, \mathcal{P}(\{\mathsf{H}, \mathsf{T}\}^\infty))$ such that $\mathbb{P}(\{\boldsymbol{\omega}\}) = 0$ for each $\boldsymbol{\omega} \in \{\mathsf{H}, \mathsf{T}\}^\infty$.*

Proof. This is a direct consequence of Ulam (1930). See App. C.1.1 for a discussion. $\blacksquare$

We will address this with well-known methods. A versatile theorem of Carathéodory provides a natural way to construct a probability space for sequences, in which prefix probabilities are well-defined. We first review two needed definitions.

Definition 3.6. *$\mathcal{A} \subseteq \mathcal{P}(\Omega)$ is called an **algebra** (or field) over Ω if*
1. *$\Omega \in \mathcal{A}$,*
2. *if $E \in \mathcal{A}$, then $E^{\mathsf{c}} \in \mathcal{A}$,*
3. *if $E_1, E_2 \in \mathcal{A}$, then $E_1 \cup E_2 \in \mathcal{A}$.*

Definition 3.7. *Let $\mathcal{A}$ be an algebra over some set Ω. A **probability pre-measure** over $(\Omega, \mathcal{A})$ is a function $\mathbb{P}_0 : \mathcal{A} \to [0,1]$ such that*

1) $\mathbb{P}_0(\Omega) = 1$,

2) if $E_1, E_2, \ldots$ is a (countable) sequence of disjoint sets in $\mathcal{A}$ whose (countable) union is also in $\mathcal{A}$, then $\mathbb{P}_0(\cup_{n=1}^\infty E_n) = \sum_{n=1}^\infty \mathbb{P}_0(E_n)$.

Note that the only difference between a σ-algebra (Def. 3.1) and an algebra is that condition 3 is weakened from countable to finite, and the only difference between a probability measure (Def. 3.2) and a pre-measure is that the latter is defined with respect to an algebra instead of a σ-algebra.

The idea behind Carathéodory's Extension Theorem is that there is often a simple construction of an algebra $\mathcal{A}$ over Ω such that there is a natural way to define a probability pre-measure. One can then extend this probability pre-measure to a probability measure that is both minimal and unique in a precise sense. For example, the standard Lebesgue measure over the the real line can be constructed in this way. For our case of infinite sequences, we will first construct an algebra over $\Omega = \overline{\Sigma}^\infty$ for some alphabet Σ. Then, assuming we are given an ASM $\bar{p}$ over $\overline{\Sigma}$, we can associate the algebra with a pre-measure that is consistent with $\bar{p}$. We will make use of the following definition to construct the algebra:

Definition 3.8. *Given any set $H \subseteq \overline{\Sigma}^k$, define its **cylinder set** (of **rank** k) to be*

$$\overline{C}(H) \stackrel{\text{def}}{=} \left\{ \boldsymbol{x}\boldsymbol{\omega} \; : \; \boldsymbol{x} \in H, \boldsymbol{\omega} \in \overline{\Sigma}^\infty \right\} \quad (7)$$

In essence, a cylinder set of rank k is the set of infinite strings that share their k-prefix with some string $\boldsymbol{x} \in H \subseteq \overline{\Sigma}^k$. For a length-$k$ string $\boldsymbol{x} = x_1 \cdots x_k$, the rank-$k$ cylinder set $\overline{C}(\boldsymbol{x}) \stackrel{\text{def}}{=} \overline{C}(\{\boldsymbol{x}\})$ is the set of all infinite strings prefixed by $\boldsymbol{x}$.[9] We denote the collection of all rank-k cylinder sets by $\overline{\mathcal{C}}_k \stackrel{\text{def}}{=} \left\{ \overline{C}(H) \; : \; H \in \mathcal{P}(\overline{\Sigma}^k) \right\}$ and define $\overline{\mathcal{C}} \stackrel{\text{def}}{=} \bigcup_{k=1}^\infty \overline{\mathcal{C}}_k$ to be the collection of all cylinder sets over Ω.[10]

Lemma 3.9. *$\overline{\mathcal{C}} \subset \mathcal{P}(\Omega)$ is an algebra over $\Omega = \overline{\Sigma}^\infty$.*

Proof. See App. C.1.2. ∎

We are now ready to define the pre-measure $\mathbb{P}_0$ for the cylinder algebra $\overline{\mathcal{C}}$. Given an ASM $\bar{p}$ and any set $\overline{C}(H) \in \overline{\mathcal{C}}$, let

$$\mathbb{P}_0(\overline{C}(H)) \stackrel{\text{def}}{=} \sum_{\boldsymbol{x} \in H} \bar{p}(\boldsymbol{x}) \quad (8)$$

where, denoting the length of $\boldsymbol{x}$ by k,

$$\bar{p}(\boldsymbol{x}) \stackrel{\text{def}}{=} \prod_{t=1}^k \bar{p}(x_t \mid \boldsymbol{x}_{<t}). \quad (9)$$

We confirm in Prop. C.2 that $\mathbb{P}_0$ is well-defined even though the cylinder set $\overline{C}(H)$ may also arise as $\overline{C}(H')$ where $H' \neq H$.[11]

Lemma 3.10. *$\mathbb{P}_0$ is a pre-measure over $\overline{\mathcal{C}}$.*

Proof. See App. C.1.2. ∎

3.4 Extension of Pre-Measure

We have now gathered all the ingredients needed to state Carathéodory's theorem.

Theorem 3.11 (Carathéodory's Extension Theorem). *Given an algebra $\mathcal{A}$ over some set Ω and a probability pre-measure $\mathbb{P}_0 : \mathcal{A} \to [0, 1]$, there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathcal{A} \subset \mathcal{F}$ and $\mathbb{P}|_{\mathcal{A}} = \mathbb{P}_0$. Furthermore, the σ-algebra $\mathcal{F}$ depends only on $\mathcal{A}$ and is minimal and unique—thus we may denote it by $\sigma(\mathcal{A})$—and the probability measure $\mathbb{P}$ is unique.*

Proof Sketch. See App. C.2.1. ∎

Applying Carathéodory's extension theorem to our cylinder algebra $\overline{\mathcal{C}}$ and pre-measure $\mathbb{P}_0$, we see that there exists a probability space $(\overline{\Sigma}^\infty, \sigma(\overline{\mathcal{C}}), \mathbb{P})$ over $\overline{\Sigma}^\infty$ that agrees with the ASM $\bar{p}$'s probabilities. It is a fair question to ask what kinds of sets are non-measurable under this construction; we discuss this in App. C.2.2.

3.5 A String-Valued Random Variable

Having constructed the probability space $(\overline{\Sigma}^\infty, \sigma(\overline{\mathcal{C}}), \mathbb{P})$, we now demonstrate how to use it to induce a probability space over $\Sigma^* \cup \Sigma^\infty$ as required by Def. 3.3. We will achieve this through the use of a random variable.

Definition 3.12 (random variable). *A mapping $X : \Omega \to S$ between two measurable spaces $(\Omega, \mathcal{F})$ and $(A, \mathcal{G})$ is an $(A, \mathcal{G})$-valued **random variable**, or a measurable mapping, if, for all $B \in \mathcal{G}$,*

$$X^{-1}(B) \stackrel{\text{def}}{=} \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}. \quad (10)$$

To construct a random variable that takes values in $\Sigma^* \cup \Sigma^\infty$, Def. 3.12 requires us to first construct a σ-algebra over $\Sigma^* \cup \Sigma^\infty$. We will do so in a similar

---

[9]This type of cylinder set, i.e., one that is generated by a singleton, is also called a **thin cylinder**.

[10]Observe that $\overline{\mathcal{C}}_1 \subset \overline{\mathcal{C}}_2 \subset \overline{\mathcal{C}}_3 \subset \cdots$.

[11]For example, in the infinite coin toss model, $\overline{C}(\mathsf{H}) = \overline{C}(\{\mathsf{HH}, \mathsf{HT}\})$.

fashion as we constructed $(\overline{\Sigma}^\infty, \overline{\mathcal{C}})$. Given $H \subseteq \Sigma^k$, define a rank-k cylinder set in $\Sigma^* \cup \Sigma^\infty$ to be

$$C(H) \stackrel{\text{def}}{=} \{ \boldsymbol{x}\boldsymbol{\omega} : \boldsymbol{x} \in H, \boldsymbol{\omega} \in \Sigma^* \cup \Sigma^\infty \}. \quad (11)$$

Let $\mathcal{C}_k$ be the set of all rank-k cylinder sets. Define $\mathcal{C} \stackrel{\text{def}}{=} \cup_{k=1}^\infty \mathcal{C}_k$. Then, $\sigma(\mathcal{C})$ is a σ-algebra by the same reasoning as in Lemma 3.9 and Thm. 3.11. We can now define the random variable X by[12]

$$X(\boldsymbol{\omega}) = \begin{cases} \boldsymbol{\omega}_{<k} & \text{if } \omega_k \text{ is the first EOS in } \boldsymbol{\omega} \\ \boldsymbol{\omega} & \text{otherwise (if EOS} \notin \boldsymbol{\omega}) \end{cases} \quad (12)$$

where $\boldsymbol{\omega} \in \overline{\Sigma}^\infty$. We claim that X is well-defined:

Proposition 3.13. *The function* X : $(\overline{\Sigma}^\infty, \sigma(\overline{\mathcal{C}})) \rightarrow (\Sigma^* \cup \Sigma^\infty, \sigma(\mathcal{C}))$ *defined in Eq. (12) is a measurable mapping.*

Proof. See App. C.3. ∎

Any measurable function induces a probability measure on the output space, called the pushforward measure (cf. §2.4 in Tao, 2011), given by

$$P(X \in E) \stackrel{\text{def}}{=} \mathbb{P}(X^{-1}(E)). \quad (13)$$

One can check that P, defined using $\mathbb{P}$, is indeed a probability measure on $(\Sigma^* \cup \Sigma^\infty, \sigma(\mathcal{C}))$ and hence $(\Sigma^* \cup \Sigma^\infty, \sigma(\mathcal{C}), P)$ is a probability space. We have therefore shown that, given any ASM, we can construct an associated sequence model as defined in Def. 3.3.

Under the formulation of a probability space together with a random variable, useful probability quantities arise naturally and intuitively. In particular, when $\boldsymbol{x} \in \Sigma^*$ is a finite string, we have

$$P(X = \boldsymbol{x}) \stackrel{\text{def}}{=} P(X \in \{\boldsymbol{x}\}) = p(\boldsymbol{x}) \quad (14)$$

with the definition of p from §2. Additionally, as we will show in the next section, the probability of the set of infinite strings $P(X \in \Sigma^\infty)$ is the probability of generating an infinite string.[13]

Deriving EOS As an aside, the preceding section allows us to motivate the EOS token in ASM as a construct that emerges naturally. Specifically, for any $\boldsymbol{x} \in \Sigma^*$, rearranging Eq. (14):

$$\bar{p}(\text{EOS} \mid \boldsymbol{x}) = \frac{P(X=\boldsymbol{x})}{\bar{p}(\boldsymbol{x})} = \frac{P(X=\boldsymbol{x})}{P(X \in C(\boldsymbol{x}))} \quad (15a)$$

---

[12]In this definition, the position $k \leq \infty$ of the first EOS—a **stopping time**—is itself a random variable.

[13]An important detail left out in this discussion is that both the singleton set $\{\boldsymbol{x}\}$ and Σ^∞ need to be measurable in $(\Sigma^* \cup \Sigma^\infty, \sigma(\mathcal{C}))$ for the above to make sense. This is addressed by Prop. C.7 and Prop. C.8.

$$= P(X = \boldsymbol{x} \mid X \in C(\boldsymbol{x})) \quad (15b)$$

where we have used $\bar{p}(\boldsymbol{x}) = \mathbb{P}(\overline{C}(\boldsymbol{x})) = \mathbb{P}(X^{-1}(C(\boldsymbol{x}))) = P(X \in C(\boldsymbol{x}))$. This means that the EOS probability in an ASM emerges as the conditional probability that, given that we must generate a string with a prefix $\boldsymbol{x} \in \Sigma^*$, the string is exactly $\boldsymbol{x}$.

4 Characterizing Tightness

Beyond the measure-theoretic formalization, a goal of this paper is to provide an exact characterization of tightness in ASMs. The results presented in this section generalize Lemma 3.2 in Welleck et al. (2020). First, we consider the event

$$A_k \stackrel{\text{def}}{=} \{ \boldsymbol{\omega} \in \overline{\Sigma}^\infty : \omega_k = \text{EOS} \} \quad (16)$$

in the probability space $(\overline{\Sigma}^\infty, \sigma(\overline{\mathcal{C}}), \mathbb{P})$. Intuitively, A_k is the event that an EOS symbol appears at position k in the string. Note that under this definition the A_k are not disjoint. For example, the string $\boldsymbol{\omega} = \text{ab EOS c EOS dd} \cdots$ lives in the intersection of A_3 and A_5 since EOS appears at both position 3 and position 5. Using Eq. (16), we can express the event consisting of all finite strings as $\bigcup_{k=1}^\infty A_k$. It follows that we can express the event of an infinite string as $(\bigcup_{k=1}^\infty A_k)^{\mathsf{c}} = \bigcap_{k=1}^\infty A_k^{\mathsf{c}}$. Thus, using the random variable X, we can express the probability of generating an infinite string as

$$P(X \in \Sigma^\infty) = \mathbb{P}(X^{-1}(\Sigma^\infty)) \quad (17a)$$
$$= \mathbb{P}\left(\bigcap_{k=1}^\infty A_k^{\mathsf{c}}\right). \quad (17b)$$

Hence, we can now formalize the notion of tightness, which we have introduced in §2 and Def. 3.4.

Definition 4.1. *A sequence model P is said to be* ***tight*** *if $P(X \in \Sigma^\infty) = 0$, in which case it is also a language model (cf. Prop. C.9). Otherwise, we say that it is* ***non-tight***.

Note that the definition of A_k only uses a string's k-prefix, and hence is a cylinder set of rank k. Recalling that the cylinder sets are measurable and so are the sets countably generated by them, we see that both the event consisting of all finite strings and the event consisting of all infinite strings are measurable. Thus, $\mathbb{P}(\cup_{k=1}^\infty A_k)$ and $\mathbb{P}(\cap_{k=1}^\infty A_k^{\mathsf{c}})$ are well defined.

4.1 A Lower Bound Result

We have characterized tightness in terms of the probability of a specific event $\mathbb{P}(\cap_{k=1}^\infty A_k^{\mathsf{c}})$, a quantity we now seek to determine.

Lemma 4.2. *If* $\sum_{n=2}^{\infty} \mathbb{P}\left(A_n \mid \cap_{m=1}^{n-1} A_m^{\mathsf{c}}\right) = \infty,$ *then* $\mathbb{P}\left(\cap_{m=1}^{\infty} A_m^{\mathsf{c}}\right) = 0.$

Proof. See App. D. ∎

Using Lemma 4.2, we can derive the following useful sufficient condition for a sequence model derived from an ASM to be tight. It applies when the probability of EOS does not decay too rapidly with the length of the prefix.

Proposition 4.3. *If* $\bar{p}(\text{EOS} \mid \boldsymbol{x}) \geq f(t)$ *for all* $t \geq 1, \boldsymbol{x} \in \Sigma^{t-1},$ *and* $\sum_{t=1}^{\infty} f(t) = \infty,$ *then* $\mathbb{P}(\cap_{k=1}^{\infty} A_k^{\mathsf{c}}) = 0.$ *In other words, P is tight.*

Proof. See App. D.2. ∎

This test implies tightness for all of the tight examples in §2, but not for the non-tight ones. Note that the lower-bounding function f depends only on the length of the prefix, not its content. f does not have to be monotonic—in the case of the even/odd example from §2, it is not.

4.2 The Borel–Cantelli Lemmata

It turns out that Prop. 4.3 admits a converse statement in which we can prove a similar property of $\bar{p}$ by assuming that the model is tight. To prove this result, we will use a fundamental inequality from probability theory—the Borel–Cantelli lemmata. The Borel–Cantelli lemmata are useful for our purposes because they relate the probability measure of sets of the form $\bigcap_{n=0}^{\infty} A_n$ or $\bigcup_{n=0}^{\infty} A_n$ to a series $\sum_{n=0}^{\infty} p_n$. We will only state the lemmata here without supplying their proofs;[14] however, we point out that Lemma 4.2 can be viewed as a parallel statement to the Borel–Cantelli lemmata and one can prove the lemmata using a very similar proof (cf. proof of Thm 2.3.7 in Durrett, 2019).

Concretely, given a sequence of events $\{A_n\}_{n=1}^{\infty}$ in some probability space, the Borel–Cantelli lemmata are statements about the event

$$\{A_n \text{ i.o.}\} \stackrel{\text{def}}{=} \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n \qquad (18)$$

where i.o. stands for "infinitely often." Intuitively, $\{A_n \text{ i.o.}\}$ is the set of outcomes that appear in infinitely many sets in the collection $\{A_n\}_{n=1}^{\infty}$ (hence the name). We will not use Borel–Cantelli directly, but they offer a probabilistic proof of a key result (Cor. 4.6) which will in turn lead to the desired statement about tightness. We formally state the first and second Borel–Cantelli lemmata below.

---

[14]See §2.3 in Durrett (2019) or §4 in Billingsley (1995) instead.

Lemma 4.4 (Borel–Cantelli I). *If* $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty,$ *then* $\mathbb{P}(A_n \text{ i.o.}) = 0.$

Lemma 4.5 (Borel–Cantelli II). *If* $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty,$ *then* $\mathbb{P}(A_n \text{ i.o.}) = 1,$ *provided that* $\{A_n\}$ *is a sequence of independent events.*

Using the Borel–Cantelli lemmata, we can prove the following useful fact.

Corollary 4.6. *Given a sequence* $\{p_n\}$ *where* $p_n \in [0, 1).$ *Then,*

$$\prod_{n=1}^{\infty} (1 - p_n) = 0 \iff \sum_{n=1}^{\infty} p_n = \infty. \quad (19)$$

Proof. See App. D.3. ∎

We now turn to proving a more general version of Prop. 4.3, which would imply its converse. First, we define the following quantity

$$\widetilde{p}_{\text{EOS}}(t) \stackrel{\text{def}}{=} \mathbb{P}(A_t \mid A_1^{\mathsf{c}} \cap \cdots \cap A_{t-1}^{\mathsf{c}}) \qquad (20)$$

which can be viewed as the EOS probability at step t, given that EOS was not generated at any earlier step. In Eq. (48a) in App. D.2, we show that, when $\widetilde{p}_{\text{EOS}}(t)$ is defined, it has the same value as

$$\widetilde{p}_{\text{EOS}}(t) = \frac{\sum_{\boldsymbol{\omega} \in \Sigma^{t-1}} \bar{p}(\boldsymbol{\omega}) \bar{p}(\text{EOS} \mid \boldsymbol{\omega})}{\sum_{\boldsymbol{\omega} \in \Sigma^{t-1}} \bar{p}(\boldsymbol{\omega})}. \qquad (21)$$

We can now completely characterize the tightness of an ASM with the following theorem.

Theorem 4.7 (Proposition 2.4 in Meister et al., 2022). *An ASM is tight if and only if* $\widetilde{p}_{\text{EOS}}(t) = 1$ *for some t or* $\sum_{t=1}^{\infty} \widetilde{p}_{\text{EOS}}(t) = \infty.$

Proof. See App. D.4. The proof uses Cor. 4.6, which accounts for the form of the condition. ∎

We remark that Thm. 4.7 is a generalization of Prop. 4.3 since if $\widetilde{p}_{\text{EOS}}(t)$ is lower-bounded by $f(t)$ whose series diverges, its own series would also diverge. However, since $\widetilde{p}_{\text{EOS}}(t)$ involves the computation of a partition function in its denominator, it can be intractable to calculate (Lin et al., 2021). Hence, Prop. 4.3 will be our main tool for determining tightness.

Finally, we note that Thm. 4.7 generalizes claims in previous work. For example, Welleck et al. (2020) require $f(t) = c > 0$ for some constant c to determine tightness. Hence, their bound is not helpful in determining the tightness in either Ex. 2.4 or Ex. 2.5, because the EOS probability can be arbitrarily small in both cases. Applying Prop. 4.3, we see that (1) the ASM in Ex. 2.4 is non-tight, because the series $\sum_{t=1}^{\infty} \frac{1}{e^t+1}$ is convergent, and (2) the ASM in Ex. 2.5 is tight, since the series $\sum_{t=1}^{\infty} \frac{1}{t+1}$ is divergent.

5 Analysis of Common Language Models

We now put into practice the foundations built up in the previous sections and discuss the tightness of several classes of ASMs.

5.1 Stochastic Finite-State Language Models

Language modeling based on n-grams has been historically influential in NLP (Jurafsky and Martin, 2009, Ch. 4). However, as Fig. 1 illustrates, n-gram language models are specific cases of the more general stochastic finite-state language models (Vidal et al., 2005). Tightness is more naturally characterized in this more general setting, as it turns out. We begin with a linear-algebraic definition of stochastic finite-state language models—or, more precisely, sequence models, since in this paper we do not consider the non-tight ones to be language models.

Definition 5.1. *A Q-state **stochastic finite-state sequence model** (SFSSM) is a quadruple $\left(\Sigma, \boldsymbol{s}, \{\mathbf{P}^{(a)}\}_{a \in \Sigma}, \boldsymbol{t}\right)$, where Σ is an alphabet of symbols, $\mathbf{P}^{(a)} \in \mathbb{R}_{\geq 0}^{Q \times Q}$ is a symbol-specific transition matrix for $a \in \Sigma$,[15] $\boldsymbol{s} \in \mathbb{R}_{\geq 0}^Q$ is a vector of initial state probabilities, and $\boldsymbol{t} \in \mathbb{R}_{\geq 0}^Q$ is a vector of termination probabilities, i.e., probabilities of generating EOS in each state.[16] We further require that $\sum_{q=1}^{Q} s_q = 1$ and that $t_q + \sum_{q'=1}^{Q} \mathbf{P}_{qq'} = 1$ for all $1 \leq q \leq Q$, where $\mathbf{P} \stackrel{\text{def}}{=} \sum_{a \in \Sigma} \mathbf{P}^{(a)}$ is the **transition sum matrix**.*

Given an SFSSM $\left(\Sigma, \boldsymbol{s}, \{\mathbf{P}^{(a)}\}_{a \in \Sigma}, \boldsymbol{t}\right)$, the probability of a string $\boldsymbol{x} \in \Sigma^*$ is defined by

$$\bar{p}(x_1 \cdots x_n) = \boldsymbol{s}^\top \left(\prod_{t=1}^n \mathbf{P}^{(x_t)}\right) \boldsymbol{t}. \tag{22}$$

Definition 5.2. *A state q of an SFSSM ($1 \leq q \leq Q$) is **accessible** if there is a positive-probability path to q from some state r with $s_r > 0$; it is **co-accessible** if there is a positive-probability path from q to some state r with $t_r > 0$. It is **useful** if it is both accessible and co-accessible, i.e., q appears on some positive-probability accepting path.*

Def. 5.2 allows a simple characterization of tight SFSSMs, namely Thm. 5.3, and a straightforward proof of Cor. 5.4.[17]

---

[15]For simplicity, we have disallowed ε-transitions.

[16]We use Q to denote the number of states as Q is the traditional notation for the set of states in a finite-state automaton.

[17]Cor. 5.4 is a special case of Chi and Geman (1998), who showed that MLE estimates of PCFGs are tight.

Theorem 5.3. *An SFSSM is tight iff all accessible states are also co-accessible.*

Proof. See App. E.1.1. ∎

Corollary 5.4. *Maximum likelihood estimates of n-gram models based on some corpus are tight.*

Proof. See App. E.1.1. ∎

In fact, we can express the termination probability of an SFSSM in simple linear algebra terms.

Definition 5.5. ***Trimming** an SFSSM means removing its non-useful (useless) states to obtain a **substochastic finite-state sequence model**.[18] This does not affect the string probabilities (22). Removing the non-useful states means removing their rows and columns from $\mathbf{P}$ as well as their rows from $\boldsymbol{s}$ and $\boldsymbol{t}$, yielding possibly smaller $\mathbf{P}'$, $\boldsymbol{s}'$ and $\boldsymbol{t}'$.*

Theorem 5.6. *Let $\mathbf{P}'$ be the transition sum matrix of a trimmed substochastic FSSM. Then $\mathbf{I} - \mathbf{P}'$ is invertible and $P(X \in \Sigma^*) = \boldsymbol{s}'^\top (\mathbf{I} - \mathbf{P}')^{-1} \boldsymbol{t}' \leq 1$.*

Proof. See App. E.1.2. ∎

The well-known matrix inversion formula used above finds the total weight of all accepting paths in any weighted graph (Tarjan, 1981).[19] The formula can be seen as a special case of Lehmann's (1977) algebraic path algorithm.

5.2 Transformer Language Models

We now prove that all Transformer language models are tight. Key to our proof of the tightness of various neural architectures, including the Transformer, is the following basic fact in topology.

Theorem 5.7. *Let X be a compact topological space and Y be any topological space. If $f : X \to Y$ is continuous, then $f(X) \subseteq Y$ is also compact.*

Proof. See App. E.2. ∎

To address the variable-length nature of modern deep NLP models, we will mathematically abstract them as a function on vector tuples,[20] $f : \left(\mathbb{R}^d\right)^+ \to \left(\mathbb{R}^d\right)^+$, that is length-preserving in the sense that $f\left(\mathbb{R}^{t \times d}\right) \subseteq \left(\mathbb{R}^{t \times d}\right)$ for all $t > 0$.

---

[18]We use the term *substochastic* rather than *stochastic* here because the trimmed model satisfies $t'_q + \sum_{q'=1}^{Q'} \mathbf{P}'_{qq'} \leq 1$, but might no longer achieve equality as required by Def. 5.1.

[19]This is assuming the total weight is finite (which we guarantee by substochasticity) and the matrix is invertible (which we guarantee by trimming)

[20]Here $\left(\mathbb{R}^d\right)^+$ is the set of nonempty tuples of vectors in $\mathbb{R}^d$. This is formally the disjoint union (coproduct) $\coprod_{t \in \mathbb{Z}_{>0}} \mathbb{R}^{t \times d}$.

Intuitively, this definition is saying that f is a function that maps a nonempty vector tuple $\{\boldsymbol{v}_i\}_{i=1}^t$ to another vector tuple $\{\boldsymbol{h}_i\}_{i=1}^t$ of the same length,

$$f(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_t) = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_t) \in \mathbb{R}^{t \times d}, \quad (23)$$

where $\boldsymbol{v}_i \in \mathbb{R}^d$ is commonly the embedding of the input symbol x_i. In particular, we can take the function $f : \left(\mathbb{R}^d\right)^+ \to \left(\mathbb{R}^d\right)^+$ to be the function defined by a stack of Transformer layers. This setup will help us state the following.

Lemma 5.8. *Let $f : \left(\mathbb{R}^d\right)^+ \to \left(\mathbb{R}^d\right)^+$ be the function defined by a finite number of Transformer layers (e.g., n layers) with any continuous activation function. Given a compact set $K \subset \mathbb{R}^d$. Then, there exists a compact set $K' \subset \mathbb{R}^d$ such that for every $t \in \mathbb{Z}_{>0}$,*

$$f\left(K^t\right) \subseteq \left(K'\right)^t. \quad (24)$$

Proof. See App. E.2. ∎

Recall that a Transformer language model—or more precisely, a Transformer ASM—defines the conditional probabilities using the softmax transformation

$$\bar{p}(x_{t+1} \mid \boldsymbol{x}_{\leq t}) = \frac{\exp(\boldsymbol{u}_{x_{t+1}}^\top \boldsymbol{h}_t)}{\sum_{y \in \overline{\Sigma}} \exp(\boldsymbol{u}_y^\top \boldsymbol{h}_t)} \quad (25)$$

where $\boldsymbol{u}_x \in \mathbb{R}^d$ is the output symbol embedding of $x \in \overline{\Sigma}$ and $\boldsymbol{h}_t$ is defined from the input embeddings of $\boldsymbol{x}_{\leq t}$ via Eq. (23). Using Lemma 5.8, together with the finiteness of the vocabulary Σ and the continuity of the softmax transformation (25), readily yields our main result on Transformers.

Theorem 5.9. *The autoregressive sequence model defined by any (fixed-depth) Transformer is tight.*

Proof. See App. E.2. ∎

5.3 Recurrent Neural Language Models

Recall that the hidden state of an RNN is typically defined by the recurrence

$$\boldsymbol{h}_t = \sigma\left(\boldsymbol{W}\boldsymbol{v}_t + \boldsymbol{U}\boldsymbol{h}_{t-1} + \boldsymbol{b}\right) \quad (26)$$

where $\boldsymbol{v}_t \in \mathbb{R}^d$ is the embedding of the input symbol x_t, as above, and $\sigma(\cdot)$ is some activation function (Elman, 1990). The conditional probabilities are usually defined in the same way as Eq. (25). Using Thm. 5.7 and the same strategy of proof as in Thm. 5.9, one can also easily prove the tightness of any RNN ASM with bounded activations (e.g.,

tanh or sigmoid). However, as we saw in Ex. 2.4, an unbounded activation function (e.g., ReLU) can indeed lead to non-tightness by making the probability of EOS decay too fast. The condition derived in Thm. 4.7 precisely determines how fast such decay can be without losing the tightness of the language model. Below, we generalize this result as well as Lemma 3.2 of Welleck et al. (2020), and show that if the norm of the activations eventually grows sub-logarithmically, the RNN is still tight.

Proposition 5.10. *Given an RNN ASM over $\overline{\Sigma}$. Again let the output symbol vector be $\boldsymbol{u}_x \in \mathbb{R}^d$ for $x \in \overline{\Sigma}$, and set $k \overset{\text{def}}{=} \sup_{x \in \Sigma} \|\boldsymbol{u}_x - \boldsymbol{u}_{\text{EOS}}\|_2$. Additionally, for each $t > 0$, let $\|\widehat{\boldsymbol{h}}_t\|_2$ be the maximum attainable hidden state norm for any context $\boldsymbol{x} \in \Sigma^t$. Such a sequence model is tight if $k\|\widehat{\boldsymbol{h}}_t\|_2 \leq \log t$ for all sufficiently large t.*

Proof. See App. E.3. ∎

This result is weaker than Thm. 5.9 because in an RNN, unlike a Transformer, the depth of the computation graph grows with the sequence length.

6 Conclusion

This paper presents a measure-theoretic treatment of language modeling and its tightness. Practical implications of our results include determining when sampling from an autoregressive sequence model is guaranteed to terminate and whether MCMC algorithms over such models will mix to the correct distribution.

To this end, we first defined various components of language modeling in measure-theoretic terminology. This in turn allows us to understand the portion of probability mass allocated to infinite-length strings. Importantly, this presentation formalizes a definition of sequence modeling under which the probability of producing an infinite-length sequence is non-zero; while today's models are often capable of producing such strings, previously there was no rigorous treatment of this case.

Indeed, such a definition is useful when considering a number of neural architectures (e.g., a simple RNN as in Elman, 1990) and language generation systems (e.g., the distribution induced by nucleus sampling; Holtzman et al., 2020). In particular, we showed that perhaps the most commonly-used NLP architecture, the Transformer language model, is indeed a language model—a tight distribution over finite strings—a property that had been called into question by previous work.

Limitations

Our discussion in this paper leaves out the consideration of computability of measures over languages. Specifically, we note that there exist works on computable measure theory developed in the context of theoretical computer science (de Leeuw et al., 1956) and probabilistic programming languages (Roy, 2011). Additional machinery needs to be developed in order for a proper treatment and we leave this for future work.

Another notable limitation is that we exclusively focused on the autoregressive production of language. Importantly, our formalism might not be compatible with other models of language production such as those induced by a PCFG.

Finally, our proofs of Thm. 5.9 and Prop. 5.10 exploit the strictly positive property of the softmax function. Importantly, they do not apply to models with sparse distributions (Martins and Astudillo, 2016; Peters et al., 2019; Martins, 2021).

Ethics

There are no ethical implications of this paper to the best knowledge of the authors.

Acknowledgments

References

Sheldon Axler. 2020. *Measure, Integration & Real Analysis*. Springer International Publishing.

Patrick Billingsley. 1995. *Probability and Measure*, 3rd edition. Wiley.

David Blackwell and Persi Diaconis. 1996. A non-measurable tail set. *Statistics, probability and game theory: Papers in honor of David Blackwell*, 30:1–5.

Taylor L. Booth and Richard A. Thompson. 1973. Applying probability measures to abstract languages. *IEEE Transactions on Computers*, C-22(5):442–450.

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Yining Chen, Sorcha Gilroy, Andreas Maletti, Jonathan May, and Kevin Knight. 2018. Recurrent neural networks as weighted language recognizers. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2261–2271, New Orleans, Louisiana. Association for Computational Linguistics.

Zhiyi Chi. 1999. Statistical properties of probabilistic context-free grammars. *Computational Linguistics*, 25(1):131–160.

Zhiyi Chi and Stuart Geman. 1998. Estimation of probabilistic context-free grammars. *Computational Linguistics*, 24(2):299–305.

Shay B. Cohen and Mark Johnson. 2013. The effect of non-tightness on Bayesian estimation of PCFGs. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1033–1041, Sofia, Bulgaria. Association for Computational Linguistics.

K. de Leeuw, E. F. Moore, C. E. Shannon, and N. Shapiro. 1956. *COMPUTABILITY BY PROBABILISTIC MACHINES*, Annals of Mathematics. Studies, no. 34, pages 183–212. Princeton University Press.

Rick Durrett. 2019. *Probability: Theory and Examples*, 5th edition. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211.

Gerald B. Folland. 1999. *Real Analysis: Modern Techniques and Their Applications*, 2nd edition. Wiley.

Charles M. Grinstead and J. Laurie Snell. 1997. *Introduction to Probability*, 2nd revised edition. American Mathematical Society.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Roger A. Horn and Charles R. Johnson. 2012. *Matrix Analysis*, 2nd edition. Cambridge University Press.

Frederick Jelinek. 1976. Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, 64(4):532–556.

Dan Jurafsky and James Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2nd edition. Pearson Prentice Hall.

A. N. Kolmogorov. 1933. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer.

Daniel J. Lehmann. 1977. Algebraic structures for transitive closure. *Theoretical Computer Science*, 4(1):59–76.

Chu-Cheng Lin. 2022. *On Expressiveness, Inference, and Parameter Estimation of Discrete Sequence Models*. Ph.D. thesis, Johns Hopkins University.

Chu-Cheng Lin, Aaron Jaech, Xin Li, Matthew R. Gormley, and Jason Eisner. 2021. Limitations of autoregressive models and their alternatives. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5147–5173, Online. Association for Computational Linguistics.

Chu-Cheng Lin and Arya D. McCarthy. 2022. On the uncomputability of partition functions in energy-based sequence models. In *International Conference on Learning Representations*.

Tianyu Liu, Yuchen Jiang, Nicholas Monath, Ryan Cotterell, and Mrinmaya Sachan. 2022. Autoregressive structure prediction with language models. In *Findings of the Association for Computational Linguistics: EMNL 2022*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Andre Martins and Ramon Astudillo. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1614–1623, New York, New York, USA. PMLR.

André F. T. Martins. 2021. Reconciling the discrete-continuous divide: Towards a mathematical theory of sparse communication. *CoRR*, abs/2104.00755.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the uniform information density hypothesis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 963–980, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2022. Locally typical sampling. *Transactions of the Association for Computational Linguistics*.

James R. Munkres. 2000. *Topology*, 2nd edition. Prentice Hall, Inc.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 807–814, Madison, WI, USA.

Mark-Jan Nederhof and Giorgio Satta. 2006. Estimation of consistent probabilistic context-free grammars. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 343–350, New York City, USA. Association for Computational Linguistics.

John C. Oxtoby. 1980. *Measure and Category: A Survey of the Analogies between Topological and Measure Spaces*. Springer New York.

Ben Peters, Vlad Niculae, and André F. T. Martins. 2019. Sparse sequence-to-sequence models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1504–1519, Florence, Italy. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 837–848,

Dublin, Ireland. Association for Computational Linguistics.

Daniel M. Roy. 2011. *Computability, Inference and Modeling in Probabilistic Programming*. Ph.D. thesis, Massachusetts Institute of Technology, USA. AAI0823858.

Halsey L. Royden. 1988. *Real Analysis*, 3rd edition. Prentice-Hall.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27:623–656.

Terence Tao. 2011. *An Introduction to Measure Theory*. American Mathematical Society.

Terence Tao. 2016. *Analysis II: Third Edition*. Texts and Readings in Mathematics. Springer Singapore.

Robert E. Tarjan. 1981. Fast algorithms for solving path problems. *Journal of the Association for Computing Machinery*, 28(3):594–614.

Stanisław Ulam. 1930. Zur masstheorie in der allgemeinen mengenlehre. *Fundamenta Mathematicae*, 16(1):140–150.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

E. Vidal, F. Thollard, C. de la Higuera, F. Casacuberta, and R.C. Carrasco. 2005. Probabilistic finite-state machines - part i. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(7):1013–1025.

Warren Weaver. 1955. Translation. In William N. Locke and A. Donald Boothe, editors, *Machine Translation of Languages*, pages 15–23. MIT Press. Reprinted from a memorandum written by Weaver in 1949.

Sean Welleck, Ilia Kulikov, Jaedeok Kim, Richard Yuanzhe Pang, and Kyunghyun Cho. 2020. Consistency of a recurrent language model with respect to incomplete decoding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5553–5568, Online. Association for Computational Linguistics.

A   Related Work

The issue of tightness has been studied extensively in the context of probabilistic context-free grammars (PCFG; Chi and Geman, 1998; Chi, 1999; Cohen and Johnson, 2013), although Chi (1999) refers to non-tight models as **improper**. Specifically, Chi (1999) gave algorithms for determining the tightness of a PCFG by formalizing a PCFG as a branching process. Chi (1999) further proved that any maximum-likelihood estimator yields a tight PCFG. Several previous works study the ability of language models to place probability mass on infinite-length strings (Booth and Thompson, 1973; Nederhof and Satta, 2006; Chen et al., 2018; Welleck et al., 2020), where they refer to the non-tight language models as **inconsistent**. In some cases, this behavior can be attributed to the discrepancy between the language model itself and the distribution induced by a (possibly stochastic) decoding algorithm: the decoder may have a lower probability of generating the EOS token. For example, on the tight bigram model of Ex. 2.3, a greedy decoder will always generate a and never EOS. Yet in other examples, it is the model itself that leaks probability mass to infinite-length strings, i.e., it may be non-tight, which is the problem we focus on in this work, providing a characterization of tightness. Notably, the conditions we propose are more general than those of Welleck et al. (2020).

Several other works consider the limitations of common neural network architectures for modeling distributions over finite sequences (strings), albeit focusing specifically on other attributes, such as their computational complexity for problems like equivalence or undecidability (Chen et al., 2018; Lin et al., 2021; Lin and McCarthy, 2022; Lin, 2022). In contrast, this work provides a formal treatment of language models by enlarging the sample space to $\Sigma^* \cup \Sigma^\infty$, although to ensure tightness, Σ^∞ must receive probability 0. Such definitions are not uncommon in probability theory. For example, while the Wiener process (i.e., the standard Brownian motion) is a distribution over *all* functions, the definition ensures that the set of discontinuous functions is assigned probability 0 (Durrett, 2019, Ch. 7).

Meister et al. (2022) similarly address the notion of a language model as a distribution over infinite sequences by casting such models as stochastic processes. They use this framing in order to motivate decoding, without providing comprehensive measure-theoretic foundations of such distributions.

B   Details for Motivating Ex. 2.3

Here, we lay out the steps to calculate $\mathbb{P}(\Sigma^*)$ from Fig. 1b:

$$\mathbb{P}(\Sigma^*) = \sum_{n=0}^{\infty} \left(\mathbb{P}(\mathsf{a}^{n+1}) + \sum_{m=0}^{\infty} \mathbb{P}(\mathsf{a}^{n+1}\mathsf{b}^{m+1}) \right) \tag{27a}$$

$$= \sum_{n=0}^{\infty} \left(1 \cdot (0.7)^n \cdot 0.1 + \sum_{m=0}^{\infty} 1 \cdot (0.7)^n \cdot 0.2 \cdot (0.9)^m \cdot 0.1 \right) \tag{27b}$$

$$= \sum_{n=0}^{\infty} (0.7)^n \cdot \left(0.1 + 0.2 \cdot \left(\sum_{m=0}^{\infty} (0.9)^m \right) \cdot 0.1 \right) \tag{27c}$$

$$= \sum_{n=0}^{\infty} (0.7)^n \cdot \left(0.1 + 0.2 \cdot \frac{1}{1 - 0.9} \cdot 0.1 \right) \tag{27d}$$

$$= \sum_{n=0}^{\infty} (0.7)^n \cdot 0.3 = \frac{0.3}{1 - 0.7} = 1 \tag{27e}$$

C   Measure Theory Details

C.1   Proofs and Details in §3.3

C.1.1   Details of Thm. 3.5

Theorem 3.5. *Assuming the Axiom of Choice and the Continuum Hypothesis, there exists no probability measure* $\mathbb{P}$ *over* $(\{\mathsf{H}, \mathsf{T}\}^\infty, \mathcal{P}(\{\mathsf{H}, \mathsf{T}\}^\infty))$ *such that* $\mathbb{P}(\{\boldsymbol{\omega}\}) = 0$ *for each* $\boldsymbol{\omega} \in \{\mathsf{H}, \mathsf{T}\}^\infty$.

This theorem is an **impossibility of measure** theorem. Generally speaking, the existence of a non-measurable set implies some form of impossibility of measure. The most famous example of non-measurable sets are Vitali sets, which exist given the Axiom of Choice. Vitali's 1905 construction is typically described in introductory texts on measure theory (Royden, 1988; Billingsley, 1995; Axler, 2020). The existence of Vitali sets shows that it is impossible to define a probability measure that satisfies translational invariance on the measurable space $\big([0,1), \mathcal{P}([0,1))\big)$. Thus, to achieve translational invariance, Lebesgue measure uses a σ-algebra smaller than $\mathcal{P}([0,1))$, in which the Vitali sets are among the non-measurable sets. However, the translational invariance desideratum is not relevant to our space of discrete sequences. A theorem by Ulam (1930) reveals a deeper reason that some sets must be non-measurable. We shall state the theorem below as given in Oxtoby (1980) and omit its proof. We refer interested readers to Chapter 5 in Oxtoby (1980), which contains an accessible proof and an excellent discussion of the theorem including its generalizations and historical context.

Theorem C.1 (Ulam, 1930). *Assuming the Axiom of Choice, a finite measure μ defined for all subsets of a set X of cardinality $\aleph_1$ vanishes identically [that is, equals zero for all subsets] if it is equal to zero for every one-element subset.*

In the statement above, $\aleph_1$ denotes the cardinality of the first uncountable ordinal. We can see that Thm. 3.5 is a straightforward consequence of Thm. C.1.

Proof of Thm. 3.5. Recall that $\mathrm{card}(\{\mathsf{H},\mathsf{T}\}^\infty) = 2^{\aleph_0}$. Assuming the Continuum Hypothesis, $2^{\aleph_0} = \aleph_1$, and hence by Thm. C.1, such a measure is uniformly 0, and hence cannot be a probability measure. ∎

C.1.2 Other Proofs in §3.3

Lemma 3.9. $\overline{\mathcal{C}} \subset \mathcal{P}(\Omega)$ *is an algebra over* $\Omega = \overline{\Sigma}^\infty$.

Proof. First, $\Omega \in \overline{\mathcal{C}}$ since it is a cylinder set of rank 0 or indeed of any rank k: $\Omega = \overline{C}(\overline{\Sigma}^k) \in \overline{\mathcal{C}}_k \subset \overline{\mathcal{C}}$. Second, $\overline{\mathcal{C}}$ is closed under complements: given a cylinder set of rank k, that is, $\overline{C}(H)$ where $H \subseteq \overline{\Sigma}^k$, its complement $\big(\overline{C}(H)\big)^\mathsf{c} = \overline{C}\big(\overline{\Sigma}^k \setminus H\big)$ is also a cylinder set of rank k. Finally, $\overline{\mathcal{C}}$ is closed under union: the union of cylinder sets of ranks $k_1 \leq k_2$ is a cylinder set of rank k_2, since both can be regarded as cylinder sets of rank k_2. Hence, $\overline{\mathcal{C}}$ is an algebra over Ω. ∎

Proposition C.2. $\mathbb{P}_0$ *as defined in Eq. (8) is a well-defined function.*

Proof. Suppose a cylinder set arises in two ways, $\overline{C}(H_1) = \overline{C}(H_2)$, where $H_1 \subseteq \overline{\Sigma}^{k_1}$ and $H_2 \subseteq \overline{\Sigma}^{k_2}$. We must show $\sum_{\boldsymbol{x} \in H_1} \bar{p}(\boldsymbol{x}) = \sum_{\boldsymbol{x}' \in H_2} \bar{p}(\boldsymbol{x}')$. Without loss of generality, assume that $k_1 \leq k_2$. The definition of $\overline{C}(H_2)$ (Def. 3.8) implies that H_2 consists of all length-k_2 prefixes of strings in $\overline{C}(H_2)$. But $\overline{C}(H_2) = \overline{C}(H_1)$, so the definition of $\overline{C}(H_1)$ (Def. 3.8) implies that its length-k_2 prefixes are exactly the strings of the form $\boldsymbol{xy}$ where $\boldsymbol{x} \in H_1, \boldsymbol{y} \in \overline{\Sigma}^{k_2-k_1}$. Hence we can write H_2 in terms of H_1 as $H_2 = \{\boldsymbol{xy} : \boldsymbol{x} \in H_1, \boldsymbol{y} \in \overline{\Sigma}^{k_2-k_1}\}$. Thus

$$\sum_{\boldsymbol{x}' \in H_2} \bar{p}(\boldsymbol{x}') = \sum_{\boldsymbol{x} \in H_1} \sum_{\boldsymbol{y} \in \overline{\Sigma}^{k_2-k_1}} \bar{p}(\boldsymbol{xy}) = \sum_{\boldsymbol{x} \in H_1} \bar{p}(\boldsymbol{x}) \tag{28}$$

where the last equality is true because $\bar{p}$ is defined by the locally normalized product (9). ∎

Lemma 3.10. $\mathbb{P}_0$ *is a pre-measure over* $\overline{\mathcal{C}}$.

For the proof of Lemma 3.10, we will mostly follow the proof of Thm 2.3 in Billingsley (1995), with the exception of invoking the Tychonoff theorem directly. This proof depends on the following lemma, which is Example 2.10 in Billingsley (1995). We repeat the statement and proof here for the reader's convenience.

Lemma C.3. *Let $\mathbb{P}_0$ be a finitely additive probability pre-measure over $\overline{\mathcal{C}}$ such that, given a decreasing sequence of sets $A_1 \supset A_2 \supset \cdots$ in $\overline{\mathcal{C}}$ where $\bigcap_{n=1}^\infty A_n = \emptyset$, $\lim_{n\to\infty} \mathbb{P}_0(A_n) = 0$. Then, $\mathbb{P}_0$ is also countably additive over $\overline{\mathcal{C}}$.*

Proof. Let $\{A_n\}$ be a sequence of disjoint sets in $\overline{\mathcal{C}}$ such that $A = \bigcup_n A_n \in \overline{\mathcal{C}}$. Then, defining $B_n = \bigcup_{m>n} A_m$, we see that $B_1 \supset B_2 \supset \cdots$ and $\bigcap_n B_n = \emptyset$. Notice that

$$A = A_1 \cup B_1 = A_1 \cup A_2 \cup B_2 = \cdots = A_1 \cup \cdots \cup A_n \cup B_n \tag{29}$$

for any n and hence by finite additivity of $\mathbb{P}_0$

$$\mathbb{P}_0(A) = \mathbb{P}_0(A_1) + \cdots + \mathbb{P}_0(A_n) + \mathbb{P}_0(B_n) \tag{30}$$

or equivalently

$$\mathbb{P}_0(A_1) + \cdots + \mathbb{P}_0(A_n) = \mathbb{P}_0(A) - \mathbb{P}_0(B_n). \tag{31}$$

Since, $B_n \downarrow \emptyset$ implies that $\mathbb{P}_0(B_n) \downarrow 0$ by assumption, taking the limits on both sides of Eq. (31) yields

$$\sum_n \mathbb{P}_0(A_n) = \lim_{n\to\infty} \sum_{i\leq n} \mathbb{P}_0(A_i) = \mathbb{P}_0(A) - \lim_{n\to\infty} \mathbb{P}_0(B_n) = \mathbb{P}_0(A) \tag{32}$$

which shows countable additivity. ∎

We also recall the Tychonoff theorem.[21]

Theorem C.4 (Tychonoff)**.** *Let $\{X_\alpha\}_{\alpha\in J}$ be an indexed family of compact topologies. Then, their product topology $\prod_{\alpha\in J} X_\alpha$ is also compact.*

We can now give the proof for Lemma 3.10.

Proof of Lemma 3.10. We first show that $\mathbb{P}_0$ is finitely additive over $\overline{\mathcal{C}}$. Let $C(H_1)$ and $C(H_2)$ be two disjoint cylinder sets. By Prop. C.2, we can assume they are of the same rank without loss of generality. Then,

$$C(H_1) \cup C(H_2) = \bigcup_{\boldsymbol{x}\in H_1} \{\boldsymbol{x}\boldsymbol{\omega} : \boldsymbol{\omega} \in \overline{\Sigma}^\infty\} \cup \bigcup_{\boldsymbol{x}\in H_2} \{\boldsymbol{x}\boldsymbol{\omega} : \boldsymbol{\omega} \in \overline{\Sigma}^\infty\} \tag{33a}$$

$$= \bigcup_{\boldsymbol{x}\in H_1\cup H_2} \{\boldsymbol{x}\boldsymbol{\omega} : \boldsymbol{\omega} \in \overline{\Sigma}^\infty\} \quad (H_1 \text{ and } H_2 \text{ equal rank and disjoint}) \tag{33b}$$

$$= C(H_1 \cup H_2) \tag{33c}$$

which leads to

$$\mathbb{P}_0(C(H_1) \cup C(H_2)) = \mathbb{P}_0(C(H_1 \cup H_2)) = \sum_{\boldsymbol{x}\in H_1\cup H_2} \bar{p}(\boldsymbol{x}) = \mathbb{P}_0(C(H_1)) + \mathbb{P}_0(C(H_2)). \tag{34a}$$

Hence, $\mathbb{P}_0$ is finitely additive.

Now, equip $\overline{\Sigma}$ with the discrete topology. Since $\overline{\Sigma}$ is finite, it is compact under the discrete topology and so is $\overline{\Sigma}^\infty$ by Thm. C.4. Then, by properties of the product topology over discrete finite spaces, all cylinder sets in $\overline{\Sigma}^\infty$ are compact. To apply Lemma C.3, let $C_1 \supset C_2 \supset \cdots$ be a decreasing sequence of cylinder sets with empty intersection. Suppose to the contrary that $\mathbb{P}_0\left(\bigcap_n C_n\right) > 0$. This would imply that all C_n are nonempty (any of these being empty would result in a measure 0). However, by Cantor's intersection theorem[22], $\bigcap_n C_n$ is nonempty, contradicting the assumption. Hence, $\mathbb{P}_0\left(\bigcap_n C_n\right) = 0$, and by Lemma C.3, $\mathbb{P}_0$ is countably additive. ∎

---

[21] See §37 in Munkres (2000) for a detailed and well-written treatise.

[22] Cantor's intersection theorem states that a decreasing sequence of nonempty compact sets have a nonempty intersection. A version of this result in introductory real analysis is the Nested Interval Theorem.

C.2 Details in §3.4

C.2.1 Carathéodory's Extension Theorem

Theorem 3.11 (Carathéodory's Extension Theorem). *Given an algebra $\mathcal{A}$ over some set Ω and a probability pre-measure $\mathbb{P}_0 : \mathcal{A} \to [0, 1]$, there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathcal{A} \subset \mathcal{F}$ and $\mathbb{P}|_{\mathcal{A}} = \mathbb{P}_0$. Furthermore, the σ-algebra $\mathcal{F}$ depends only on $\mathcal{A}$ and is minimal and unique—thus we may denote it by $\sigma(\mathcal{A})$—and the probability measure $\mathbb{P}$ is unique.*

Proof Sketch. First, construct an outer measure by approximation with countable coverings. Then, show that the collection of sets that is measurable with respect to this outer measure is a σ-algebra $\mathcal{F}$ that contains $\mathcal{A}$. Finally, restricting the outer measure to this σ-algebra, one is then left with a probability space. To show minimality, one can show that $\mathcal{F}$ is contained in any σ-algebra that contains $\mathcal{A}$. Uniqueness is given by applying Dynkin's π-λ theorem (Theorem 3.2 in Billingsley, 1995).

Great care must be taken in each step involved in the outline above. To address these is well beyond the scope of this paper and we refer reader to the many excellent texts with a proof of this theorem, such as Chapter 12 in Royden (1988) and Chapter 11 in Billingsley (1995). ∎

C.2.2 The Space of Non-measurable Sets

Non-measurable sets are, in general, difficult to find. Even when we can exhibit such sets, they tend to be very abstract and counter-intuitive. Vitali's and Bernstein's sets are two prominent examples for the Lebesgue measure. Blackwell and Diaconis (1996) offers a construction of a non-measurable set in the cylinder σ-algebra.[23]

As another approach to understand this better, we can consider how our collection $\sigma(\overline{\mathcal{C}})$ of all measurable sets, i.e., our σ-algebra, is constructed from our algebra $\overline{\mathcal{C}}$ of cylinder sets (as opposed to simply knowing from Carathéodory's Extension Theorem that it exists). Concretely, as in §1.6 in Folland (1999), we can intuitively consider the following process to build from the collection of cylinder sets $\overline{\mathcal{C}}$, which is a countable collection, all the way up to its generated σ-algebra, whose cardinality is unknown just yet:
- Let $\overline{\mathcal{C}}_0 = \overline{\mathcal{C}}$,
- Let $\overline{\mathcal{C}}_1$ be the set that includes all countable unions of sets in $\overline{\mathcal{C}}_0$ or the complements of such,
- Repeat this process to build $\overline{\mathcal{C}}_n$ for every $n \in \mathbb{N}$.

One might then take the union $\bigcup_{n \in \mathbb{N}} \overline{\mathcal{C}}_n$ of this increasing sequence of collections of sets, and ask if it is the same as $\sigma(\overline{\mathcal{C}})$. In general, the answer is no (as one might expect if one is familiar with the Borel Hierarchy). However, we can obtain $\sigma(\overline{\mathcal{C}})$ if we perform this construction for every countable ordinal α. Abbreviating the operation in the second step above as δ, i.e., $\overline{\mathcal{C}}_1 = \delta(\overline{\mathcal{C}}_0)$, and letting ω_1 be the collection of all countable ordinals,[24] we can define

$$\overline{\mathcal{C}}_\alpha = \begin{cases} \delta(\overline{\mathcal{C}}_\beta) & \text{if } \alpha = \beta + 1 \text{ for some } \beta \in \omega_1, \\ \bigcup_{\beta \in \omega_1 : \beta < \alpha} \overline{\mathcal{C}}_\beta & \text{otherwise.} \end{cases} \tag{35}$$

This will give us the desired set as follows:

Proposition C.5 (Proposition 1.23, Folland, 1999). $\sigma(\overline{\mathcal{C}}) = \bigcup_{\alpha \in \omega_1} \overline{\mathcal{C}}_\alpha$.

Next, we recall the following basic fact from cardinality theory.

Proposition C.6 (Proposition 0.14, Folland, 1999). *If $\operatorname{card}(A) \leq 2^{\aleph_0}$ and $\operatorname{card}(X_\alpha) \leq 2^{\aleph_0}$ for all $\alpha \in A$, then $\operatorname{card}\left(\bigcup_{\alpha \in A} X_\alpha\right) \leq 2^{\aleph_0}$.*

Noting that $\operatorname{card}(\omega_1) \leq 2^{\aleph_0}$ and $\operatorname{card}(\overline{\mathcal{C}}) = \aleph_0$, we can conclude that $\operatorname{card}(\sigma(\overline{\mathcal{C}})) \leq 2^{\aleph_0}$ from Prop. C.5 and Prop. C.6. In other words, the cardinality of $\sigma(\overline{\mathcal{C}})$ is at most that of the continuum, and since $\operatorname{card}(\mathcal{P}(\overline{\Sigma}^\infty)) = 2^{2^{\aleph_0}}(= \beth_2)$, $\sigma(\overline{\mathcal{C}})$ is, in terms of cardinality, an almost negligible subset of $\mathcal{P}(\overline{\Sigma}^\infty)$! That is, most subsets in $\overline{\Sigma}^\infty$ are non-measurable—though explicit examples have rarely been constructed

---

[23]The following assumes basic familiarity with the theory of ordinal numbers. Readers without such background may skip to the last paragraph for conclusion.

[24]ω_1 is, in fact, the same as the first uncountable ordinal. Its existence (and hence the existence of the collection of all countable ordinals) can be guaranteed by exhibiting a well-ordered uncountable set using the Axiom of Choice.

(Blackwell and Diaconis, 1996). App. C.3 below establishes that common subsets of $\overline{\Sigma}^\infty$ that we work with are measurable.

C.3 Proofs in §3.5

Proposition 3.13. *The function* $X : (\overline{\Sigma}^\infty, \sigma(\overline{\mathcal{C}})) \to (\Sigma^* \cup \Sigma^\infty, \sigma(\mathcal{C}))$ *defined in Eq. (12) is a measurable mapping.*

Proof. To show that X is measurable, it suffices to show the measurability of preimages of a generating set[25] of the σ-algebra $\sigma(\mathcal{C})$ on $\Sigma^* \cup \Sigma^\infty$. Such a generating set is formed by the thin cylinders $C(\boldsymbol{x}) \overset{\text{def}}{=} C(\{\boldsymbol{x}\})$ for $\boldsymbol{x} \in \Sigma^*$. (Recall that cylinders in $\Sigma^* \cup \Sigma^\infty$ are defined by Eq. (11).) Given $\boldsymbol{x} \in \Sigma^*$:

$$X^{-1}(C(\boldsymbol{x})) = X^{-1}(\{\boldsymbol{x}\boldsymbol{\omega} : \boldsymbol{\omega} \in \Sigma^* \cup \Sigma^\infty\}) \tag{36a}$$

$$= X^{-1}(\{\boldsymbol{x}\boldsymbol{\omega} : \boldsymbol{\omega} \in \Sigma^*\}) \cup X^{-1}(\{\boldsymbol{x}\boldsymbol{\omega} : \boldsymbol{\omega} \in \Sigma^\infty\}) \tag{36b}$$

$$= \left(\bigcup_{\boldsymbol{\omega} \in \Sigma^*} \overline{C}(\boldsymbol{x}\boldsymbol{\omega}\,\text{EOS}) \right) \cup \left(\overline{C}(\boldsymbol{x}) \cap \bigcap_{k=1}^\infty A_k^{\mathsf{c}} \right) \tag{36c}$$

Note that the set A_k above, defined by Eq. (16), is a cylinder of $\overline{\Sigma}^\infty$, representing the event of terminating by step k. Then, from the derivation above, we can see that $X^{-1}(C(\boldsymbol{x}))$ is formed by countable operations over measurable sets (cylinders) of $\overline{\Sigma}^\infty$, and is hence measurable. So X is a measurable function. ∎

Proposition C.7. *In measure space* $(\Sigma^* \cup \Sigma^\infty, \sigma(\mathcal{C}))$, $\{\boldsymbol{x}\}$ *is measurable for all* $\boldsymbol{x} \in \Sigma^*$.

Proof. We will show that $\{\boldsymbol{x}\} = C(\boldsymbol{x}) \setminus \bigcup_{a \in \Sigma} C(\boldsymbol{x}a)$ and hence is measurable. By definition in Eq. (11), for any $\boldsymbol{x} \in \Sigma^*$,

$$C(\boldsymbol{x}) = \{\boldsymbol{x}\boldsymbol{\omega} : \boldsymbol{\omega} \in \Sigma^* \cup \Sigma^\infty\} \tag{37a}$$

$$= \{\boldsymbol{x}\boldsymbol{\omega} : \boldsymbol{\omega} \in \Sigma^*\} \cup \{\boldsymbol{x}\boldsymbol{\omega} : \boldsymbol{\omega} \in \Sigma^\infty\} \tag{37b}$$

where

$$\{\boldsymbol{x}\boldsymbol{\omega} : \boldsymbol{\omega} \in \Sigma^*\} = \{\boldsymbol{x}\} \cup \bigcup_{a \in \Sigma} \{\boldsymbol{x}a\boldsymbol{\omega} : \boldsymbol{\omega} \in \Sigma^*\} \tag{38a}$$

and

$$\{\boldsymbol{x}\boldsymbol{\omega} : \boldsymbol{\omega} \in \Sigma^\infty\} = \bigcup_{a \in \Sigma} \{\boldsymbol{x}a\boldsymbol{\omega} : \boldsymbol{\omega} \in \Sigma^\infty\}. \tag{39}$$

So

$$C(\boldsymbol{x}) = \{\boldsymbol{x}\} \cup \bigcup_{a \in \Sigma} \left(\{\boldsymbol{x}a\boldsymbol{\omega} : \boldsymbol{\omega} \in \Sigma^*\} \cup \{\boldsymbol{x}a\boldsymbol{\omega} : \boldsymbol{\omega} \in \Sigma^\infty\} \right) \tag{40a}$$

$$= \{\boldsymbol{x}\} \cup \bigcup_{a \in \Sigma} C(\boldsymbol{x}a) \tag{40b}$$

where the union is disjoint. This implies $\{\boldsymbol{x}\} = C(\boldsymbol{x}) \setminus \bigcup_{a \in \Sigma} C(\boldsymbol{x}a)$ as desired. ∎

Proposition C.8. *In the measure space* $(\Sigma^* \cup \Sigma^\infty, \sigma(\mathcal{C}))$, Σ^∞ *is measurable.*

Proof. First, $\Sigma^* \cup \Sigma^\infty$ is the entire outcome space, which is measurable by the definition of σ-algebra. Notice that

$$\Sigma^\infty = (\Sigma^* \cup \Sigma^\infty) \setminus \bigcup_{\boldsymbol{x} \in \Sigma^*} \{\boldsymbol{x}\}. \tag{41}$$

Since each $\{\boldsymbol{x}\}$ in the above is measurable by Prop. C.7 and Σ^* is a countable set, Σ^∞ is then measurable. ∎

---

[25]A set G is said to be a generating set of a σ-algebra $\mathcal{F}$ if $\mathcal{F}$ is the smallest σ-algebra that contains G.

The measurability of Σ^∞ in $(\Sigma^* \cup \Sigma^\infty, \sigma(\mathcal{C}))$ (Prop. C.8) was assumed by our definition of tightness (Def. 4.1). As we have also established that each $\{x\}$ is measurable (Prop. C.7), we can give an alternative characterization.

Proposition C.9. *A sequence model* $(\Sigma^* \cup \Sigma^\infty, \sigma(\mathcal{C}), P)$ *is tight if and only if* $\sum_{x \in \Sigma^*} P(\{x\}) = 1$.

Proof. We defined a sequence model to be tight if and only if $P(\Sigma^\infty) = 0$ (Def. 4.1). By Propositions C.7 and C.8, we can write

$$1 = P(\Sigma^* \cup \Sigma^\infty) = P(\Sigma^\infty) + P(\Sigma^*) \qquad \text{(finite additivity)} \qquad (42a)$$

$$= P(\Sigma^\infty) + \sum_{x \in \Sigma^*} P(\{x\}). \qquad \text{(countable additivity)} \qquad (42b)$$

Hence, a sequence model is tight if and only if $\sum_{x \in \Sigma^*} P(\{x\}) = 1$. ∎

D Proofs on Characterizing Tightness (§4)

D.1 Proof of Lemma 4.2

The result below is stated without proof as Exercise 4.3.5 in Durrett (2019).

Lemma 4.2. *If* $\sum_{n=2}^\infty \mathbb{P}\left(A_n \mid \cap_{m=1}^{n-1} A_m^\mathsf{c}\right) = \infty$, *then* $\mathbb{P}\left(\cap_{m=1}^\infty A_m^\mathsf{c}\right) = 0$.

Proof. First, recall an elementary inequality that for $x > 0$,

$$x - 1 \geq \log x \quad \Leftrightarrow \quad 1 - x \leq \log \frac{1}{x}. \qquad (43)$$

Note that $\mathbb{P}(\cap_{m=1}^n A_m^\mathsf{c}) > 0$ for any n, for otherwise the conditional probabilities would be undefined. Let $p_n \stackrel{\text{def}}{=} \mathbb{P}(\cap_{m=1}^n A_m^\mathsf{c})$. Then we have that $p_n > 0$ for all n, and

$$\infty = \sum_{n=2}^\infty \mathbb{P}(A_n \mid \cap_{m=1}^{n-1} A_m^\mathsf{c}) \qquad (44a)$$

$$= \sum_{n=2}^\infty 1 - \mathbb{P}(A_n^\mathsf{c} \mid \cap_{m=1}^{n-1} A_m^\mathsf{c}) \qquad (44b)$$

$$= \lim_{N \to \infty} \sum_{n=2}^N 1 - \mathbb{P}(A_n^\mathsf{c} \mid \cap_{m=1}^{n-1} A_m^\mathsf{c}) \qquad (44c)$$

$$\leq \lim_{N \to \infty} \sum_{n=2}^N \log 1/\mathbb{P}(A_n^\mathsf{c} \mid \cap_{m=1}^{n-1} A_m^\mathsf{c}) \qquad \text{(by Eq. (43))} \qquad (44d)$$

$$= \lim_{N \to \infty} \sum_{n=2}^N \log \frac{\mathbb{P}(\cap_{m=1}^{n-1} A_m^\mathsf{c})}{\mathbb{P}(\cap_{m=1}^n A_m^\mathsf{c})} \qquad (44e)$$

$$= \lim_{N \to \infty} \sum_{n=2}^N \log \frac{p_{n-1}}{p_n} \qquad (44f)$$

$$= \lim_{N \to \infty} \sum_{n=2}^N (\log p_{n-1} - \log p_n) \qquad (44g)$$

$$= \lim_{N \to \infty} (\log p_1 - \log p_N) \qquad (44h)$$

$$= \log p_1 - \lim_{N \to \infty} \log p_N \qquad (44i)$$

which implies that

$$\lim_{N \to \infty} \log p_N = -\infty \qquad (45a)$$

$$\Leftrightarrow \quad \lim_{N \to \infty} p_N = 0 \tag{45b}$$

$$\Leftrightarrow \quad \lim_{N \to \infty} \mathbb{P}(\cap_{m=1}^N A_m^{\mathsf{c}}) = 0 \tag{45c}$$

$$\Leftrightarrow \quad \mathbb{P}(\cap_{m=1}^\infty A_m^{\mathsf{c}}) = 0. \qquad \text{(by continuity of measure)} \tag{45d}$$

∎

D.2 Proof of Prop. 4.3

Proposition 4.3. *If $\bar{p}(\text{EOS} \mid x) \geq f(t)$ for all $t \geq 1, x \in \Sigma^{t-1}$, and $\sum_{t=1}^\infty f(t) = \infty$, then $\mathbb{P}(\cap_{k=1}^\infty A_k^{\mathsf{c}}) = 0$. In other words, P is tight.*

Proof. In the proof, we rename the index t to n to match the usual presentation of the Borel-Cantelli lemmata. We are given that $\bar{p}(\text{EOS} \mid x) \geq f(n)$ for all $x \in \Sigma^{n-1}$. To apply Lemma 4.2, we observe that

$$A_n \cap (A_1^{\mathsf{c}} \cap \cdots \cap A_{n-1}^{\mathsf{c}}) = \{\boldsymbol{\omega} \in \overline{\Sigma}^\infty : \omega_n = \text{EOS}\} \cap \left(\bigcap_{i=1}^{n-1} \{\boldsymbol{\omega} \in \overline{\Sigma}^\infty : \omega_i \neq \text{EOS}\} \right) \tag{46a}$$

$$= \{\boldsymbol{\omega} \in \overline{\Sigma}^\infty : \boldsymbol{\omega} = \text{EOS}, \forall\, i < n, \boldsymbol{\omega} \neq \text{EOS}\} \tag{46b}$$

$$= \{\boldsymbol{\omega} \in \overline{\Sigma}^\infty : \boldsymbol{\omega}\text{'s first EOS is at position } n\} \tag{46c}$$

and similarly

$$A_1^{\mathsf{c}} \cap \cdots \cap A_{n-1}^{\mathsf{c}} = \{\boldsymbol{\omega} \in \overline{\Sigma}^\infty : \text{there is no EOS in } \boldsymbol{\omega}\text{'s first } n-1 \text{ positions}\} \tag{47}$$

Setting $G \overset{\text{def}}{=} \{\boldsymbol{\omega}\,\text{EOS} : \boldsymbol{\omega} \in \Sigma^{n-1}\} \subset \overline{\Sigma}^n$, we get

$$\mathbb{P}(A_n \mid A_1^{\mathsf{c}} \cap \cdots \cap A_{n-1}^{\mathsf{c}}) = \frac{\mathbb{P}(A_n \cap (A_1^{\mathsf{c}} \cap \cdots \cap A_{n-1}^{\mathsf{c}}))}{\mathbb{P}(A_1^{\mathsf{c}} \cap \cdots \cap A_{n-1}^{\mathsf{c}})} \tag{48a}$$

$$= \frac{\mathbb{P}(\overline{C}(G))}{\mathbb{P}(\overline{C}(\Sigma^{n-1}))} \qquad \text{(definition of } G\text{)} \tag{48b}$$

$$= \frac{\sum_{\boldsymbol{\omega} \in \Sigma^{n-1}} \bar{p}(\text{EOS} \mid \boldsymbol{\omega})\bar{p}(\boldsymbol{\omega})}{\sum_{\boldsymbol{\omega} \in \Sigma^{n-1}} \bar{p}(\boldsymbol{\omega})} \qquad \text{(by Eq. (8))} \tag{48c}$$

$$\geq \frac{\sum_{\boldsymbol{\omega} \in \Sigma^{n-1}} f(n)\bar{p}(\boldsymbol{\omega})}{\sum_{\boldsymbol{\omega} \in \Sigma^{n-1}} \bar{p}(\boldsymbol{\omega})} \qquad \text{(definition of } f(n)\text{)} \tag{48d}$$

$$= f(n)\frac{\sum_{\boldsymbol{\omega} \in \Sigma^{n-1}} \bar{p}(\boldsymbol{\omega})}{\sum_{\boldsymbol{\omega} \in \Sigma^{n-1}} \bar{p}(\boldsymbol{\omega})} \tag{48e}$$

$$= f(n). \tag{48f}$$

Since $\sum_{n=1}^\infty f(n) = \infty$ and hence $\sum_{n=2}^\infty f(n) = \infty$, the above inequality shows that the condition of Lemma 4.2 holds. Hence by Lemma 4.2, the event of a string never terminating, i.e., $\cap_{k=1}^\infty A_k^{\mathsf{c}}$, has probability measure $\mathbb{P}(\cap_{k=1}^\infty A_k^{\mathsf{c}}) = 0$.

In summary, if the EOS probability of a language model is lower-bounded at ever steps by the terms of a divergent series, then the event that this language model terminates has probability 1. ∎

D.3 Proof of Cor. 4.6

To show Cor. 4.6, we first show the following simple consequence of Borel–Cantelli.

Corollary D.1. *If $\mathbb{P}(A_n \text{ i.o.}) = 1$, then $\sum_{n=1}^\infty \mathbb{P}(A_n) = \infty$.*

Proof. Suppose to the contrary that $\sum_{n=1}^\infty \mathbb{P}(A_n) < \infty$, then, by Borel–Cantelli I (Lemma 4.4), $\mathbb{P}(A_n \text{ i.o.}) = 0$, which contradicts the assumption. Hence, $\sum_{n=1}^\infty \mathbb{P}(A_n) = \infty$.

∎

Cor. 4.6 below is also stated without proof as Exercise 4.3.4 in Durrett (2019).

Corollary 4.6. *Given a sequence $\{p_n\}$ where $p_n \in [0, 1)$. Then,*

$$\prod_{n=1}^{\infty}(1 - p_n) = 0 \iff \sum_{n=1}^{\infty} p_n = \infty. \tag{19}$$

Proof. We can use a product measure to construct a sequence of independent events $\{A_n\}_{n=1}^{\infty}$ such that $\mathbb{P}(A_n) = p_n$. (The product measure ensures independence.) Then, by definition in Eq. (18),

$$\{A_n \text{ i.o.}\}^{\mathsf{c}} = \bigcup_{m=1}^{\infty} \bigcap_{n \geq m} A_n^{\mathsf{c}} \tag{49}$$

So,

$$1 - \mathbb{P}(A_n \text{ i.o.}) = \mathbb{P}\left(\bigcup_{m} \bigcap_{n \geq m} A_n^{\mathsf{c}}\right) \tag{50a}$$

$$= \lim_{m \to \infty} \mathbb{P}\left(\bigcap_{n \geq m} A_n^{\mathsf{c}}\right) \tag{50b}$$

$$= \lim_{m \to \infty} \prod_{n \geq m} \mathbb{P}(A_n^{\mathsf{c}}) \qquad (A_n \text{ are independent by construction}) \tag{50c}$$

$$= \lim_{m \to \infty} \prod_{n \geq m} (1 - p_n) \tag{50d}$$

$(\Rightarrow)$: Assume $\prod_{n=1}^{\infty}(1 - p_n) = 0$. Then, for any m,

$$0 = \prod_{n \geq 1}(1 - p_n) = \underbrace{\left(\prod_{1 \leq n < m}(1 - p_n)\right)}_{>0}\left(\prod_{n \geq m}(1 - p_n)\right) \tag{51}$$

So it must the case that, for any m, $\prod_{n \geq m}(1 - p_n) = 0$. Therefore,

$$1 - \mathbb{P}(A_n \text{ i.o.}) = \lim_{m \to \infty} \prod_{n \geq m}(1 - p_n) = 0 \tag{52}$$

which implies $\mathbb{P}(A_n \text{ i.o.}) = 1$. Cor. D.1 implies that $\sum_{n=1}^{\infty} p_n = \infty$.

$(\Leftarrow)$: Assume $\sum_{n=1}^{\infty} p_n = \infty$. Then by Borel–Cantelli II (Lemma 4.5), $\mathbb{P}(A_n \text{ i.o.}) = 1$ which implies

$$0 = 1 - \mathbb{P}(A_n \text{ i.o.}) = \lim_{m \to \infty} \prod_{n \geq m}(1 - p_n) \tag{53}$$

Observe that $\left\{\prod_{n \geq m}(1 - p_n)\right\}_m$ is a non-decreasing sequence in m; to see this, note that as m grows larger we multiply strictly fewer values $(1 - p_n) \in (0, 1]$. However, since we know the sequence is non-negative and tends to 0, it follows that for *any* m, we have

$$\prod_{n \geq m}(1 - p_n) = 0. \tag{54}$$

It follows that, for any m, we have

$$\prod_{n=1}^{\infty}(1 - p_n) = \prod_{n < m}(1 - p_n) \underbrace{\prod_{n \geq m}(1 - p_n)}_{=0} = \prod_{n < m}(1 - p_n) \cdot 0 = 0. \tag{55}$$

∎

D.4 Proof of Thm. 4.7

Theorem 4.7 (Proposition 2.4 in Meister et al., 2022)**.** *An ASM is tight if and only if $\widetilde{p}_{\text{EOS}}(t) = 1$ for some t or $\sum_{t=1}^{\infty} \widetilde{p}_{\text{EOS}}(t) = \infty$.*

Proof. Recall the definition of $\widetilde{p}_{\text{EOS}}$, as previously defined in Eq. (20), is

$$\widetilde{p}_{\text{EOS}}(t) \overset{\text{def}}{=} \mathbb{P}(A_t \mid A_1^{\mathsf{c}} \cap \cdots \cap A_{t-1}^{\mathsf{c}}). \tag{56}$$

Case 1. Suppose that $\widetilde{p}_{\text{EOS}}(t) < 1$ for all t. Consider the termination probability again:

$$\mathbb{P}\left(\bigcap_{t=1}^{\infty} A_t^{\mathsf{c}}\right) = \lim_{T \to \infty} \mathbb{P}\left(\bigcap_{t=1}^{T} A_t^{\mathsf{c}}\right) \tag{57a}$$

$$= \lim_{T \to \infty} \prod_{t=1}^{T} \mathbb{P}(A_t^{\mathsf{c}} \mid A_1^{\mathsf{c}} \cap \cdots \cap A_{t-1}^{\mathsf{c}}) \tag{57b}$$

$$= \lim_{T \to \infty} \prod_{t=1}^{T} (1 - \widetilde{p}_{\text{EOS}}(t)) \tag{57c}$$

$$= \prod_{t=1}^{\infty} (1 - \widetilde{p}_{\text{EOS}}(t)). \tag{57d}$$

In the above, we have assumed that $\mathbb{P}(A_1^{\mathsf{c}} \cap \cdots \cap A_t^{\mathsf{c}}) > 0$ for all t, which is true by assumption that $\widetilde{p}_{\text{EOS}}(t) < 1$.. Hence, by Cor. 4.6, Eq. (57d) is 0 if and only if $\sum_t \widetilde{p}_{\text{EOS}}(t) = \infty$.

Case 2. If $\widetilde{p}_{\text{EOS}}(t) = 1$ is true for some $t = t_0$, then $\mathbb{P}(A_1^{\mathsf{c}} \cap \cdots \cap A_{t_0}^{\mathsf{c}}) = 0$ and hence $\mathbb{P}\left(\bigcap_{t=1}^{\infty} A_t^{\mathsf{c}}\right) = 0$ and such a language model is guaranteed to terminate at t_0. ∎

E Proofs for Analyses of Common Language Models (§5)

E.1 Proofs for FSSMs (§5.1)

E.1.1 Proofs for Stochastic FSSMs

Theorem 5.3. *An SFSSM is tight iff all accessible states are also co-accessible.*

Proof. We refer to a state q as **initial** if $s_q > 0$ and as **final** if $t_q > 0$. (These are sometimes called source and sink states.) We prove each direction of the theorem in turn:

($\Rightarrow$): Assume the SFSSM is tight. Let q be an accessible state. Since the SFSSM has at least one positive-probability path from an initial state, there is a positive probability of reaching q during generation. If there were no positive-probability path from q to a final state, then the SFSSM would never terminate on the occasions when it reached q, contradicting the assumption of tightness. Hence q must be co-accessible.

($\Leftarrow$): Assume that all accessible states are co-accessible. We construct a Markov chain whose states are the SFSSM's accessible states $Q_A \subseteq \{1, \ldots, Q\}$ together with an EOS state. In this Markov chain, the initial probability of q is given by s_q when $q \in Q_A$ and by 0 when $q = $ EOS; the transition probability from q to q' is given by $\mathbf{P}_{qq'}$ when $q, q' \in Q_A$, by t_q when $q \in Q_A$ and $q' = $ EOS, by 1 when $q = q' = $ EOS, and by 0 otherwise. The probability that the Markov chain is in state $q \in Q_A$ after t steps equals the probability that the SFSSM is in state q after t steps (note that the SFSSM never reaches any state $q \notin Q_A$). The probability that it is in state EOS after t steps equals the probability that the SFSSM has terminated after $\leq t$ steps.

Clearly EOS is an **absorbing state** of the Markov chain, meaning that once the Markov chain reaches this state, it never leaves. A fundamental result on finite-state Markov chains (Grinstead and Snell, 1997, Theorem 11.3) is that if every state can reach an absorbing state, then with probability 1, the chain reaches an absorbing state ("is absorbed") in finite time. Every state can in fact reach EOS, by coaccessibility of Q_A. This further implies that EOS is the *only* absorbing state (as an absorbing state cannot reach any

other state). So by the result cited above, the Markov chain reaches EOS with probability 1 in finite time. Consequently, the SFSSM terminates after finitely many steps with probability 1; that is, the SFSSM is tight. ∎

Corollary 5.4. *Maximum likelihood estimates of n-gram models based on some corpus are tight.*

Proof. The SFSSM for an n-gram model has states that correspond to $(n-1)$-grams and transitions that correspond to characters (unigrams), as illustrated by Fig. 1. When the SFSSM's probabilities are estimated with MLE, the accessible states are $(n-1)$-grams that have appeared in some string in the corpus. Such states must also be co-accessible so that they can generate the rest of that string. Hence, by Thm. 5.3, this SFSSM is tight. ∎

E.1.2 Proofs for Substochastic FSSMs

To prove Thm. 5.6, we will make use of the following useful lemma.

Lemma E.1. *Let $\mathbf{P}'$ be the transition sum matrix of a trimmed substochastic FSSM. Then $\rho(\mathbf{P}') < 1$ where $\rho(\cdot)$ denotes the spectral radius.*

Proof. To begin with, we wish to apply the following result, which connects the row sums of a matrix to its spectral radius. Below, M_n denotes the set of $n \times n$ matrices, and $\|A\|_\infty = \max_{1 \le i \le n} \sum_{j=1}^n |A_{ij}|$ denotes the operator ∞-norm.

Proposition E.2 (§6.2.P8, Horn and Johnson, 2012). *For any $A \in M_n$, $\rho(A) \le \|A\|_\infty$. Additionally, if A is irreducible and not all absolute row sums of A are equal, then $\rho(A) < \|A\|_\infty$.*

However, the transition sum matrix $\mathbf{P}$ of a substochastic FSSM may be reducible, whereas the irreducibility condition in Prop. E.2 cannot be dropped. Hence, we need to "decompose" $\mathbf{P}'$ in a way that recovers irreducibility. We use the *Frobenius normal form* (also known as *irreducible normal form*) to achieve this.

Proposition E.3 (§8.3.P8, Horn and Johnson, 2012). *Let $A \in M_n$ be non-negative. Then, either A is irreducible or there exists a permutation matrix Π such that*

$$\Pi^\top A \Pi = \begin{bmatrix} A_1 & & * \\ & \ddots & \\ \mathbf{0} & & A_k \end{bmatrix} \tag{58}$$

*is block upper triangular, and each diagonal block is irreducible (possibly a 1×1 zero matrix). This is called a **Frobenius normal form** (or **irreducible normal form**) of A. Additionally, $\lambda(A) = \lambda(A_1) \cup \cdots \cup \lambda(A_k)$ where $\lambda(\cdot)$ denotes the set of eigenvalues of a matrix.*

Notice that the permutation in the Frobenius normal form merely renumbers the states of the trimmed FSSM. We may check that as a result, the termination probability given in Thm. 5.6 is unchanged:[26]

$$(\Pi^\top \boldsymbol{s}')^\top (\Pi^\top \mathbf{P}' \Pi)^k (\Pi^\top \boldsymbol{t}') = (\boldsymbol{s}'^\top \Pi)(\Pi^\top \mathbf{P}'^k \Pi)(\Pi^\top \boldsymbol{t}') = \boldsymbol{s}'^\top \mathbf{P}'^k \boldsymbol{t}' \tag{59}$$

Hence, with an appropriate renumbering, we may assume without loss of generality that $\mathbf{P}$ is already given in Frobenius normal form

$$\mathbf{P}' = \begin{bmatrix} \mathbf{P}'_1 & & * \\ & \ddots & \\ \mathbf{0} & & \mathbf{P}'_k \end{bmatrix} \tag{60}$$

where each $\mathbf{P}'_i$ is irreducible.

Since the transition sum matrix $\mathbf{P}'$ of a trimmed substochastic FSSM is a substochastic matrix, each $\mathbf{P}'_i$ is also substochastic. In fact, each $\mathbf{P}'_i$ is *strictly* substochastic, meaning that there is at least one row that

---

[26]The equalities here use the fact that the inverse of a permutation matrix Π is its transpose: $\Pi \Pi^\top = I$.

sums to less than 1. To see this, suppose to the contrary that there is a stochastic $\mathbf{P}'_i$. Since the FSSM is trimmed, every state is both accessible and co-accessible. Being accessible implies that there is a positive probability of reaching every state in $\mathbf{P}'_i$. However, the stochasticity of $\mathbf{P}'_i$ forces the corresponding $\boldsymbol{t}'$ entries to be 0. Hence, none of these states can transition to EOS, meaning that they're not co-accessible, contradicting the assumption. Hence, every $\mathbf{P}'_i$ is strictly substochastic. Then, either all row sums of $\mathbf{P}'_i$ are less than 1 (in which case $\|\|\mathbf{P}'_i\|\|_\infty < 1$) or some row sums are 1 and some are less than 1 (in which case $\|\|\mathbf{P}'_i\|\|_\infty = 1$ and $\mathbf{P}'$ has unequal absolute row sums). In either case, Prop. E.2 implies that $\rho(\mathbf{P}'_i) < 1$, for all $1 \le i \le k$.

Finally, the last sentence of Prop. E.3 entails that $\rho(\mathbf{P}') = \max\{\rho(\mathbf{P}'_1), \dots, \rho(\mathbf{P}'_k)\}$. Since each $\rho(\mathbf{P}'_i) < 1$, we have $\rho(\mathbf{P}') < 1$. ∎

Theorem 5.6. *Let $\mathbf{P}'$ be the transition sum matrix of a trimmed substochastic FSSM. Then $\mathbf{I} - \mathbf{P}'$ is invertible and $P(X \in \Sigma^*) = \boldsymbol{s}'^\top(\mathbf{I} - \mathbf{P}')^{-1}\boldsymbol{t}' \le 1$.*

Proof. By Lemma E.1, $\rho(\mathbf{P}') < 1$, in which case $\mathbf{I} - \mathbf{P}'$ is invertible and the Neumann series $\sum_{k=0}^\infty \mathbf{P}'^k = \mathbf{I} + \mathbf{P}' + \mathbf{P}'^2 + \cdots$ converges to $(\mathbf{I} - \mathbf{P}')^{-1}$ (Horn and Johnson, 2012, §5.6). Thus

$$P(\Sigma^*) = \sum_{k=0}^\infty P(\Sigma^k) = \sum_{k=0}^\infty \boldsymbol{s}'^\top \mathbf{P}'^k \boldsymbol{t}' = \boldsymbol{s}'^\top \left(\sum_{k=0}^\infty \mathbf{P}'^k \right) \boldsymbol{t}' = \boldsymbol{s}'^\top (\mathbf{I} - \mathbf{P}')^{-1} \boldsymbol{t}'. \tag{61}$$

∎

E.2 Proofs for Transformer Result (§5.2)

Again, the following theorem is well-known:

Theorem 5.7. *Let X be a compact topological space and Y be any topological space. If $f : X \to Y$ is continuous, then $f(X) \subseteq Y$ is also compact.*

Proof. Let $\{U_\alpha\}_{\alpha \in \mathcal{A}}$ be any open cover of $f(X)$. By continuity, $f^{-1}(U_\alpha) \subset X$ is open for any $\alpha \in \mathcal{A}$, and hence $\{f^{-1}(U_\alpha)\}_{\alpha \in \mathcal{A}}$ is also an open cover of X. By the compactness of X, there is a finite sub-cover $\{f^{-1}(U_{\alpha_i})\}_{i=1}^n$, in which case $\{U_{\alpha_i}\}_{i=1}^n$ forms a finite sub-cover for $f(X)$. ∎

Lemma 5.8. *Let $f : (\mathbb{R}^d)^+ \to (\mathbb{R}^d)^+$ be the function defined by a finite number of Transformer layers (e.g., n layers) with any continuous activation function. Given a compact set $K \subset \mathbb{R}^d$. Then, there exists a compact set $K' \subset \mathbb{R}^d$ such that for every $t \in \mathbb{Z}_{>0}$,*

$$f(K^t) \subseteq (K')^t. \tag{24}$$

Note. We make use of the following notations in the proof below: $\triangle_{t-1} = \{\boldsymbol{y} \in \mathbb{R}^t : \boldsymbol{y} \ge 0, \boldsymbol{1}^\top \boldsymbol{y} = 1\}$ denotes the $(t-1)$-dimensional simplex; $B_r(\boldsymbol{z}) = \{\boldsymbol{v} \in \mathbb{R}^n : \text{dist}(\boldsymbol{z}, \boldsymbol{v}) < r\}$ denotes the open ball centered at $\boldsymbol{z}$ with radius r; $\overline{A}$ denotes the closure of set A.

Proof. Let $K_0 = K$. In an autoregressive transformer, each of the n layers consists of two blocks: a self-attention block and a feedforward block. We will use induction on the $2n$ blocks to build up compact sets $K_1, K_2, \dots, K_{2n}$ that contain the output vectors of these respective blocks, and then take $K' = K_{2n}$.

The self-attention block is a function on $(\mathbb{R}^d)^+ \to (\mathbb{R}^d)^+$. So, let $t \in \mathbb{Z}_{>0}$ be arbitrary and consider any sequence of input vectors $(\boldsymbol{v}_1, \dots, \boldsymbol{v}_t)$ such that for all i, $\boldsymbol{v}_i \in K_0$. Denote the output vectors of the attention block with $(\boldsymbol{v}'_1, \dots, \boldsymbol{v}'_t)$. By definition of attention, each output vector $\boldsymbol{v}'_j = \sum_{i=1}^t \alpha_i^{(j)} \boldsymbol{v}_i$ where $\boldsymbol{\alpha}^{(j)} \in \triangle_{t-1}$ are the attention weight vectors obtained through the softmax function. Compact sets in $\mathbb{R}^d$ are bounded (by the Heine–Borel theorem), and hence there exists $M > 0$ such that $K_0 \subseteq \overline{B_M(0)}$. Noting that the norm function $\| \cdot \|$ on $\mathbb{R}^d$ is convex, we have the following

$$\|\boldsymbol{v}'_j\| = \left\| \sum_{i=1}^t \alpha_i^{(j)} \boldsymbol{v}_i \right\| \tag{62a}$$

$$\leq \sum_{i=1}^{t} \alpha_i^{(j)} \|\boldsymbol{v}_j\| \tag{$*$}$$

$$\leq \sum_{i=1}^{t} \alpha_i^{(j)} M = M \tag{62b}$$

where ($*$) results from Jensen's inequality. Eq. (62b) shows that each of the output vectors $\boldsymbol{v}_j'$ lies in $\overline{B_M(0)}$ which is compact. Hence, setting $K_1 = \overline{B_M(0)}$, we have shown that, for any $t \in \mathbb{Z}_{>0}$, the attention block maps K_0^t into K_1^t.

Note that we *cannot* use Thm. 5.7 here because the attention block defines a different function on $\mathbb{R}^{t \times d} \to \mathbb{R}^{t \times d}$ for each t, and Thm. 5.7 only implies that there exists a separate *length-dependent* output compact set $K_t \subset \mathbb{R}^{t \times d}$ for each t, which is different from this lemma's statement.

The feedforward function is a continuous function on $\mathbb{R}^d \to \mathbb{R}^d$, and therefore, by Thm. 5.7, maps its input compact set K_1 to an output compact set, which we call K_2.

Finally, residual connections and layer norms are also continuous functions acting on each of the input vectors, and hence by the same reasoning would also preserve compactness.

Now we can use induction and show that there exist compact sets $K_3, K_4, \ldots, K_{2n-1}, K_{2n}$ where K_{2n} contains the output set of the final layer. Set $K' = K_{2n}$ and we have proven the statement. ∎

Theorem 5.9. *The autoregressive sequence model defined by any (fixed-depth) Transformer is tight.*

Proof. Given the Transformer, there exists a fixed compact set K that will contain all inputs $\boldsymbol{v}_i \in \mathbb{R}^d$ to the first layer. This is true because each $\boldsymbol{v}_i$ is the sum of a word embedding, which falls in a finite set since $\overline{\Sigma}$ is finite, and a position embedding, which lies in the compact set $[-1, 1]^d$. Hence, by Lemma 5.8, there exists a fixed compact set K' that contains all output embedding vectors (regardless of how long the sequence is).

The final output probability is given by a multiplication with the word embedding matrix followed by the softmax function as in Eq. (25). This process amounts to composing two continuous functions. In particular, we can extract the EOS probability as a continuous $\mathbb{R}$-valued function $g_{\text{EOS}} : K' \to (0, 1)$ (neither 0 or 1 is in the range of the softmax function). By continuity of g_{EOS} and Thm. 5.7, $K'' \overset{\text{def}}{=} g_{\text{EOS}}(K') \subseteq (0, 1)$ is compact. Since K'' is compact, and hence closed, $\inf K'' \in K''$. Thus $\inf K'' \in (0, 1)$ and in particular $\inf K'' > 0$. Therefore, taking $\epsilon = \inf K''$, we have shown that the EOS probability of a Transformer is bounded below by some $\epsilon > 0$ (regardless of the length of the sequence).

Hence, by Prop. 4.3, any Transformer ASM is tight and thus defines a language model. ∎

E.3 Proofs for RNN Result (§5.3)

Proposition 5.10. *Given an RNN ASM over $\overline{\Sigma}$. Again let the output symbol vector be $\boldsymbol{u}_x \in \mathbb{R}^d$ for $x \in \overline{\Sigma}$, and set $k \overset{\text{def}}{=} \sup_{x \in \Sigma} \|\boldsymbol{u}_x - \boldsymbol{u}_{\text{EOS}}\|_2$. Additionally, for each $t > 0$, let $\|\widehat{\boldsymbol{h}}_t\|_2$ be the maximum attainable hidden state norm for any context $\boldsymbol{x} \in \Sigma^t$. Such a sequence model is tight if $k \|\widehat{\boldsymbol{h}}_t\|_2 \leq \log t$ for all sufficiently large t.*

Proof. Let $X_t(\boldsymbol{\omega})$ be the random variable that is equal to the tth token in an outcome $\boldsymbol{\omega} \in \Omega$. Also let $\boldsymbol{h}_{\boldsymbol{x}}$ be the hidden representation of the RNN after processing some finite list of tokens $\boldsymbol{x} \in \Sigma^*$. Further, let $\boldsymbol{u}_x \in \mathbb{R}^d$ be the output embedding of $x \in \overline{\Sigma}$, Then for any $t \in \mathbb{N}$ and any $\boldsymbol{x} \in \Sigma^t$, we have:

$$\mathbb{P}(X_{t+1} = \text{EOS} \mid X_{\leq t} = \boldsymbol{x}) = \frac{\exp \boldsymbol{u}_{\text{EOS}}^\top \boldsymbol{h}_{\boldsymbol{x}}}{\sum_{y \in \overline{\Sigma}} \exp \boldsymbol{u}_y^\top \boldsymbol{h}_{\boldsymbol{x}}} \tag{63a}$$

$$= \frac{1}{\sum_{y \in \overline{\Sigma}} \exp \boldsymbol{u}_y^\top \boldsymbol{h}_{\boldsymbol{x}} / \exp \boldsymbol{u}_{\text{EOS}}^\top \boldsymbol{h}_{\boldsymbol{x}}} \tag{63b}$$

$$= \frac{1}{1 + \sum_{y \in \Sigma} \exp(\boldsymbol{u}_y - \boldsymbol{u}_{\text{EOS}})^\top \boldsymbol{h}_{\boldsymbol{x}}} \tag{63c}$$

9767

$$\geq \frac{1}{1 + \sum_{y \in \Sigma} \exp\left(\|\boldsymbol{u}_y - \boldsymbol{u}_{\text{EOS}}\|_2 \|\boldsymbol{h}_{\boldsymbol{x}}\|_2\right)} \qquad \text{(Cauchy–Schwarz)} \qquad \text{(63d)}$$

$$\geq \frac{1}{1 + \sum_{y \in \Sigma} \exp(k\|\boldsymbol{h}_{\boldsymbol{x}}\|_2)} \qquad \text{(63e)}$$

$$= \frac{1}{1 + |\Sigma| \exp(k\|\boldsymbol{h}_{\boldsymbol{x}}\|_2)} \qquad \text{(63f)}$$

Now define $\|\widehat{\boldsymbol{h}}_t\|_2 \overset{\text{def}}{=} \sup_{\boldsymbol{x} \in \Sigma^t} \|\boldsymbol{h}_{\boldsymbol{x}}\|_2$. We then have that $\forall t \in \mathbb{N}$ and $\forall \boldsymbol{x} \in \Sigma^t$:

$$\mathbb{P}(X_{t+1} = \text{EOS} \mid X_{\leq t} = \boldsymbol{x}) \geq \frac{1}{1 + |\Sigma| \exp(k\|\widehat{\boldsymbol{h}}_t\|_2)} \qquad (64)$$

Now, by Prop. 4.3, we have that if $\sum_{t=0}^{\infty} \frac{1}{1+|\Sigma|\exp(k\|\widehat{\boldsymbol{h}}_t\|_2)}$ diverges, then the language model is tight. We will show that this condition holds if $\exists N \in \mathbb{N}$ such that $\forall t \geq N$, $k\|\widehat{\boldsymbol{h}}_t\|_2 \leq \log t$.

First, note that $\lim_{t \to \infty} \frac{1}{t} \frac{1+|\Sigma|t}{1} = \lim_{t \to \infty} \frac{1}{t} + |\Sigma| = |\Sigma| \in (0, \infty)$. Hence, by the limit comparison test, since $\sum_{t=1}^{\infty} \frac{1}{t}$ diverges, this means $\sum_{t=1}^{\infty} \frac{1}{1+|\Sigma|t}$ must also diverge.

Now, suppose there exists N such that that $k\|\widehat{\boldsymbol{h}}_t\|_2 \leq \log t$ for all $t \geq N$. This implies that for $t \geq N$ we have $\frac{1}{1+|\Sigma|\exp(k\|\widehat{\boldsymbol{h}}_t\|_2)} \geq \frac{1}{1+|\Sigma|t}$, which combined with the above and the comparison test, implies that $\sum_{t=N}^{\infty} \frac{1}{1+|\Sigma|\exp(k\|\widehat{\boldsymbol{h}}_t\|_2)}$ diverges. This in turn means that $\sum_{t=0}^{\infty} \frac{1}{1+|\Sigma|\exp(k\|\widehat{\boldsymbol{h}}_t\|_2)}$ diverges.

Hence, if $k\|\widehat{\boldsymbol{h}}_t\|_2 \leq \log t$ for all sufficiently large t (that is, for all $t \geq N$), then the RNN ASM is tight and thus defines a language model.

■

A For every submission:

☑ A1. Did you describe the limitations of your work?
7

☒ A2. Did you discuss any potential risks of your work?
Our paper provides theoretical analysis of language models. We do not expect potential risks as a result of this work.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
No response.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☒ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No response.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.