When is a Language Process a Language Model?

Li Du[1] **Holden Lee**[1] **Jason Eisner**[1] **Ryan Cotterell**[2]
[1]Johns Hopkins University [2]ETH Zürich
leodu@cs.jhu.edu hlee283@jhu.edu jason@cs.jhu.edu
ryan.cotterell@inf.ethz.ch

Abstract

A language model may be viewed as a Σ-valued stochastic process for some alphabet Σ. However, in some pathological situations, such a stochastic process may "leak" probability mass onto the set of infinite strings and hence is not equivalent to the conventional view of a language model as a distribution over ordinary (finite) strings. Such ill-behaved language processes are referred to as *non-tight* in the literature. In this work, we study conditions of tightness through the lens of stochastic processes. In particular, by regarding the EOS symbol as marking a stopping time and using results from martingale theory, we give characterizations of tightness that generalize our previous work (Du et al., 2023).

1 Introduction

Mathematically, there are two views of language models. From a formal language perspective, a language model is viewed as a distribution over Σ^*, the set of (finite-length) strings over a finite alphabet Σ (Booth and Thompson, 1973). From a probabilistic perspective, a language model is viewed as a discrete stochastic process $\{X_t \in \Sigma : t \in \mathbb{Z}^+\}$ (Markov, 1913; Shannon, 1948). Under the latter view, to signal the end of the string, it is common to augment the alphabet with a special end-of-string symbol, EOS, and define the X_t to be $\overline{\Sigma}$-valued random variables where $\overline{\Sigma} = \Sigma \cup \{\text{EOS}\}$. Although X_t is defined for all $t > 0$, the random string consists only of the symbols X_t preceding the first EOS. We call such a stochastic process a *language process*, following Meister et al. (2022).

While the literature rarely distinguishes between the two views, there is an important difference: the probability mass of a language process can "leak" onto the set of infinite strings, meaning that < 1 of the probability mass is assigned to the set of (finite) strings.[1] When such leakage happens, the language process is said to be *non-tight* (Chi, 1999; Cohen and Johnson, 2013) and cannot be regarded as a language model (i.e., a distribution over Σ^*). Non-tightness may cause language processes to generate infinite sequences (Welleck et al., 2020) and bias Monte Carlo samplers over strings (Lew et al., 2023). This phenomenon raises a natural question: How can we mathematically determine whether a language process is a language model?

As demonstrated by several examples in Du et al. (2023, Section 2), a precise answer to this question requires some probability theory (Kolmogorov, 1933). To this end, Du et al. (2023) formalize language processes (which they call "sequence models") as extensions of pre-measures, and then obtain characterizations of tightness using the Borel–Cantelli lemmas.

In this work, we revisit the problem of characterizing tightness through the lens of stochastic processes, an intuitive, high-level construct in measure-theoretic probability. We begin by presenting language processes using the Kolmogorov Extension Theorem (§2), which establishes their existence more simply and directly than the construction of Du et al. (2023). We then formulate EOS as a stopping time, yielding a natural understanding of EOS that connects the tightness property to martingales (§3.1).[2] Applying results from martingale theory, we are able to characterize tightness more generally than in previous work (§3.2).

2 Language Models as Stochastic Processes

In probability theory, a *stochastic process* is a collection of random variables $\{X_t : t \in T\}$ on some probability space $(\overline{\Omega}, \overline{\mathscr{F}}, P)$ over some index set T.[3] Language models are often framed in terms of stochastic processes, usually implicitly through

[1]A string has finite length by definition (Sipser, 2013), so an "infinite string" is really a sequence, not a string.

[2]The study of optional stopping is a major motivation for the development of martingale theory (Doob, 1953).

[3]See [M1]. Throughout this paper, we will freely use concepts from measure-theoretic probability. To be self-contained, we summarize necessary definitions and background in App. B and refer to it when appropriate.

Findings of the Association for Computational Linguistics ACL 2024, pages 11083–11094
August 11-16, 2024

the notation X_t to refer to the t^{th} word. We first make this intuitive notation rigorous by formulating X_t as a random variable over a certain probability space, which we must construct. Previous papers simply assume the existence of such a probability space, occasionally explicitly (Meister et al., 2022).

It is popular to specify a language model in terms of a parametric family of autoregressive conditional probabilities $\{p(x_t \mid \boldsymbol{x}_{<t}) : x_t \in \overline{\Sigma}, \boldsymbol{x}_{<t} \in \overline{\Sigma}^*\}$. However, we need to show that the stochastic process $\{X_t : t \in T\}$ actually exists—that is, there is a distribution over infinite sequences in $\overline{\Sigma}^\infty$ that has the given conditional probabilities. We will use the canonical tool for this, the Kolmogorov extension theorem. The theorem states that a collection of distributions over finite subsets of the variables X_t can be extended to a joint distribution over all of the infinitely many variables X_t, provided that the distributions in the collection agree on the variables where they overlap (a consistency condition).

As is the case for many fundamental theorems, the Kolmogorov extension theorem has many variants. Most probability texts (Chung, 1974; Billingsley, 1995; Durrett, 2019) only state and prove the theorem for $\mathbb{R}$-valued stochastic processes, whereas we are interested in constructing a $\overline{\Sigma}$-valued stochastic process. For completeness and rigor, we state a discrete version of the theorem below and provide a proof.

Theorem 2.1 (Kolmogorov Extension Theorem)**.** *Let T be an arbitrary index set, and $(\Omega, \mathscr{F})$ be a finite measurable space where $\mathscr{F}$ is the discrete σ-algebra.*[4] *Define $\mathscr{F}^k$ and $\mathscr{F}^T$ as in [M3]. Given a system of measures $\{\mu_*\}$ where for each k-tuple $(t_1, \ldots, t_k)$ of distinct elements in T, $\mu_{t_1,\ldots,t_k}$ is a measure over $(\Omega^k, \mathscr{F}^k)$, and where for every choice of $H_1, \ldots H_k \in \mathscr{F}$, these measures satisfy*

(1) For all permutations π,

$$\mu_{t_1,\ldots,t_k}(H_1 \times \cdots \times H_k) = \mu_{t_{\pi(1)},\ldots,t_{\pi(k)}}(H_{\pi(1)} \times \cdots \times H_{\pi(k)}) \tag{1}$$

(2) For arbitrary distinct k-tuples $(t_1, \ldots, t_k)$,

$$\mu_{t_1,\ldots,t_{k-1}}(H_1 \times \cdots \times H_{k-1}) = \mu_{t_1,\ldots,t_{k-1},t_k}(H_1 \times \cdots \times H_{k-1} \times \Omega). \tag{2}$$

Then, there is a unique probability measure P on $(\Omega^T, \mathscr{F}^T)$ such that the coordinate random variables $\{X_t : t \in T\}$ have $\mu_{t_1,\ldots,t_k}$ as their finite-dimensional distributions, i.e.,

$$P\big((X_{t_1}, \ldots, X_{t_k}) \in A\big) = \mu_{t_1,\ldots,t_k}(A) \tag{3}$$

for all $A \in \mathscr{F}^k$.

Proof. App. C. □

In Theorem 2.1, the outcomes in the probability space may be denoted as $\boldsymbol{\omega} \in \Omega^T$ and the random variables X_t are defined by $X_t(\boldsymbol{\omega}) = \boldsymbol{\omega}_t \in \Omega$, which extract the respective elements of $\boldsymbol{\omega}$.[5]

We now show that Theorem 2.1 enables a simple construction of the language process from the autoregressive conditionals, where $\Omega = \overline{\Sigma}$ and the index set $T = \mathbb{Z}^+ = \{1, 2, \ldots\}$.

For $t \geq 0$ and $\boldsymbol{x} \in \overline{\Sigma}^t$, define $p^t(\boldsymbol{x}) = \prod_{s=1}^{t} p(x_s \mid \boldsymbol{x}_{<s})$. These finite-dimensional distributions naturally yield a system of finite-dimensional measures $\{\mu_*\}$ where each $\mu_{t_1,\ldots,t_k}$ is derived from some p^t by sorting the indices and marginalizing over skipped indices. This process is detailed below.

We first define the measures for consecutive index tuples of the form $(1, \ldots, t)$:

$$\mu_{1,\ldots,t}(H_1 \times \cdots \times H_t) \stackrel{\text{def}}{=} \sum_{\boldsymbol{x} \in H_1 \times \cdots \times H_t} p^t(\boldsymbol{x}). \tag{4}$$

We now use marginalization to extend to arbitrary sorted k-tuples $(t_1, \ldots, t_k)$ where $t_1 < \cdots < t_k$, by defining

$$\mu_{t_1,\ldots,t_k}(H_1 \times \cdots \times H_k) \stackrel{\text{def}}{=} \mu_{1,\cdots,t_k}\left(\prod_{t=1}^{t_k} G_t\right) \tag{5}$$

where $G_t = H_t$ if $t = t_i$ for some i and otherwise $G_t = \overline{\Sigma}$. Finally, we use permutation to extend to unsorted tuples. For an arbitrary k-tuple $(t_1, \ldots, t_k)$ of distinct elements, we define

$$\mu_{t_1,\ldots,t_k}(H_1 \times \cdots \times H_k) \stackrel{\text{def}}{=} \mu_{t_{\pi(1)},\ldots,t_{\pi(k)}}(H_{\pi(1)} \times \cdots \times H_{\pi(k)}) \tag{6}$$

where π is the unique permutation that sorts the elements: $t_{\pi(1)} < \cdots < t_{\pi(k)}$.

By construction, the family of measures $\{\mu_*\}$ defined above satisfies both conditions in Theorem 2.1. Hence, Theorem 2.1 implies that there exists a unique probability measure

[4] That is, $\mathscr{F}$ is the power set of Ω.

[5] It is possible to state a weaker version of Theorem 2.1 without the first condition by requiring $t_1 < \cdots < t_k$ in the system of measures $\{\mu_*\}$, as is the case in Durrett (2019). However, this requires choosing a total order of the index set T.

$(\overline{\Sigma}^{\mathbb{Z}^+}, \mathscr{F}^{\mathbb{Z}^+}, P)$ such that the coordinate random variables $\{X_t : t \in \mathbb{Z}^+\}$ have $\{\mu_*\}$ as its finite-dimensional distributions. Since $\{\mu_*\}$ is directly derived from a family of autoregressive conditionals $\{p(\cdot \mid \boldsymbol{x}_{<t}) : \boldsymbol{x}_{<t} \in \overline{\Sigma}^*\}$, the collection of random variables $\{X_t : t \in \mathbb{Z}^+\}$ *is* the corresponding language process.

3 Characterizing Tightness

Having established the existence of a language process with given autoregressive conditional probabilities, we can readily apply tools from the theory of stochastic processes. We will see that two key subjects of our investigation, EOS termination and tightness, are naturally characterized by stopping times. Stopping times are central to the study of stochastic processes (Doob, 1953).[6] This connection will allow us to derive tightness conditions that generalize Du et al. (2023).

3.1 Stopping Time

To define stopping time, we first review the necessary definition of filtration. Given a stochastic process $\{X_t : t \in \mathbb{Z}^+\}$ on a probability space $(\Omega^T, \mathscr{F}^T, P)$, let $(\overline{\Omega}, \overline{\mathscr{F}}) = (\Omega^T, \mathscr{F}^T)$ denote its underlying measurable space. For *any* measurable space $(\overline{\Omega}, \overline{\mathscr{F}})$ (not necessarily with the above structure), a sequence of sub-σ-algebras $\{\overline{\mathscr{F}}_t : t \in \mathbb{N}\}$ is called a *filtration* of the space if

$$\overline{\mathscr{F}}_0 \subseteq \overline{\mathscr{F}}_1 \subseteq \cdots \subseteq \overline{\mathscr{F}}. \tag{7}$$

Our stochastic process is said to be *adapted* to a filtration $\{\overline{\mathscr{F}}_t : t \in \mathbb{N}\}$ of its measurable space if for each t, X_t is $\overline{\mathscr{F}}_t$-measurable and not merely $\overline{\mathscr{F}}$-measurable. This means that the "prefix distribution" $P(X_1, \ldots, X_t)$ does not depend on the entire measure function P, but is determined by the restriction of P to just the sets in $\overline{\mathscr{F}}_t$. This is because $\overline{\mathscr{F}}_t \subseteq \overline{\mathscr{F}}$ includes enough of the events of the full measurable space $\overline{\mathscr{F}}$. In particular, in our discrete-variable setting where $\mathscr{F}$ is the discrete σ-algebra, $\overline{\mathscr{F}}_t$ contains all the events of the form $X_t = x_t$ and thus (by Eq. (7)) all events of the form $X_1 = x_1 \wedge \cdots \wedge X_t = x_t$.

Intuitively, a filtration describes an evolution of representational power. Each $\overline{\mathscr{F}}_t$ is fine-grained enough to characterize the information conveyed by $X_1, \ldots, X_t$, whereas $\overline{\mathscr{F}}_{t-1} \subseteq \overline{\mathscr{F}}_t$ may not be able to do so. Then an adapted process $\{X_t\}$ is one where each prefix distribution $P(X_1, \ldots, X_t)$ can be defined over the simplified measurable space $(\overline{\Omega}, \overline{\mathscr{F}}_t)$: the events of the prefix distribution remain measurable there.

We are now ready to give the formal definition of a stopping time.

Definition 3.1. In a filtered probability space $(\overline{\Omega}, \overline{\mathscr{F}}, P, \{\overline{\mathscr{F}}_t\})$, an $\mathbb{N} \cup \{\infty\}$-valued random variable τ is called a *stopping time* if $\{\tau = t\} \in \overline{\mathscr{F}}_t$ for all $t \in \mathbb{N}$.

This means that $P(\tau = t)$ can be defined over the simplified measurable space $(\overline{\Omega}, \overline{\mathscr{F}}_t)$. So can $P(\tau = t')$ for any $t' < t$, thanks to Eq. (7).

In the above setting where the filtered probability space is a discrete stochastic process that is adapted to its filtration, it follows that for $t' \leq t$, the joint probabilities $P(\tau = t', X_1 = x_1, \ldots, X_t = x_t)$ and thus the conditional probabilities $P(\tau = t' \mid X_1 = x_1, \ldots, X_t = x_t)$ can also be defined using the simplified measurable space.

In the case of language processes, let us define

$$\tau_{\text{EOS}} \stackrel{\text{def}}{=} \inf\{t : X_t = \text{EOS}\}, \tag{8}$$

which is the first time at which EOS appears in the sequence. Observe that $\{\tau_{\text{EOS}} = t\}$ is the event of getting any string of length $t-1$ and $\{\tau_{\text{EOS}} = \infty\}$ is the event of getting any infinite string. Let us construct a filtration such that τ_{EOS} is a stopping time.

We use the natural filtration (cf. [M4]) of $(\overline{\Sigma}^{\mathbb{Z}^+}, \mathscr{F}^{\mathbb{Z}^+})$, defined by

$$\overline{\mathscr{F}}_t \stackrel{\text{def}}{=} \left\{ \left\{ \boldsymbol{x}\boldsymbol{\omega} : \boldsymbol{\omega} \in \overline{\Sigma}^{\mathbb{Z}^+} \right\} : \boldsymbol{x} \in \overline{\Sigma}^t \right\}. \tag{9}$$

It is straightforward to verify that $\{X_t\}$ is adapted to $\{\overline{\mathscr{F}}_t\}$ and that τ_{EOS} is indeed a stopping time.[7] Notice also that for $t' \leq t$, $P(\tau_{\text{EOS}} = t' \mid X_1 = x_1, \ldots, X_t = x_t)$ is always 0 or 1: this is not guaranteed for stopping times in general, but holds here because $\tau_{\text{EOS}} = t'$ is fully determined by $X_1, \ldots X_t$ (Eq. (8)).

The stopping time characterization of EOS gives rise to an understanding of the language process as a string-valued random variable, i.e., the formal language view of language model we introduced in §1. Define the *stopped language process* by $Y_t = X_{t \wedge \tau_{\text{EOS}}}$ (recall that τ_{EOS} is itself a random variable).[8] In other words, $\{Y_t\}$ agrees with $\{X_t\}$ up

[6] They also happen to be practically important in designing Monte Carlo samplers for language models (Lew et al., 2023).

[7] We provide a proof of this fact in Proposition D.1 in App. D.

[8] It is common to use lattice-theoretic notation in stochastic processes, i.e., $x \wedge y = \min\{x, y\}$ and $x \vee y = \max\{x, y\}$.

through the first EOS and continues with EOS thereafter. The set of sequences $\{y_t\}$ in which EOS is followed by any non-EOS symbol is assigned probability 0. Hence, there exists a bijection between the support of $\{Y_t\}$ and $\Sigma^* \cup \Sigma^{\mathbb{Z}^+}$ where $\Sigma^{\mathbb{Z}^+}$ is the set of infinite strings.[9] If the language process is tight, i.e., $P(\overline{\Sigma}^{\mathbb{Z}^+}) = 0$ as well, then the stopped process $\{Y_t\}$ can be regarded as a string-valued random variable (one that almost surely takes values in Σ^*).

Finally, we can use τ_{EOS} to give a straightforward characterization of tightness. Recall that a tight language process is where EOS termination occurs with probability 1. This corresponds to the following definition.

Definition 3.2. A language process is said to be *tight* if $P(\tau_{\text{EOS}} < \infty) = 1$.

A tight language process specifies a distribution $\{Y_t\}$ over Σ^*, that is, a language model.

3.2 Tightness Results

We now derive concrete conditions for when a language process is tight. The construct of filtration allows us to apply results from martingale theory, which yields generalizations of results from previous work (Du et al., 2023). Specifically, we recall the Lévy–Borel–Cantelli theorem from martingale theory.

Theorem 3.3 (Lévy–Borel–Cantelli Theorem)**.**[10] *Let $(\Omega, \mathscr{F}, P, \{\mathscr{G}_t\}_{t\geq 0})$ be a filtered probability space with $\mathscr{G}_0 = \{\varnothing, \Omega\}$ and $\{A_t\}_{t\geq 1}$ be a sequence of events with $A_t \in \mathscr{G}_t$. Then*

$$\{A_t \; i.o.\} \overset{a.s.}{=} \left\{ \sum_{t=1}^{\infty} P(A_t \mid \mathscr{G}_{t-1}) = \infty \right\}. \quad (10)$$

where { } indicates an event, i.o. is "infinitely often," and a.s. is "almost surely" (see [M6]). That is, if an outcome $\boldsymbol{\omega}$ satisfies infinitely many of the properties A_t, it is almost surely an outcome in which the A_t have infinite total conditional probability, and vice-versa. The definition of "conditional probability" here is subtle: $P(A_t \mid \mathscr{G}_{t-1})$ denotes a Kolmogorov conditional (see [M5]). In the case where $\mathscr{G}_{t-1}$ is a finitely generated σ-algebra, we can say that this conditional probability of A_t is conditioned on the *most specific* event of $\mathscr{G}_{t-1}$ in which $\boldsymbol{\omega}$ falls (which is well-defined in this case). Hence it is conditioned on *at least* whether $\boldsymbol{\omega}$ satisfies $A_1, \ldots, A_{t-1}$, since $A_1, \ldots, A_{t-1} \in \mathscr{G}_{t-1}$.

We remark that both the first and the second Borel–Cantelli lemmas from elementary measure theory, which were used in Du et al. (2023), can be straightforwardly derived from Theorem 3.3 (see, e.g., Williams, 1991, §12.5).

To relate Theorem 3.3 to tightness, let $(\overline{\Omega}, \overline{\mathscr{F}}, P)$ be our language process, with $\{\mathscr{G}_t\}$ being some filtration of its measurable space. (The language process is not necessarily adapted to the filtration.) We choose

$$A_t \overset{\text{def}}{=} \{\tau_{\text{EOS}} \leq t\}. \quad (11)$$

A little thought shows that $\{A_t \; i.o.\} = \{\tau_{\text{EOS}} < \infty\}$, since $\{A_t \; i.o.\}$ means that there are infinitely many times t beyond the first EOS, and $\{\tau_{\text{EOS}} < \infty\}$ means that there is a first EOS; both are true just when $\boldsymbol{\omega}$ corresponds to a finite string. (See Proposition D.2 in the Appendix.) Thus, the sequence of events $\{A_t\}_{t\geq 1}$ connects the tightness property with Theorem 3.3 in the following way.

Theorem 3.4. *Let $\{\mathscr{G}_t\}_{t\geq 0}$ be any filtration over a language process with $\mathscr{G}_0 = \{\varnothing, \Omega\}$ and $A_t \in \mathscr{G}_t$ for all $t \geq 1$. Then,*

$$\{\tau_{\text{EOS}} < \infty\} \overset{a.s.}{=} \left\{ \sum_{t=1}^{\infty} P(A_t \mid \mathscr{G}_{t-1}) = \infty \right\}. \quad (12)$$

Due to its abstract form, Theorem 3.4 is very general. It yields different characterizations of tightness depending on the chosen filtration $\{\mathscr{G}_t\}$, which need only satisfy $\mathscr{G}_0 = \{\varnothing, \Omega\}$ and $A_t \in \mathscr{G}_t$ for each $t \geq 1$. For example, applying Theorem 3.4 to the natural filtration defined in Eq. (9), we obtain the following characterization of tightness:

$$\text{tight} \;\Leftrightarrow\; \sum_{t=1}^{\infty} P(A_t \mid \mathscr{F}_{t-1}) \overset{\text{a.s.}}{=} \infty. \quad (13)$$

In other words, a language process is tight if it has probability 1 of drawing a sequence $\boldsymbol{\omega}$ for which $\sum_t P(\tau_{\text{EOS}} \leq t \mid X_1, \ldots, X_{t-1})$ is infinite.[11] Using a different filtration leads to a more practical condition to determine tightness, as we state below.

[9] The mapping that maps from $\Sigma^* \cup \Sigma^{\mathbb{Z}^+}$ to the support of $\{Y_t\}$ by adding an infinite sequence of EOS to elements of Σ^* is this bijection.

[10] See, e.g., Kallenberg, 2021, Corollary 9.21 or Durrett, 2019, Theorem 4.3.4.

[11] Which is to say, $\boldsymbol{\omega}$ either contains EOS, or avoids generating EOS at every step despite EOS often having a high probability given the previous symbols. The latter event has probability 0, so this is equivalent to saying that the language process has probability 1 of drawing a sequence that contains EOS—i.e., it is tight.

Corollary 3.5 (Theorem 4.7 in Du et al., 2023)**.** *A language process is tight if and only if $s_t = 1$ for some t or $\sum_t s_t = \infty$, where s_t is defined as*

$$s_t \stackrel{\text{def}}{=} P(\tau_{\text{EOS}} \leq t \mid \tau_{\text{EOS}} > t-1) \tag{14}$$

$$= \frac{\sum_{\boldsymbol{\omega} \in \Sigma^{t-1}} p(\text{EOS} \mid \boldsymbol{x}) p(\boldsymbol{x})}{\sum_{\boldsymbol{\omega} \in \Sigma^{t-1}} p(\boldsymbol{x})}. \tag{15}$$

That is, s_t is the probability that a prefix of length $t-1$ that does not contain EOS *will be immediately followed by* EOS.

Proof. Apply Theorem 3.4 to the filtration $\mathscr{G}_t = \sigma(\{A_1, \ldots, A_t\})$ (defined in [M2]) and then compute $P(A_t \mid \mathscr{G}_{t-1})$. See App. E for details. □

Note that the quantity s_t in Cor. 3.5 involves a partition function in its denominator, which may be intractable to compute (Lin et al., 2021). We therefore derive the following condition which is easier to verify in practice.

Corollary 3.6 (Proposition 4.3 in Du et al., 2023)**.** *If $p(\text{EOS} \mid \boldsymbol{x}) \geq f(t)$ for all $t \geq 1$, $\boldsymbol{x} \in \Sigma^{t-1}$, and $\sum_{t=1}^{\infty} f(t) = \infty$, then the language process induced by p is tight.*

Proof. A direct consequence of Cor. 3.5. See App. E. □

In particular, the lower bound function $f(t)$ in Cor. 3.6 can be established for specific architectures such as Transformers or RNNs. We refer to Du et al. (2023, Section 5) for these results.

We conclude with some simple examples of applying these tightness conditions. If a language model satisfies $p(\text{EOS} \mid \boldsymbol{x}) \geq \varepsilon/(t+1)$ for $\boldsymbol{x} \in \Sigma^t$ and $\varepsilon > 0$, the corresponding language process is tight by Cor. 3.6 since the series $\sum_{t=0}^{\infty} \frac{1}{t+1}$ diverges. On the other hand, if $p(\text{EOS} \mid \boldsymbol{x}) \leq 2^{-(t+1)}$ for $\boldsymbol{x} \in \Sigma^t$, the corresponding language process is non-tight by Cor. 3.5 since $\sum_{t=0}^{\infty} 2^{-(t+1)} < \infty$.

4 Conclusions

This paper presents a formal treatment of language model and its tightness using the theory of stochastic processes. We discuss the formalization of language model as a stochastic process by applying a discrete version of the Kolmogorov Extension theorem. We then give a more intuitive formal understanding of EOS by characterizing it as a stopping time, a key construct in stochastic processes. Finally, the stopping formulation allows us to connect tightness to martingale theory and obtain more general conditions on tightness.

Limitations

Theorem 3.4 is a new general result about tightness, but the only applications we gave of this theorem (namely Corollaries 3.5 and 3.6) were to derive characterizations of tightness that were previously known. In this article, we only discussed the language process as derived from an autoregressive language model. However, there are alternative models of production of language, such as PCFG. It is possible to formally derive a stochastic process for PCFG based on variants of branching processes and, indeed, several previous works have made use of such formalisms to varying extents (Booth and Thompson, 1973; Miller and O'Sullivan, 1992; Chi, 1999). However, our construction using the Kolmogorov Extension theorem does not directly extend to these cases.

We also refrained from discussing specific parameterizations of language models, such as finite-state, recurrent, or Transformer-based language models. Interested readers can refer to Du et al. (2023, §5) which contains a comprehensive analysis of various concrete parametrizations of language models.

Ethics Statement

The paper provides formalisms and theoretical analysis of language models. We do not foresee ethical implications of this paper.

Acknowledgements

We thank members of Rycolab and anonymous reviewers for giving valuable feedback to our draft.

References

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

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

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.

Kai Lai Chung, editor. 1974. *A Course in Probability Theory*, second edition. Academic Press, San Diego.

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.

Łukasz Dębowski. 2020. *Information Theory Meets Power Laws: Stochastic Processes and Language Models*. John Wiley & Sons, Ltd.

Joseph L. Doob. 1953. *Stochastic processes*. John Wiley & Sons, New York.

Li Du, Lucas Torroba Hennigen, Tiago Pimentel, Clara Meister, Jason Eisner, and Ryan Cotterell. 2023. A measure-theoretic characterization of tight language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9744–9770, Toronto, Canada. Association for Computational Linguistics.

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

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

Olav Kallenberg. 2021. *Foundations of Modern Probability*, 3rd edition. Springer International Publishing, Cham.

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

Dexter Kozen. 2016. Kolmogorov extension, martingale convergence, and compositionality of processes. In *Proceedings of the 31st Annual ACM/IEEE Symposium on Logic in Computer Science*, LICS '16, page 692–699, New York, NY, USA. Association for Computing Machinery.

Alexander K. Lew, Tan Zhi-Xuan, Gabriel Grand, and Vikash Mansinghka. 2023. Sequential Monte Carlo steering of large language models using probabilistic programs. In *ICML 2023 Workshop: Sampling and Optimization in Discrete Space*.

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.

Alexander A Markov. 1913. An example of statistical investigation of the text Eugene Onegin concerning the connection of samples in chains. *Royal Academy of Sciences, St. Petersburg*. Lecture at the physical-mathematical faculty.

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

M.I. Miller and J.A. O'Sullivan. 1992. Entropies and combinatorics of random branching processes and context-free languages. *IEEE Transactions on Information Theory*, 38(4):1292–1310.

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

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.

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

Michael Sipser. 2013. *Introduction to the Theory of Computation*, 3 edition. Cengage Learning.

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

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.

David Williams. 1991. *Probability with Martingales*. Cambridge University Press.

A Related Work

Formal constructions of language models in axiomatic probability theory (e.g., measure-theoretic probability à la Kolmogorov, 1933) have not been studied until recently (Dębowski, 2020, Ch. 2; Du et al., 2023). Du et al. (2023) directly constructs and extends a pre-measure from strings. Closest to our treatment is Dębowski (2020, §2.2), which states the existence of the process by citing the Kolmogorov Extension theorem. However, they cite a proof of the $\mathbb{R}$-valued version of the Kolmogorov Extension theorem and omit how to apply the theorem (in §2.3, below Definition 2.11).[12]

We also note that, there are statements of more general versions of the Kolmogorov extension theorem in terms of projective limits, e.g., Kallenberg (2021, Theorem 8.23) or Tao (2011, Thereom 2.4.3). The projective limit version of the theorem is stated for Borel spaces, and hence encapsulates both the continuous case and the discrete case. On a more abstract level, Kozen (2016) points out the category-theoretic connection between the Kolmogorov extension theorem and Doob's martingale convergence theorem. However, we deem these versions of the Kolmogorov extension theorem to be unnecessarily abstract for our purposes and hence state and prove the more intuitive version with the consistency conditions in our §2.

Tightness in language models has been studied in various contexts, most recently in the context of neural autoregressive language models (Chen et al., 2018; Welleck et al., 2020; Lin et al., 2021; Meister et al., 2022; Du et al., 2023). In particular, Welleck et al. (2020) prove the simple fact that a nonzero lower bound of EOS probability implies tightness. Later, Du et al. (2023) prove the more general sufficient and necessary condition. In this article, we generalize this result using martingale theory to almost sure equalities of events.

Tightness in PCFGs has also been extensively studied (Booth and Thompson, 1973; Chi and Geman, 1998; Cohen and Johnson, 2013; Chi, 1999; Nederhof and Satta, 2006). It is here that the techniques from stochastic processes prove to be extensively useful. In Booth and Thompson (1973); Chi (1999); Chi and Geman (1998), the theory of branching processes is used to obtain conditions of tightness in PCFGs. We note that martingale theory is also useful in obtaining the main results in branching processes (Durrett, 2019, §4.3).

B Measure Theory Background

We put the definitions and background from measure-theoretic probability in this section and refer to it when we encounter needed definitions in the main text.

[M1] *Probability Space.* A *probability space* is a triple $(\Omega, \mathscr{F}, P)$ where Ω is the sample space, $\mathscr{F}$ is a σ-algebra over Ω and P is a probability measure over $(\Omega, \mathscr{F}, P)$.

[M2] *Generated σ-algebra.* Let $\mathscr{C}$ be a collection of subsets in Ω. The *σ-algebra generated by $\mathscr{C}$*, denoted by $\sigma(\mathscr{C})$, is the smallest σ-algebra over Ω containing $\mathscr{C}$. We also often consider the generated σ-algebra over a collection of random variables $\{X_t\}_{t\in T}$ in a probability space $(\Omega, \mathscr{F}, P)$. In this case, $\sigma(\{X_t\}_{t\in T})$ is taken to be the smallest σ-algebra in which all X_t are measurable with respect to it.

[M3] *Product σ-algebra.* Let T be an arbitrary index set and let $(\Omega, \mathscr{F})$ be a measurable space. The product σ-algebra $\mathscr{F}^T$ is a σ-algebra over Ω^T and is the defined to be the σ-algebra generated by the one-dimensional cylinder sets $\mathscr{F}^T = \sigma(\{\{\boldsymbol{\omega} \in \Omega^T : \boldsymbol{\omega}_t \in H\} : t \in T, H \in \mathscr{F}\})$ (cf. [M2]). In our discussion, Ω is taken to be a finite set and $\mathscr{F}$ is taken to be the power set of Ω. A caveat here is that, $\mathscr{F}^T$ is generally not equal to the power set of Ω^T. To see why, one notes that the two sets have different cardinalities (Folland, 1999, Proposition 1.23).

[M4] *Natural/Canonical Filtration.* In a stochastic process $X_1, X_2, \ldots$ in $(\Omega, \mathscr{F}, P)$, the *natural filtration* or *canonical filtration* $\mathscr{F}_t$ is defined as $\mathscr{F}_t = \sigma(\{X_1, \ldots, X_t\})$ (cf. [M2]).

[M5] *Kolmogorov Conditionals.* Let $\mathscr{G}$ be a sub-σ-algebra in a probability space $(\Omega, \mathscr{F}, P)$. The Kolmogorov conditional expectation of a random variable X with $E|X| < \infty$ with respect to $\mathscr{G}$ is denoted as $E(X \mid \mathscr{G})$, which is itself a random variable that is $\mathscr{G}$ measurable and satisfies

[12] It is possible to use the $\mathbb{R}$-valued Kolmogorov Extension theorem for our purposes by embedding the discrete points over $\mathbb{R}$, resulting in atomic measures. These details are, however, missing from Dębowski (2020, §2.2).

$\int_A E(X \mid \mathscr{G})dP = \int_A X dP$ for all $A \in \mathscr{G}$. The Kolmogorov conditional probability $P(A \mid \mathscr{G})$ is defined to be $E(\mathbf{1}_A \mid \mathscr{G})$.

[M6] *Almost Sure Equality.* In a probability space $(\Omega, \mathscr{F}, P)$, two random variables X and Y are almost surely equal, denoted by $X \overset{\text{a.s.}}{=} Y$, if $P(X = Y) \overset{\text{def}}{=} P(\{X = Y\}) = 1$. Two events E and F are almost surely equal, denoted by $E \overset{\text{a.s.}}{=} F$, if $P(\mathbf{1}_E = \mathbf{1}_F) = 1$.

C Kolmogorov Extension Theorem Proof

Theorem 2.1 (Kolmogorov Extension Theorem)**.** *Let T be an arbitrary index set, and $(\Omega, \mathscr{F})$ be a finite measurable space where $\mathscr{F}$ is the discrete σ-algebra.*[13] *Define $\mathscr{F}^k$ and $\mathscr{F}^T$ as in [M3]. Given a system of measures $\{\mu_*\}$ where for each k-tuple $(t_1, \ldots, t_k)$ of distinct elements in T, $\mu_{t_1,\ldots,t_k}$ is a measure over $(\Omega^k, \mathscr{F}^k)$, and where for every choice of $H_1, \ldots H_k \in \mathscr{F}$, these measures satisfy*

(1) For all permutations π,

$$\begin{aligned} &\mu_{t_1,\ldots,t_k}(H_1 \times \cdots \times H_k) = \\ &\quad \mu_{t_{\pi(1)},\ldots,t_{\pi(k)}}(H_{\pi(1)} \times \cdots \times H_{\pi(k)}) \end{aligned} \tag{1}$$

(2) For arbitrary distinct k-tuples $(t_1, \ldots, t_k)$,

$$\begin{aligned} &\mu_{t_1,\ldots,t_{k-1}}(H_1 \times \cdots \times H_{k-1}) = \\ &\quad \mu_{t_1,\ldots,t_{k-1},t_k}(H_1 \times \cdots \times H_{k-1} \times \Omega). \end{aligned} \tag{2}$$

Then, there is a unique probability measure P on $(\Omega^T, \mathscr{F}^T)$ such that the coordinate random variables $\{X_t : t \in T\}$ have $\mu_{t_1,\ldots,t_k}$ as their finite-dimensional distributions, i.e.,

$$P\big((X_{t_1}, \ldots, X_{t_k}) \in A\big) = \mu_{t_1,\ldots,t_k}(A) \tag{3}$$

for all $A \in \mathscr{F}^k$.

Proof. We adapt the strategy of the $\mathbb{R}$-valued stochastic process proof in Billingsley (1995, Sec. 36): We first define a pre-measure-like set function, which requires a proof that such a function is well-defined; We then prove that this function is indeed a pre-measure, allowing us to apply the Carathéodory extension theorem.

Definition of Pre-Measure. We first note that sets of the form

$$\{x \in \Omega^T : (x_{t_1}, \ldots, x_{t_k}) \in H\} \tag{16}$$

for some $H \in \mathscr{F}^k$ form an algebra (but not a σ-algebra). We call this algebra $\mathscr{F}_0^T$. Here and afterwards, we will abbreviate these kind of sets with the notation $\{(x_{t_1}, \ldots, x_{t_k}) \in H\}$ since the underlying space is always $\mathbb{R}^T$ so there is no confusion of this notation. Recall that our goal is to try to define a probability pre-measure P_0 over $(\Omega^T, \mathscr{F}_0^T)$ and then apply the Carathéodory theorem. We define

$$P_0(\{(x_{t_1}, \ldots, x_{t_k}) \in H\}) = \mu_{t_1,\ldots,t_k}(H). \tag{17}$$

Consistency of Pre-Measure. An immediate issue is whether the definition of P_0 is consistent, i.e., whether it is well-defined. In other words, we need to show the following

Proposition C.1. *If*

$$\{(x_{t_1}, \ldots, x_{t_k}) \in H\} = \{(x_{s_1}, \ldots, x_{s_\ell}) \in I\}, \tag{18}$$

then

$$\mu_{t_1,\ldots,t_k}(H) = \mu_{s_1,\ldots,s_\ell}(I). \tag{19}$$

[13] That is, $\mathscr{F}$ is the power set of Ω.

Proposition C.1 shows that P_0 is a consistent function that maps the same input to the same output, and hence is consistent.

Proof of Proposition C.1. Let $(t_1, \ldots, t_k), H$ and $(s_1, \ldots, s_\ell), I$ be such that $\{(x_{t_1}, \ldots, x_{t_k}) \in H\} = \{(x_{s_1}, \ldots, x_{s_\ell}) \in I\}$. Then, we can find indices $(r_1, \ldots, r_m)$ such that $\{r_1, \ldots, r_m\} = \{t_1, \ldots, t_k\} \cup \{s_1, \ldots, s_\ell\}$ and $(r_1, \ldots, r_k) = (t_1, \ldots, t_k)$. Moreoever, we can find a permutation $\pi \in \mathrm{Sym}(m)$ such that

$$(r_{\pi(1)}, \ldots, r_{\pi(\ell)}) = (s_1, \ldots, s_\ell). \tag{20}$$

We begin by rewriting the LHS and RHS of Eq. (18). First, consider the more straightforward direction of LHS

$$\begin{aligned}
&\{(x_{t_1}, \ldots, x_{t_k}) \in H\} && (21)\\
&= \{(x_{r_1}, \ldots, x_{r_k}) \in H\} && (22)\\
&= \{(x_{r_1}, \ldots, x_{r_k}, x_{r_{k+1}}, \ldots, x_{r_m}) \in H \times \Omega^{m-k}\} && (23)\\
&= \{(x_{r_1}, \ldots, x_{r_m}) \in H \times \Omega^{m-k}\} && (24)
\end{aligned}$$

Next, we rewrite RHS,

$$\begin{aligned}
&\{(x_{s_1}, \ldots, x_{s_\ell}) \in I\} && (25)\\
&= \{(x_{r_{\pi(1)}}, \ldots, x_{r_{\pi(\ell)}}) \in I\} && (26)\\
&= \{(x_{r_{\pi(1)}}, \ldots, x_{r_{\pi(\ell)}}, x_{r_{\pi(\ell+1)}}, \ldots, x_{r_{\pi(m)}}) \in I \times \Omega^{m-\ell}\} && (27)\\
&= \{(x_{r_{\pi(1)}}, \ldots, x_{r_{\pi(m)}}) \in I \times \Omega^{m-\ell}\} && (28)\\
&= \{(x_{r_1}, \ldots, x_{r_m}) \in \pi^{-1}(I \times \Omega^{m-\ell})\} && (29)
\end{aligned}$$

The two sets are equal from Eq. (18), so their equivalent forms Eq. (24) and Eq. (29) are also equal (since they are the same set), i.e.

$$\{(x_{r_1}, \ldots, x_{r_m}) \in H \times \Omega^{m-k}\} = \{(x_{r_1}, \ldots, x_{r_m}) \in \pi^{-1}(I \times \Omega^{m-\ell})\} \tag{30}$$

For this to be true, it must be the case that

$$H \times \Omega^{m-k} = \pi^{-1}(I \times \Omega^{m-\ell}) \in \mathscr{F}^m. \tag{31}$$

We now turn to showing our actual goal, which is $\mu_{t_1,\ldots,t_k}(H) = \mu_{s_1,\ldots,s_\ell}(I)$. We start from the RHS

$$\begin{aligned}
&\mu_{s_1,\ldots,s_\ell}(I) && && (32)\\
&= \mu_{r_{\pi(1)},\ldots,r_{\pi(\ell)}}(I) && \text{(by Eq. (20))} && (33)\\
&= \mu_{r_{\pi(1)},\ldots,r_{\pi(\ell)},r_{\pi(\ell+1)},\ldots,r_{\pi(m)}}(I \times \Omega^{m-\ell}) && \text{(consistency)} && (34)\\
&= \mu_{r_1,\ldots,r_m}(\pi^{-1}(I \times \Omega^{m-\ell})) && \text{(by permutation invaraince)} && (35)\\
&= \mu_{r_1,\ldots,r_m}(H \times \Omega^{m-k}) && \text{(by Eq. (31))} && (36)\\
&= \mu_{t_1,\ldots,t_k,r_{k+1},\ldots,r_m}(H \times \Omega^{m-k}) && \text{(by definition of } r_i \text{ and } t_i) && (37)\\
&= \mu_{t_1,\ldots,t_k}(H) && \text{(by consistency)} && (38)
\end{aligned}$$

Note that, in the assumption, permutation invariance is only assumed for product of cylinder sets, so we need to use Dynkin's π-λ theorem to show that this property extends to all measurable sets in $\mathscr{F}^k$. Combined, the above concludes the proof and shows that the definition of P_0 is consistent. □

Proof of Pre-Measure. Before the application of Carathéodory extension theorem, the final step is to show that P_0 is indeed a pre-measure. A pre-measure is a set function that satisfies countable additivity in an algebra (instead of a σ-algebra).[14] A common strategy to show some set function is a pre-measure is to show that its continuity at $\varnothing$ (Billingsley, 1995, Example 2.10), which we do so below.

Let $A_n \downarrow \varnothing$ where $A_n \in \mathscr{F}_0^T$, we need to show that $P_0(A_n) \downarrow 0$ to show P_0 is countably additive. Suppose to the contrary that $P_0(A_n) \not\downarrow 0$, then there exists some $\epsilon > 0$ such that $P_0(A_n) > \epsilon$ for all n. This means that $A_n \neq \varnothing$ for all n. We now set up to invoke Cantor's intersection theorem. Equip Ω with the discrete topology, then Ω^T is compact by the Tychonoff theorem (Munkres, 2000, Ch. 37). Then $\{A_n\}$ where $A_n \in \mathscr{F}_0^T$ form a nested sequence of nonempty compact sets. Each A_n is also closed since it's a finite intersection of closed sets ($\mathscr{F}_0^T$ contains only finite-index sets). Since closed subsets of compact space is compact, each A_n is also compact in the product discrete topology of Ω^T. By Cantor's intersection theorem, $\bigcap_n A_n \neq \varnothing$, contradicting our earlier assumption that $P_0(A_n) \not\downarrow 0$. Hence, $P_0(A_n) \downarrow 0$.

Extension. So far, we have shown that P_0 is a probability pre-measure over $(\Omega^T, \mathscr{F}_0^T)$, which generates $\mathscr{F}^T$. The claim then follows from Carathéodory's Extension theorem, which includes uniqueness (by Dynkin's π-λ theorem).

□

D Stopping Time Proofs

Proposition D.1. *The $\mathbb{N} \cup \{\infty\}$-valued random variable τ_{EOS} defined as*

$$\tau_{\text{EOS}} \stackrel{\text{def}}{=} \inf\{t : X_t = \text{EOS}\}, \tag{8}$$

is a stopping time.

Proof. To show τ_{EOS} is a stopping time is to show that $\{\tau_{\text{EOS}} = t\} \in \mathscr{F}_t$ for all $t \in \mathbb{N}$. By definition,

$$\begin{aligned}
\{\tau_{\text{EOS}} = t\} &= \big\{\inf\{t' : X_{t'} = \text{EOS}\} = t\big\} && (39)\\
&= \{X_1 \neq \text{EOS} \wedge \cdots \wedge X_{t-1} \neq \text{EOS} \wedge X_t = \text{EOS}\} && (40)\\
&= \{X_1 \neq \text{EOS}\} \cap \cdots \cap \{X_{t-1} \neq \text{EOS}\} \cap \{X_t = \text{EOS}\} && (41)\\
&= \underbrace{X_1^{-1}(\Sigma)}_{\in \mathscr{F}_1} \cap \cdots \cap \underbrace{X_{t-1}^{-1}(\Sigma)}_{\in \mathscr{F}_{t-1}} \cap \underbrace{X_t^{-1}(\{\text{EOS}\})}_{\in \mathscr{F}_t} && (\{X_t\} \text{ is adapted}) \quad (42)\\
&\in \mathscr{F}_t && (\mathscr{F}_1 \subseteq \mathscr{F}_2 \subseteq \cdots) \quad (43)
\end{aligned}$$

which concludes the proof. □

Proposition D.2. *Let $A_t = \{\tau_{\text{EOS}} \leq t\}$, then $\{A_t \; i.o.\} = \{\tau_{\text{EOS}} < \infty\}$.*

Proof.

$$\{A_t \; i.o.\} = \bigcap_{m=1}^{\infty} \bigcup_{t \geq m}^{\infty} A_t = \bigcap_{m=1}^{\infty} \bigcup_{t \geq m}^{\infty} \{\tau_{\text{EOS}} \leq t\} = \bigcap_{m=1}^{\infty} \{\tau_{\text{EOS}} < \infty\} = \{\tau_{\text{EOS}} < \infty\} \tag{44}$$

□

E Tightness Proofs

Corollary 3.5 (Theorem 4.7 in Du et al., 2023)**.** *A language process is tight if and only if $s_t = 1$ for some t or $\sum_t s_t = \infty$, where s_t is defined as*

$$\begin{aligned}
s_t &\stackrel{\text{def}}{=} P(\tau_{\text{EOS}} \leq t \mid \tau_{\text{EOS}} > t-1) && (14)\\
&= \frac{\sum_{\boldsymbol{\omega} \in \Sigma^{t-1}} p(\text{EOS} \mid \boldsymbol{x}) p(\boldsymbol{x})}{\sum_{\boldsymbol{\omega} \in \Sigma^{t-1}} p(\boldsymbol{x})}. && (15)
\end{aligned}$$

[14] See, e.g., (Billingsley, 1995, §2, Sec. Probability Measures), who refers to a pre-measure as "a probability measure on a field (algebra)".

That is, s_t is the probability that a prefix of length $t-1$ that does not contain EOS *will be immediately followed by* EOS.

Proof. As mentioned in §3.2, we define

$$\mathscr{G}_t = \sigma(\{A_1, \ldots, A_t\}). \tag{45}$$

Then, apply Theorem 3.4 to $\{\mathscr{G}_t\}$ implies that

$$\{\tau_{\text{EOS}} < \infty\} \overset{\text{a.s.}}{=} \left\{ \sum_{t=1}^{\infty} P(A_t \mid \mathscr{G}_{t-1}) = \infty \right\}. \tag{46}$$

To unpack this, we need to compute the quantity $P(A_t \mid \mathscr{G}_{t-1})$ next. Since $\mathscr{G}_t$'s are finitely generated, we can use the finite case of evaluating the Kolmogorov conditionals (Billingsley, 1995, Sec. 33). Notice that, $A_t = \{\tau_{\text{EOS}} \leq t\} \subseteq \{\tau_{\text{EOS}} \leq t+1\} = A_{t+1}$, so we have

$$A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots. \tag{47}$$

This means that, if ω is in any of $A_1, \ldots, A_{t-1}$, then $\omega \in A_t$ and hence $P(A_t \mid \mathscr{G}_{t-1})(\omega) = 1$. Hence, we only need to consider the case where ω isn't in any of $A_1, \ldots, A_{t-1}$, i.e., $\omega \in A_1^c \cap \cdots \cap A_{t-1}^c = A_{t-1}^c$. For $\omega \in A_{t-1}^c$, we have

$$P(A_t \mid \mathscr{G}_{t-1})(\omega) = P(A_t \mid A_{t-1}^c) = P(t \leq t \mid \tau_{\text{EOS}} > t-1) \tag{48}$$

$$= \frac{P(\tau_{\text{EOS}} \leq t \wedge \tau_{\text{EOS}} > t-1)}{P(\tau_{\text{EOS}} > t-1)} \left(= \frac{P(\tau_{\text{EOS}} = t)}{P(\tau_{\text{EOS}} > t-1)} \right) \tag{49}$$

$$= \frac{\sum_{\omega \in \Sigma^{t-1}} p(\text{EOS} \mid \omega) p(\omega)}{\sum_{\omega \in \Sigma^{t-1}} p(\omega)} \overset{\text{def}}{=} s_t. \tag{50}$$

To summarize,

$$P(A_t \mid \mathscr{G}_{t-1})(\omega) = \begin{cases} s_t & \omega \in A_{t-1}^c \\ 1 & \text{otherwise} \end{cases}. \tag{51}$$

We now use what we have calculated so far to prove the result.

(⇐). If $s_t = 1$, then $P(\tau_{\text{EOS}} \leq t) = 1$, hence $P(\tau_{\text{EOS}} < \infty) = 1$. If $\sum_t s_t = \infty$, then, for all ω,

$$\sum_{t=1}^{\infty} P(A_t \mid A_{t-1}^c)(\omega) \geq \sum_{t=1}^{\infty} s_t = \infty. \tag{52}$$

By Theorem 3.4, $\{\tau_{\text{EOS}} < \infty\} = \Omega$ and hence $P(\tau_{\text{EOS}} < \infty) = P(\Omega) = 1$.

(⇒). Assume $P(\tau_{\text{EOS}} < \infty) = 1$. If $s_t = 1$ for any t, then the result is true. So we assume $s_t < 1$ for all t. Assume to the contrary that $\sum_t s_t < \infty$, then there exists m such that $\sum_{t \geq m} s_t < 1$. Then,

$$P(\tau_{\text{EOS}} = \infty) = P(\tau_{\text{EOS}} > m \wedge \tau_{\text{EOS}} > m+1 \wedge \ldots) \tag{53}$$

$$= P(A_m^c \cap A_{m+1}^c \cap \cdots) \tag{54}$$

$$= P(A_{m+1}^c \mid A_m^c) P(A_{m+2}^c \mid A_{m+1}^c) \cdots \quad \text{(continuity of measure)} \tag{55}$$

$$\geq 1 - \sum_{t \geq m} P(A_{t+1} \mid A_t^c) \quad \text{(See footnote[15])} \tag{56}$$

$$= 1 - \sum_{t \geq m} s_t > 0 \tag{57}$$

which contradicts our hypothesis that $P(\tau_{\text{EOS}} < \infty) = 1$ since $P(\tau_{\text{EOS}} < \infty) = 1 - P(\tau_{\text{EOS}} = \infty)$. □

[15] This is an abstract application of the union bound (aka sub-additivity). Abstractly, this is saying the fact of $\prod_t p_t \geq 1 - \sum_t (1 - p_t)$. If we imagine a sequence of independent events E_t each with probability $P(E_t) = 1 - p_t$, then $\prod_t p_t = P(\bigcap_t E_t^c) = 1 - P(\bigcup_t E_t) \geq 1 - \sum_t P(E_t) = 1 - \sum_t (1 - p_t)$.

Corollary 3.6 (Proposition 4.3 in Du et al., 2023)**.** *If* $p(\text{EOS} \mid \boldsymbol{x}) \geq f(t)$ *for all* $t \geq 1$*,* $\boldsymbol{x} \in \Sigma^{t-1}$*, and* $\sum_{t=1}^{\infty} f(t) = \infty$*, then the language process induced by* p *is tight.*

Proof. If $p(\text{EOS} \mid t) \geq f(t)$, then $s_t \geq f(t)$ for all t. This means that $\sum_t s_t \geq \sum_t f(t) = \infty$. By Cor. 3.5, the language process is tight. □