Measures on Languages

1 Introduction

1.1 Notation

Let Σ denote a non-empty, countable alphabet. Unless otherwise specified, assume $|\Sigma| < \infty$. Write ε to mean the empty string.

Let $\Sigma^\star$ be the set of all finite strings from Σ. Write strings as w or s, whichever happens to be more convenient.

Let L denote a language, implicitly over Σ. In other words, $L \subseteq \Sigma^\star$.

We treat the empty language $\emptyset$ as distinct from the language with a single empty string $\{\varepsilon\}$.

Let $\mathcal{L}$ be a family of languages.

2 Measures

Given a family of languages $\mathcal{L}$, let $\sigma(\mathcal{L})$ be the σ-algebra generated on $\mathcal{L}$ satisfying the following:

(1)

$$\emptyset, \Sigma^\star \in \sigma(\mathcal{L})$$

(2)

$$L \in \sigma(\mathcal{L}) \implies L^c = \Sigma^\star \setminus L \in \sigma(\mathcal{L})$$

(3)

$$L_0, L_1, \ldots \in \sigma(\mathcal{L}) \implies \bigcup_{k=0}^{\infty} L_k \in \sigma(\mathcal{L})$$

Then $(\mathcal{L}, \sigma(\mathcal{L}))$ is a measurable space.

Remark 1. If $\mathcal{L}$ happened to be a family of regular languages, there is no guarantee that $\sigma(\mathcal{L})$ will still be a family of regular languages. A counter example is the following:

$$L_0 = \{\varepsilon\} \qquad L_1 = \{ab\} \qquad L_2 = \{aabb\} \qquad \dots \qquad L_k = \left\{a^k b^k\right\} \qquad \dots$$

But taking the countable union yields:

$$\bigcup_{k=0}^{\infty} L_k = \left\{a^k b^k \,:\, k \in \mathbb{Z}_{\geq 0}\right\}$$

Which is not regular. □

2.1 Defining Measures

2.1.1 From Non-negative Integers

We first consider the non-negative integers $\mathbb{Z}_{\geq 0}$. Let η be a σ-finite measure on $\mathbb{Z}_{\geq 0}$. The σ-finite conditions ensures that no strange singularities occur for any integers under consideration. We may later restrict η to be finite if necessary, if we want nicer conditions.

Observe that, by abuse of notation:

$$\Sigma^\star = \bigcup_{k=0}^{\infty} \Sigma^k$$

In English: $\Sigma^\star$ is the union of the set (language) of finite strings of length k, denoted Σ^k.

Because we assumed $|\Sigma| < \infty$, this also means that: $\left|\Sigma^k\right| = |\Sigma|^k$.

Consider now some language $L \in \sigma(\mathcal{L})$. Also decompose L into disjoint sub-languages by length as follows:

$$L = \bigcup_{k=0}^{\infty} L_k$$

Of course, $L_k \subseteq \Sigma^k$.

Because we are able to precisely calculate $\left|\Sigma^k\right|$, one "natural" way of defining a measure λ on the measurable space $(\mathcal{L}, \sigma(\mathcal{L}))$ is as follows:

$$\lambda(L) = \sum_{k=0}^{\infty} \lambda(L_k) = \sum_{k=0}^{\infty} \frac{|L_k|}{\left|\Sigma^k\right|} \eta(k)$$

We claim that $(\mathcal{L}, \sigma(\mathcal{L}), \lambda)$ forms a measure space.

Theorem 1. *λ is a measure.*

Proof. We check (1) measure under empty set is zero and (2) countable additivity, which will satisfy the requirements of a measure.

(1) Observe that $\lambda(\emptyset) = 0$ because the sum will be trivial.

(2) Let $L_0, L_1, L_2, \ldots$ be a countable collection of pairwise disjoint languages. We decompose each of these languages into a countably indexed set, where $L_{j,k}$ is the jth language's sub-language that only contains strings of length k. In other words:

$$L_j = \bigcup_{k=0}^{\infty} L_{j,k}$$

Observe that by (de)-construction, for any fixed j and for all $k_1 \neq k_2$, we have L_{j,k_1} and L_{j,k_2} are pairwise disjoint.

However, we have a stronger condition because each L_j is assumed to be pairwise disjoint. Thus, for all $j_1 \neq j_2$ and $k_1 \neq k_2$, L_{j_1,k_1} and L_{j_2,k_2} are disjoint. Then:

$$\begin{aligned}
\lambda\left(\bigcup_{j=0}^{\infty} L_j\right) &= \lambda\left(\bigcup_{j=0}^{\infty}\bigcup_{k=0}^{\infty} L_{j,k}\right) \\
&= \sum_{j=0}^{\infty} \lambda\left(\bigcup_{k=0}^{\infty} L_{j,k}\right) \\
&= \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} \frac{|L_{j,k}|}{|\Sigma^k|}\eta(k) \\
&= \sum_{j=0}^{\infty} \lambda(L_j)
\end{aligned}$$

This shows that λ is indeed a measure. □