Measures on Languages

1 Introduction

1.1 Notation

Let Σ denote a countable alphabet. Unless otherwise specified, assume $|\Sigma| < \infty$.

Let $\Sigma^\star$ be the set of all finite words from Σ. Write words as w.

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

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)

$$L_1, L_2, \ldots \in \sigma(\mathcal{L}) \implies \bigcup_{k=1}^{\infty} L_k \in \sigma(\mathcal{L})$$

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:

$$L_1 = \{ab\} \qquad L_2 = \{aabb\} \qquad L_3 = \{aaabbb\} \qquad \ldots \qquad L_k = \left\{a^k b^k\right\} \qquad \ldots$$

But taking the countable union yields:

$$\bigcup_{k=1}^{\infty} L_k = \{a^n b^n \,:\, n \in \mathbb{Z}_{\geq 0}\}$$

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:

$$\Sigma^{\star} = \bigcup_{k=1}^{\infty} \Sigma^k$$

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:

$$L = \bigcup_{k=1}^{\infty} L_k$$

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:

$$\lambda(L) = \sum_{k=1}^{\infty} \lambda(L_k) = \sum_{k=1}^{\infty} \frac{|L_k|}{|\Sigma^k|} \eta(k)$$