

Slicing-Based Measures on Formal Languages

1 Introduction

In this sketch we study measure spaces on formal languages and a few consequences.

2 Preliminaries

2.1 Measure Theory

Let X be a set. A σ -algebra $\Sigma \subseteq 2^X$ is a set that satisfies the following:

- (a) $\emptyset \in \Sigma$ and $X \in \Sigma$.
- (b) If $A \in \Sigma$, then $X \setminus A \in \Sigma$.
- (c) If $(A_n) \subseteq \Sigma$ is a countable collection of sets, then:

$$\bigcup_{n=1}^{\infty} A_n \in \Sigma$$

The pairing (X, Σ) is called a measureable space.

A measure is a function $\mu: \Sigma \rightarrow \mathbb{R}^{\geq 0}$ such that:

- (a) $\mu(\emptyset) = 0$
- (b) If $(A_n) \subseteq \Sigma$ is a countable collection of pairwise disjoint sets, then:

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n)$$

Together (X, Σ, μ) is called a measure space.

2.2 Formal Language Theory

An alphabet Σ is a finite set of symbols. A finite word w is a finite sequence of symbols from Σ , and let \emptyset denote the empty word, which is also finite. The length of a word is written $|w|$. Write Σ^n to mean the set of all words of length n , and write Σ^* to mean the set of all finite words. A language $L \subseteq \Sigma^*$ is a set of words.

2.3 Metric Spaces

A metric space (M, d) is a set M with a distance $d: M \times M \rightarrow \mathbb{R}^{\geq 0}$ such that for all $x, y, z \in M$:

- (a) $d(x, y) \geq 0$
- (b) $d(x, y) = 0$ if and only if $x = y$
- (c) $d(x, y) = d(y, x)$
- (d) $d(x, z) \leq d(x, y) + d(y, z)$

3 Slicing-Based Measures

Consider a language $L \subseteq \Sigma^*$. The n -splice of a language written as L^n is defined as:

$$L^n = L \cap \Sigma^n$$

We then have the following relations:

$$L = \bigcup_{n=0}^{\infty} L^n = \bigcup_{n=0}^{\infty} (L \cap \Sigma^n) = L \cap \bigcup_{n=0}^{\infty} \Sigma^n = L \cap \Sigma^* = L$$

Suppose that $(\mathbb{N}, 2^{\mathbb{N}}, \eta)$ is a probability measure space on \mathbb{N} with the probability measure η , one way to define a measure λ_η is as follows:

$$\lambda_\eta(L) = \sum_{n=0}^{\infty} \frac{|L^n|}{|\Sigma^n|} \eta(n)$$

Theorem 1. $(\Sigma^*, 2^{\Sigma^*}, \lambda_\eta)$ is a measure space.

Proof. Since the 2^{Σ^*} is the largest σ -algebra on Σ^* , it suffices to show that λ_η is a measure.

To see that \emptyset is mapped to 0:

$$\lambda_\eta(\emptyset) = \sum_{n=0}^{\infty} \frac{|\emptyset|}{|\Sigma^n|} \eta(n) = \sum_{n=0}^{\infty} 0 = 0$$

Now take $(A_n) \subseteq \Sigma^*$ to be a countable collection of disjoint sets. Write A_n^k to denote the k splice of the n th set. In other words:

$$A_n = \bigcup_{k=0}^{\infty} A_n^k$$

Observe that all such A_n^k are pairwise disjoint by construction, and so:

$$\lambda_{\eta} \left(\bigcup_{n=0}^{\infty} A_n \right) = \lambda_{\eta} \left(\bigcup_{n=0}^{\infty} \bigcup_{k=0}^{\infty} A_n^k \right) = \sum_{n=0}^{\infty} \lambda_{\eta} \left(\bigcup_{k=0}^{\infty} A_n^k \right) = \sum_{n=0}^{\infty} \sum_{k=0}^{\infty} \frac{|A_n^k|}{|\Sigma^n|} \eta(k) = \sum_{n=0}^{\infty} \lambda_{\eta}(n)$$

We conclude that $(\Sigma^*, 2^{\Sigma^*}, \lambda_{\eta})$ forms a measure space.

□

We can generalize this more. Suppose that $\nu = (\nu_n)$ is a countable collection of measures where each ν_n is defined on the splice Σ^n . Then we can extend a definition of $\lambda_{\eta, \nu}$ as:

$$\lambda_{\eta, \nu}(A) = \sum_{n=0}^{\infty} \nu(A^n) \eta(n)$$

Theorem 2. $(\Sigma^*, 2^{\Sigma^*}, \lambda_{\eta, \nu})$ is a measure space.

Proof. As with before, we only show that $\lambda_{\eta, \nu}$ is a measure.

For \emptyset we have again:

$$\lambda_{\eta, \nu}(\emptyset) = \sum_{n=0}^{\infty} 0 = 0$$

Again take $(A_n) \subseteq \Sigma^*$ to be a countable disjoint collection of sets, and A_n^k to be the k splice of A_n . Then:

$$\lambda_{\eta, \nu} \left(\bigcup_{n=0}^{\infty} A_n \right) = \lambda_{\eta, \nu} \left(\bigcup_{n=0}^{\infty} \bigcup_{k=0}^{\infty} A_n^k \right) = \sum_{n=0}^{\infty} \lambda_{\eta, \nu} \left(\bigcup_{k=0}^{\infty} A_n^k \right) = \sum_{n=0}^{\infty} \sum_{k=0}^{\infty} \nu_k(A_n^k) \eta(k) = \sum_{n=0}^{\infty} \lambda_{\eta, \nu}(A_n)$$

This shows that $(\Sigma^*, 2^{\Sigma^*}, \lambda_{\eta, \nu})$ is a measure space.

□

4 Measure Induced Metrics

For a measure space (X, Σ, μ) , an interesting consequence is that a metric space can be defined on Σ^* as follows:

$$d(A, B) = \mu(A \triangle B)$$

Where \triangle is the symmetric set difference. We now set out to show this.

Lemma 1. $(A \Delta C) \subseteq (A \Delta B) \cup (B \Delta C)$.

Proof. Observe that we may rewrite the above as follows:

$$(A \setminus C) \cup (C \setminus A) \subseteq [(A \setminus B) \cup (B \setminus C)] \cup [(B \setminus A) \cup (C \setminus B)]$$

It then suffices to show that:

$$A \setminus C \subseteq (A \setminus B) \cup (B \setminus C) \quad C \setminus A \subseteq (B \setminus A) \cup (C \setminus B)$$

We take turns examining these.

If $x \in A \setminus C$, then this implies that $x \in A$ and $x \notin C$. There are now two cases, where $x \in B$ or $x \notin B$. First assume that $x \in B$, which will imply that $x \in B \setminus C$. Now assume that $x \notin B$, which will imply that $x \in A \setminus B$. Either way, the implication is that $x \in (A \setminus B) \cup (B \setminus C)$, and so it follows that $A \setminus C \subseteq (A \setminus B) \cup (B \setminus C)$.

If $x \in C \setminus A$, then this implies that $x \in C$ and $x \notin A$. The argument is similar to the above, in which either $x \in B$ or $x \notin B$. If $x \in B$, then $x \in B \setminus A$, and otherwise if $x \notin B$ implies that $x \in C \setminus B$. Collectively, the two imply that $C \setminus A \subseteq (B \setminus A) \cup (C \setminus B)$.

Collectively, this shows that $(A \Delta C) \subseteq (A \Delta B) \cup (B \Delta C)$. □

Theorem 3. If (X, Σ, μ) is a measure space, then for $d: \Sigma \times \Sigma \rightarrow \mathbb{R}^{\geq 0}$ defined as:

$$d(A, B) = \mu(A \Delta B)$$

Is a metric function.

Proof. We prove the conditions necessary for a metric: identity, symmetry, and triangle inequality.

As μ is a measure, then for any $A \in \Sigma$:

$$d(A, A) = \mu(A \Delta A) = \mu(\emptyset) = 0$$

By the symmetry of symmetric set difference, for any $A, B \in \Sigma$:

$$d(A, B) = \mu(A \Delta B) = \mu(B \Delta A) = d(B, A)$$

For any $A, B, C \in \Sigma$, we have by convexity as shown in the lemma above:

$$A \Delta C \subseteq (A \Delta B) \cup (B \Delta C)$$

Then by sub-additivity of measures:

$$d(A, C) = \mu(A \Delta C) \leq \mu((A \Delta B) \cup (B \Delta C)) \leq \mu(A \Delta B) + \mu(B \Delta C) = d(A, B) + d(B, C)$$

□

Because $(\Sigma^*, 2^{\Sigma^*}, \lambda_{\eta, \nu})$ is a measure space, the consequence is that for any languages $L_1, L_2 \subseteq \Sigma^*$, we also have a metric space as defined above.