

# A mathematical theory of Language

Heinrich Hartmann

April 28, 2014

## Contents

<b>1 Preliminaries</b>	<b>1</b>
1.1 Spaces of Sequences . . . . .	1
1.2 Spaces of probability measures . . . . .	2
1.3 Probability measures on sequences . . . . .	3
1.3.1 Unigram measures . . . . .	3
1.3.2 $N$ -gram measures . . . . .	4
1.4 Markov measures . . . . .	5
1.4.1 The case of $\langle \text{undef} \rangle$ . . . . .	5
1.4.2 0-Markov measures . . . . .	6

## 1 Preliminaries

### 1.1 Spaces of Sequences

Let  $A$  be a finite set. In Examples, this will be the set of characters or the set of words. We define the set of Sequences over  $A$  as

$$\Sigma A = \{(a_1, \dots, a_l) \mid a_i \in A, l \geq 0\}.$$

We introduce the following notation:

- The empty sequence is denoted by  $\epsilon \in \Sigma A$ .
- The length function is denoted by

$$\text{length} : \Sigma A \longrightarrow \mathbb{N}_0, \quad (a_1, \dots, a_l) \mapsto l.$$

- We have the following canonical decomposition by sequence length

$$\Sigma A = A^0 \cup A^1 \cup A^2 \dots$$

and denote by  $i_N : A^N \rightarrow \Sigma A$  the inclusion of the length- $i$  sequences into  $\Sigma A$ .

Furthermore for each index  $i \geq 0$  we have a projection

$$\pi_i : \Sigma A \longrightarrow A_+, \quad (a_1, \dots, a_l) \mapsto \begin{cases} a_i & l \geq i \\ \langle \text{undef} \rangle & l < i \end{cases}.$$

Here  $A_+ := A \cup \{\langle \text{undef} \rangle\}$ . This space allows  $\pi_i$  to be defined on whole of  $\Sigma A$ .

Furthermore if  $N \geq 0$  and  $i \geq 1$  are given, we define the  $i$ -th  $N$ -gram projection to be:

$$\pi_i^N : \Sigma A \longrightarrow A_+^N, \quad (a_1, \dots, a_l) \mapsto \begin{cases} (a_{i+0}, \dots, a_{i+N-1}) & l \geq i + N - 1 \\ \langle \text{undef} \rangle & l < i + N - 1 \end{cases}.$$

Note that we get back  $\pi_i$  as  $\pi_i^1$ . Moreover  $\pi_i^0$  is the canonical projection of  $\Sigma A$  to the one point space  $A^0 \subset A_+^0$ .

We have the following two filtrations of  $\Sigma A$

$$A^0 = \Sigma^{\leq 0} A \subset \Sigma^{\leq 1} A \subset \Sigma^{\leq 2} A \subset \dots \subset \Sigma A$$

and

$$\Sigma A = \Sigma_{\geq 0} A \supset \Sigma_{\geq 1} A \supset \Sigma_{\geq 2} A \supset \dots \supset \bigcap \Sigma_{\geq i} A = \emptyset$$

where

$$\Sigma^{\leq i} A = A^0 \cup \dots \cup A^i$$

and

$$\Sigma_{\geq i} A = A^i \cup A^{i+1} \cup \dots$$

## 1.2 Spaces of probability measures

Let  $X$  be an at most countable<sup>1</sup> set. We denote by

$$\mathcal{M}(X) = \{\mu : X \longrightarrow \mathbb{R}_{\geq 0} \mid \sum_{x \in X} \mu(x) < \infty\}$$

---

<sup>1</sup>The assumption of countability could be dropped, at the expense of a slightly more technical treatment of infinite sums.

the space of all finite measures on  $X$ . For  $A \subset X$  we write  $\mu[A] = \sum_{x \in A} \mu(x)$ . This definition agrees with the usual definition, of measures in the case of discrete spaces with maximal  $\sigma$ -algebra.

The set of probability measures is defined as

$$\mathcal{P}(X) = \{\mu \mid \mu[X] = 1\} \subset \mathcal{M}(X).$$

We get a normalization map,

$$\mathcal{M}(X) \setminus \{0\} \longrightarrow \mathcal{P}(X), \quad \mu \mapsto \frac{1}{\mu[X]} \mu$$

which is defined for non-zero measures  $\mu$ .

Let  $f : X \longrightarrow Y$  be a map of sets, then we get a natural map

$$f_* : \mathcal{M}(X) \longrightarrow \mathcal{M}(Y), \quad f_*(\mu)(y) = \mu[f^{-1}(\{y\})] = \sum_{x:f(x)=y} \mu(x)$$

as well as  $f_* : \mathcal{P}(X) \rightarrow \mathcal{P}(Y)$ .

If  $f$  has finite fibers, then we also have the following map:

$$f^* : \mathcal{M}(Y) \longrightarrow \mathcal{M}(X), \quad \mu \mapsto \mu \circ f,$$

however the total volume of  $Y$  is not preserved, so that no map on  $\mathcal{P}$  is induced. In particular case, that  $\iota : A \subset X$  is the inclusion of a subspace we write  $\mu|_A$  for  $\iota^*(\mu)$ . If for  $P \in \mathcal{P}(X)$  the restriction  $P|_A$  is not necessary a probability measure. If  $P[A] \neq 0$ , then  $P|_A$  can be normalized to

$$P[\_ | A] = \frac{1}{P[A]} P|_A,$$

the conditional probability measure on  $A$ .

If  $\mu \in \mathcal{M}_f(X)$  and  $g : X \rightarrow \mathbb{R}$ , we define the *expectation* of  $g$  as:

$$E_\mu[f] := \sum_{x \in X} g(x) \mu(x).$$

This sum is well defined since  $\mu$  is finitely supported.

### 1.3 Probability measures on sequences

In this section we study the relationship between  $\mathcal{P}(A)$  and  $\mathcal{P}(\Sigma A)$ .

For  $i \geq 1$  and  $N \geq 0$  we have the following maps:

$$\pi_{i*} : \mathcal{P}(\Sigma A) \longrightarrow \mathcal{P}(A_+), \quad \pi_i^N : \mathcal{P}(\Sigma A) \longrightarrow \mathcal{P}(A_+^N),$$

as well as

$$length_* : \mathcal{P}(\Sigma A) \longrightarrow \mathcal{P}(\mathbb{N}_0).$$

### 1.3.1 Unigram measures

Hence, for each probability measure  $P$  on  $\Sigma A$ , we have  $\pi_{i*}P$  which is a measure on  $A_+$ . Note that

$$\pi_{i*}P(\langle \text{undef} \rangle) = P[\{s \mid \text{length}(s) < i\}] = P[\text{length} < i].$$

In the case, that  $\pi_{i*}P(\langle \text{undef} \rangle) \neq 1$  we can normalize  $\pi_{i*}P$  to the  $i$ -th element distribution

$$D_i P = \frac{1}{P[\text{length} \geq i]} \pi_{i*}P = \pi_{i*}P[\_ | \text{length} \geq i] \in \mathcal{P}(A)$$

on  $A$ .

To define a total distribution of all elements, we want to take the following infinite sum of  $i$ -th element distributions

$$\sum_{i \geq 1} \pi_{i*}P$$

However, this sum is not necessarily finite for a general measure  $P \in \mathcal{P}(\Sigma A)$ , therefore we make the additional assumption that  $P$  is finitely supported and define

$$M^1 : \mathcal{M}_f(\Sigma A) \longrightarrow \mathcal{M}(A), \quad \mu \mapsto \sum_{i \geq 1} \pi_{i*}\mu.$$

Note that the measure is only defined on  $A \subset A_+$ , since we have

$$\sum_{i \geq 1} \pi_{i*}(\langle \text{undef} \rangle) = \sum_{i \geq 1} P[\text{length} < i] = \infty.$$

We calculate the total volume to be

$$M^1 \mu(A) = \sum_{i \geq 1} \mu[\text{length} \geq i] = E_\mu[\text{length}].$$

Hence, for  $P \in \mathcal{P}_f(\Sigma A)$  with  $P[\text{length} = 0] \neq 1$ , we can normalize the measure  $M^1 P$ , and define *unigram distribution* on  $A$  as

$$D^1 P := \frac{1}{E_P[\text{length}]} \sum_{i \geq 1} \pi_{i*}P \in \mathcal{P}(A).$$

### 1.3.2 $N$ -gram measures

More generally, for integers  $i \geq 1$  and  $N \geq 0$  we get a measure  $\pi_i^N(P)$  on  $A_+^N$ , for which

$$\pi_i^N(P)(\langle \text{undef} \rangle) = P[\text{length} < i + N - 1].$$

If this number is not equal to one, we define the  $i$ -th  $N$ -gram distribution to be

$$D_i^N P = \frac{1}{P[\text{length} \geq i + N - 1]} \pi_i^N(P) \in \mathcal{P}(A^N)$$

For the global  $N$ -gram distributions we take

$$M^N : \mathcal{M}_f(\Sigma A) \longrightarrow \mathcal{M}(A^N), \quad \mu \mapsto \sum_{i \geq 1} \pi_i^N \mu.$$

Again, this measure is only defined on  $A^N \subset A_+^N$ . We calculate the total volume as

$$M^N(\mu)(A^N) = \sum_{i \geq 1} \mu[\text{length} \geq i + (N - 1)] \quad (1)$$

$$= \sum_{j=N} (j - (N - 1)) \mu[\text{length} = j] \quad (2)$$

$$= E_\mu[\max\{0, \text{length} - (N - 1)\}] \quad (3)$$

Note that this number depends only on the length distribution  $\text{length}_*(\mu) \in \mathcal{M}_f(\mathbb{N}_0)$  and is non-zero if and only if  $\mu[\text{length} \geq N] \neq 0$ .

Hence, for  $P \in \mathcal{P}_f(\Sigma A)$ , with  $P[\text{length} \geq N] > 0$  we can define the *total  $N$ -gram distribution* as

$$D^N P = \frac{1}{E[\max\{0, \text{length} - (N - 1)\}]} \sum_{i \geq 1} \pi_i^N P \in \mathcal{P}(A^N).$$

Note, that the special case of  $N = 1$  reduces to our earlier definition.

### 1.4 Markov measures

A probability measure  $P$  on  $\Sigma A$  is called  $K$ -Markov if for all  $l \geq K$ ,  $b_0, \dots, b_l \in A$  and  $n > l$  the condition

$$\begin{aligned} & P[\pi_n = b_0 \mid \pi_{n-1} = b_1, \dots, \pi_{n-K} = b_K, \dots, \pi_{n-l} = b_l] \\ &= P[\pi_n = b_0 \mid \pi_{n-1} = b_1, \dots, \pi_{n-K} = b_K] \end{aligned}$$

holds whenever both sides are well-defined, i.e.  $P[\pi_{n-1} = b_1, \dots, \pi_{n-l} = b_l]$  is non-zero.

#### 1.4.1 The case of $\langle \text{undef} \rangle$

The above definition, does not specify a condition in the case one or more of the  $b_i$  are  $\langle \text{undef} \rangle$ . For  $b_0 = \langle \text{undef} \rangle$  is unproblematic. In the case that  $b_0 \in A$  and  $b_j = \langle \text{undef} \rangle$  for one  $j > 0$ , the condition is empty since  $\pi_{i-j} = \langle \text{undef} \rangle$  implies  $\pi_i = \langle \text{undef} \rangle$ . For the remaining case of  $b_0 = \langle \text{undef} \rangle$  and  $b_j = \langle \text{undef} \rangle$  for one or more  $j > 0$ , the naive-condition does not extend. To see this, we take  $K = 0$  and  $l = 1$  with  $b_1 = \langle \text{undef} \rangle$ , so that the extended condition reads

$$P[\pi_n = \langle \text{undef} \rangle \mid \pi_{n-1} = \langle \text{undef} \rangle] = P[\pi_n = \langle \text{undef} \rangle]. \quad (4)$$

This implies  $P[\pi_{n-1} = \langle \text{undef} \rangle] = 1$ , which does not always hold.

#### 1.4.2 0-Markov measures

In the special case of  $K = 0$  we find

$$\begin{aligned} P[\pi_n = b_0] &= P[\pi_n = b_0 \mid \pi_{n-1} = b_1] \\ \Leftrightarrow P[\pi_n = b_0, \pi_{n-1} = b_1] &= P[\pi_n = b_0]P[\pi_{n-1} = b_1] \end{aligned}$$

which is the definition of  $P$ -independent between  $\pi_n$  and  $\pi_{n-1}$  random variables, except that the case  $\langle \text{undef} \rangle$  is excluded. We can account for that by using conditional probabilities. Assume that  $P[\text{length} \geq n] > 0$  then,  $\pi_n, \pi_{n-1}$  are  $P[\_ \mid \text{length} \geq n]$  independent random variables on  $\Sigma_{\geq n}A$ .

Similarly, we see that the full collection  $\{\pi_i\}_{i \leq n}$  is  $P[\_ \mid \text{length} \geq n]$ -independent on  $\Sigma_{\geq n}A$ . Indeed,

$$P[\pi_0 = a_0, \dots, \pi_n = a_n \mid \text{length} \geq n] = \frac{1}{P[\text{length} \geq n]} P[\pi_0 = a_0, \dots, \pi_n = a_n]$$

and

$$\begin{aligned} P[\pi_0 = a_0, \dots, \pi_n = a_n] &= P[\pi_0 = a_0, \dots, \pi_n = a_n] \\ &= P[\pi_n = a_n \mid \pi_{n-1} = a_{n-1}, \dots, \pi_0 = a_0] \\ &\quad P[\pi_{n-1} = a_{n-1} \mid \pi_{n-2} = a_{n-2}, \dots, \pi_0 = a_0] \\ &\quad \dots \\ &\quad P[\pi_0 = a_0] \end{aligned}$$