REVUE D'ANALYSE NUMÉRIQUE ET DE THÉORIE DE L'APPROXIMATION
Tome 21, Nº 1, 1992, pp. 53–58

APPROXIMATION PROBLEMS IN LANGUAGES

CIREŞICA JALOBEANU
(Cluj-Napoca)

Abstract. A metric is introduced into a free semigroup L with unity, generated by a finite set A of symbols. If $L_n \subset L$ is the set of words with maximum length n over the alphabet A, the approximation of a word from L by elements from L_n is discused. If $D \subset L_n$ is a sublanguage for pattern description, the problem of approximation of a word from L_n by words from $D \subset L_n$ is considered. A condition for existence and uniqueness of the best approximation word is done.

Introduction

The automatic identification of curves used in many applications as mass spectrometry or in diagnosis based on electrocardiograms implies the comparison of two curves : a catalogued curve, with a measured one. But the measured curves are subject to measurement errors and the comparison point by point is not suitable. A global comparison can be done using the linguistic analysis of curves [1], [2]. In this case to a segment of a curve a letter from an alphabet is put in correspondence, to a curve — a word, and to a set of experimental curves, a set of words.

The measurement errors make the words corresponding to the same phenomenon differ by a letter or by the order of the letters. The identification problem is to find in a dictionary of the catalogued curves a word, the most likely to a word corresponding to the measured curve.

An approach to this problem can be done by organizing the language as a metric space and using the distance between words as a measure of the similarity. The following question arises : what kind of conditions must be accomplished by the description language so that a correct identification of the word could be done when the unknown word has a number of differences from the words of the dictionary ?

In order to answer this question in this work a study of the approximation problem of a word from a language by the words of a sublanguage is done, when the language is organised as a metric space.

1. The metric space (L, ρ)

Let A be a finite set of symbols, the alphabet, and L the free semigroup with unity over A. In L it is considered the metric

$$\rho : L \times L \to R^+$$

defined by

$$\rho(x,y) = \sum_{i=1}^{\max(n,k)} 1/2^i \, \sigma_i(x, y) \qquad (1)$$

where

$$\sigma_i(x,y) = \begin{cases} 1 \text{ if } x_i \neq y_i \\ 0 \text{ if } x_i = y_i \end{cases}$$

and $x, y \in L$, $x = x_1 \dots x_k$, $y = y_1 \dots y_n$. Such a metric has been used in [3].

The metric ρ induces in L a topology where the open sphere with center $x_0 \in L$ and radius r, $r \in R^+$, is the set

$$S(x_0, r) = \{x : x \in L, \ \rho(x_0, x) < r\}$$

We must observed that in the induced topology the principal right ideals from L are open sets.

Definition 1.1. The subset $H \subset L$ is a right ideal generated by $P \subset L$ if for all $x \in H$, there exists $u \in P$ and $v \in L$ so that $x = uv$.

Definition 1.2. The principal right ideal generated by $\alpha \in L$ is the ideal generated by the set $\{\alpha\}$. So, the right principal ideal generated by $\alpha \in L$ contains all the words from L for which α is one of the prefixes.

From definition 1.1, using the metric ρ the following statement is obvious.

Theorem 1.1. *If H is the right principal ideal generated by $\alpha \in L$, then $\rho(x,y) < 1/2^{|\alpha|}$ for all $x,y \in H$, where $|\alpha|$ is the length of the word $\alpha \in L$.*

We can establish now the

Theorem 1.2. *The principal right ideal generated by $\alpha \in L$ is an open subset in (L, ρ).*

Proof. The right principal ideal generated by $\alpha \in L$ is $H = \alpha L$. Then, for all $y \in H$, $y = \alpha z$, where $z \in L$. The set

$$\{\alpha z\xi : \xi \in L\} \subset H, \text{ hence}$$

$$S(y, 1/2^{|\alpha|+z}) = \{\alpha z\xi : \xi \in L\} \subset H, \text{ and } H \text{ is open.}$$

Now, let L_n be the set of words with maximum length n over the alphabet A.

Definition 1.3. A word $x_0 \in L_n$ is called a best approximation word for $y \in L$ iff

$$\rho(x_0, y) = \inf_{x \in L_n} \rho(x,y).$$

Theorem 1.3. *If $L_n \subset L$, there exists a word and only one in L_n which is the best approximation word for any $y \in L$.*

Proof. Let $|y| = m$

— if $m > n$, then the first n letters from y is a word $x_0 \in L_n$ and $\rho(x_0,y) = 1/2^n - 1/2^m$.

For all $x \in L_n$, $x \neq x_0$ there is at least one different letter on the i position such that

$$\rho(x, y) = 1/2^i + (1/2^n - 1/2^m) > \rho(x_0, y).$$

Hence

$$\rho(x_0, y) = \inf_{x \in L_n} \rho(x, y).$$

— if $m \leqslant n$, then $y \in L_n$ and y is this best approximation element. As the elements from L_n are distinct, the best approximation element is unique.

Definition 1.4. The word m_2 is a left divisor of the word m_1, when $m_1, m_2 \in L$, if there exists $x \in L$, such that $m_1 = m_2 x$.

From theorem 1.3, we can derive the

Theorem 1.4. *If $y \in L$, $D \subset L_n$ and D has at least one left divisor of the word $y \in L$, then there exists in D at least one best approximation word for $y \in L$.*

2. The approximation of the words from L_n by the words of a sublanguage D

Definition 2.1. The metric space (E, d) is called q-discreet metric space if there exists a real number $q \in R^+$, $q \neq 0$, such that for any $x, y \in E$, $x \neq y$, $d(x, y) \geqslant q$. The number q is called the quanta of the discreet metric space.

We will consider now the language $L_n \subset L$, when L_n is the set of words with maximum length n over the alphabet A. It is clear that (L_n, ρ) is a q-discreet metric space with the quanta $q = 1/2^n$. Indeed, let $x, y \in L_n$, if $x = y$, $\rho(x,y) = 0$, but if $x \neq y$ and x, y have all the letter in common, except one, then $\rho(x, y) \geqslant 1/2^n$. Hence, for all $x \neq y$, $x, y \in L_n$, $\rho(x, y) \geqslant 1/2^n$.

In this section we investigate the best approximation problem for words from L_n by words from an arbitrary sublanguage $D \subset L_n$.

Definition 2.2. If $y \in L_n$, then a word $x_0 \in D$ is a best approximation word for $y \in L_n$ if

$$\rho(x_0, y) = \inf_{x \in L_n} \rho(x, y) < 1 - q.$$

Here $\rho(x_0, y) < 1 - q$ means that the best approximation word $x_0 \in D$ has at least a letter in common with $y \in L_n$.

In the next theorem we will note by Pref(y) the set of all words which are prefixes for the word y.

Theorem 2.1. *If $D \subset L_n$ has the property that:*

for all $y \in L_n$ and $\alpha \in Pref(y)$, $H \cap D \neq \emptyset$, when $H = \alpha L$ is the right principal ideal generated by α, then in D there exists at least one best approximation word for all $y \in L_n$.

Proof. This follows immediately from theorem 1.1.

The following theorem can be proved on the uniqueness of the best approximation word when $D \subset L_n$ and for all $z \in L_n$ $\rho(z, D) \leqslant k \leqslant 1/2 - q$, when $k \in R^+$ and $k \neq 0$.

Theorem 2.2. *If $D \subset L_n$ has the property that for all $x, y \in D$, $\rho(x, y) \geqslant 2k + q$, $k \in R^+$ and q is the quanta of the space L_n, then for all $z \in L_n$ such that $\rho(z, D) \leqslant k$, there exists a word and only one of best approximation for $z \in L_n$.*

Proof. As $\rho(z, D) \leqslant k$ and $k \leqslant 1/2 - q$, from definition 2.2 and because D is a finite sublanguage, results the existence of the best approximation element $x \in D$.

As the uniqueness is concerned, let us suppose that there exists two elements $x \in D$ and $y \in D$, such that $\rho(x, z) = \inf_{u \in D} \rho(u, z) \leqslant k$ and $\rho(y, z) = \inf_{u \in D} \rho(u, z) \leqslant k$.

From the triangle inequality results $\rho(x, y) \leqslant \rho(y, z) + \rho(z, x)$ and so $2k + q \leqslant 2k$. This last inequality is false and it follows that $x = y$ contradicting the supposition.

Looking for conditions necessary and sufficient for the existence and uniqueness of the best approximation word we will introduce the following definition.

Definition 2.3. The q-discreet metric space (E, d) having the property that for all $x, y \in E$, $x \neq y$ and $d(x, y) > q$, there exists $z \in E$ such that $d(x, y) = d(x, z) + d(y, z)$, will be called M-convex discreet metric space. (Convex in Menger sense [4]).

First of all we must observe that the discreet metric space (L_n, ρ) is not M-convex. Indeed, let $x, y \in (L_n, \rho)$, $x \neq y$ and $\rho(x, y) = 2q = 1/2^{n-1}$; it means that if $x = x_1 \ldots x_{n-1}x_n$, then $y = x_1 \ldots y_{n-1}x_n$. Now, for all $z \in L_n$, $z \neq x$, $z \neq y$, such that $\rho(x, z) = q = 1/2^n$, $z = x_1 \ldots x_{n-1}z_n$ and so $\rho(y, z) = 1/2^{n-1} + 1/2^n$. Hence $\rho(x, z) + \rho(z, y) = 1/2^{n-2} \neq \rho(x, y)$.

It can be verified that the space L_n becomes M-convex in respect with the following metric, derived by Hamming's metric

$$d : L_n \times L_n \to R^+$$

$$d(x, y) = 1/n \sum_{i=1}^{n} \sigma_i(x, y), \text{ where}$$

$$\sigma_i = \begin{cases} 1 & \text{if } x_i \neq y_i \\ 0 & \text{if } x_i = y_i \end{cases} \quad i = 1, \ldots, n.$$

We are now ready to establish the theorem on the existence and uniqueness of the best approximation word.

Theorem 2.3. *Let (L_n, d) be a discreet metric space M-convex and $z \in L_n$ such that $d(z, D) \leqslant tq \leqslant 1/2 - q/2$. There exists one and only one best approximation word in D for $z \in L_n$, iff $d(x, y) \geqslant q(2t + 1)$, for $t \in N$ a natural number and for all $x, y \in D$.*

Proof. The condition of the theorem are sufficient, as it results from theorem 2.2. In order to prove the necessity of the conditions, we will prove the following.

Lemma 2.1. *If (L_n, d) is a q-discreet metric space and M-convex, then for every $x, y \in L_n$ such that $d(x, y) = 2qt$, when $t \in N$, there exists $z \in L_n$ such that $d(x, y) = tq$ and $d(z, y) = tq$.*

Indeed, if $x, y \in L_n$, $x \neq y$ and $d(x, y) = 2q$, then there exists $z \in L_n$, $z \neq x$, $z \neq y$ such that $d(x, y) + d(z, y) = 2q$, as L_n is M-convex. From the fact that L_n is discreet, there results $d(x, z) = d(y, z) = q$.

If $x, y \in L_n$, $x \neq y$ and $d(x, y) = 2qt$, the space being discreet and M-convex, there are $z_1, z_2, \ldots, z_{2t-1}$ distinct elements from L_n such that

$$d(x, z_1) = d(z_1, z_2) = \ldots = d(z_{2t-1}, y) = q.$$

Hence for $z_i \in L_n$, $d(x, z_i) = d(z_i, y) = qt$.

Based on this lemma, a proof of theorem 2.3 is the following.

Let us consider $u, v \in D$ such that $d(u, v) = \inf_{x,y \in D} d(x, y) = 2qt$. Then by the M-convexity of the space L_n and by lemma 2.1, there results the existence of an element $w \in L_n$ such that $d(u, w) = d(w, v) = qt$ and hence $d(w, D) \leqslant qt$. But this is impossible because
— if $d(w, D) = qt$, the best approximation word is not unique, in contradiction with the hypothesis.
— if $d(w, D) < qt$, there exists a word $y \in D$ such that $d(w, y) < qt$, then, $d(u, v) \leqslant d(u, w) + d(w, y) < 2qt$, in contradiction with the hypothesis $\inf_{x,y \in D} d(x, y) = 2qt$.

Theorem 2. 3 gives a reply to our question : How must a description catalog D be organized in order to recover errors in words with at most t errors? Indeed, a catalog is a sublanguage $D \subset L_n$. To recover a word with some errors means to find in the catalog D one word and only one, the most likely to it, i. e. the best approximation word.

Let us introduce the following definition

Definition 2.4. A word $z \in L_n$ has at most t errors in respect with a set $D \subset L_n$, if $d(z, D) \leqslant tq$.

Using this definition, from theorem 2.3, the following result is obvious.

Theorem 2.4. *If (L_n, d) is a discreet metric space M-convex and $D \subset L_n$, a word $z \in L_n$ with at most t errors in respect with D can be recovered iff $d(x, y) \geqslant (2t + 1)q$, for all $x, y \in D$.*

This last sentence in a new form of the well-known Hamming's theorem on the linear codes [5]. But here the result takes place for finite languages if a metric is considered in respect with which L_n is a q-discreet metric space M-convex.

3. Conclusions

The description language D used in data bases for identification is a sublanguage of the discreet metric space L_n. The identification problem is then a problem of best approximation of $y \in L_n$ with respect to $D \subset L_n$.

Sufficient conditions for the existence of a unique best approximation word have been given. It has been proved that the same conditions are necessary and sufficient if the discreet metric space L_n is M-convex.

These results provide cnditions for the description language $D \subset L_n$ which permit the correct identification of a word with not more than a given number of errors.

Acknowledgements. The author is indebted to Prof. Dr. Doc. Elena Popoviciu for many helpful discussions during the elaboration of this work.

REFERENCES

1. K. S. Fu, *Syntactic methods in pattern recognition*, Acad. Press New-York London, 1974.
2. V. V. Mottl, I. B. Muchnik, *Linguistic analysis of experimental curves*, Proc. of the IEEE *5*, (1979), 714—736.
3. T. Aizawa, T. Ebara, K. Ozeki, Y. Uesaca, *Sur l'espace topologique lié à une nouvelle théorie de l'apprentisage*, Kybernetik *14*, (1974), 141—149.
4. W. V. Peterson, *Error-correcting codes*, MIT Press, New-York, 1961.

Received 1.V.1991

Received 1.V.1991

Institutul Politehnic
Str. Emil Isac 15
R—3400 Cluj-Napoca