Some remarks on prefix and suffix codes

G. Pirillo

IASI CNR, V.le G. B. Morgagni 67/A, Firenze, Italy
Université de Marne-la-Vallée 5,
Boulevard Descartes Champs sur Marne, 77454 Marne-la-Vallée Cedex2
pirillo@math.unifi.it

Abstract

We introduce some new classes of codes and we prove that their elements are the well known unbordered words. We recently proved that a **finite** *code is circular if and only if it satisfies a property $\mathcal{P}$, see* Definition 2.6. *Here, we show also that in the new classes we introduce here there are codes that are* **infinite**, *circular and do not satisfy the property $\mathcal{P}$. So not only the argument of our previous result required the finiteness hypothesis but also the statement requires it in order to be correct.*
Keywords: prefix codes, suffix codes, bifix codes, circular codes, unbordered words.
AMS Subj. Classification: 94A45

1 Introduction

We introduced in [8] the notion of "tiling" (tessellation) of a word w and using it we presented there a general result (contained "in nuce" in [7]) which is the following: there is a hierarchy of codes $\mathcal{P}_0$, $\mathcal{P}_1$, ..., $\mathcal{P}_n$, ... such that $\mathcal{P}_0$ is the class of comma-free codes and a finite code X is circular if and only if there exists a positive integer n such that X is in class $\mathcal{P}_n$.

As we already remarked in [8], this result has pratically been suggested by the concrete study of the combinatorial properties of trinucleotides that, on the 4 letter genetic alphabet, naturally consitute a finite set of words. Moreover, the finiteness condition on the code was necessary for the use of our argument in the proof of the result of [8].

Here we introduce the property $\mathcal{L}$ (roughly speaking a subset X has property $\mathcal{L}$ when, if u and v of X have a strict overlap on a non empty word w, then $w = v$) and the property $\mathcal{R}$ (roughly speaking a subset X has property $\mathcal{R}$ when, if u and v of X have a strict overlap on a non empty word w, then $w = u$).

We prove here that the infinite set of words, better defined hereafter, $X = \{1, 2, \ldots, n, n+1, \ldots, 0a, 01a, \ldots 012\ldots na, 012\ldots n(n+1)a, \ldots\}$ which has property $\mathcal{L}$ and property $\mathcal{R}$, is infinite, circular and does not have property $\mathcal{P}$. The same statement holds for the infinite set of words $X' = \{1, 2, \ldots, n, n+1, \ldots, a0, a10, \ldots, an\ldots 210, a(n+1)n\ldots 210, \ldots\}$.

So not only the proof of our proposition of [8] requires the finiteness condition on the code but also the statement itself should be wrong without the same hypothesis of finiteness.

2 Preliminary definitions and properties

We denote by A an *alphabet*, by A^* the *free monoid* on A, by A^+ the *free semigroup* on A, by ϵ the *empty word*, and, finally, by $|u|$ the *length* of a word $u \in A^*$. We consider a word u of length $k \geq 1$ as a map $u : \{1, 2, \ldots, k-1, k\} \to A$; we write $u = u(1)u(2)\cdots u(i)\cdots u(k-1)u(k)$. A word u is a *factor* of a word v if there exist two words $u', u'' \in A^*$ such that $v = u'uu''$. When $u' = \epsilon$ (resp. $u'' = \epsilon$) we say that u is a *prefix* (resp. *suffix*) of v. A *proper factor* (resp. *proper prefix*, *proper suffix*) u of v is a factor (resp. prefix, suffix) u of v such that $0 < |u| < |v|$. See [5].

A (right) *infinite word* on A is a map q from the set of positive integers into A. We write $q = q(1)q(2)\cdots q(i)\cdots$. A word u is a *factor* of q if there exist a word u' and an infinite word q' such that $q = u'uq'$. If $u' = \epsilon$ we say that u is a *prefix* of q. A non-empty word u may be a

factor of another (finite or infinite) word w in more than one way. So it is useful to speak about occurrences. For this reason, let i, j be integers such that $1 \leq i \leq j$ (with $j \leq |w|$ if w is a finite word) and let us denote by $w(i,j)$ the factor $w(i) \cdots w(j)$ of w. We say that the pair of integers (i,j) is an *occurrence* of the factor u in the word w if $u = w(i,j)$. Given a subset X of A^* we denote by X^n the set of the words on A which are product of n words of X. We denote by A^ω the set of the infinite words on A and by X^ω the set of the infinite words on A which have the form $x_1 x_2 \cdots x_i \cdots$ with $x_i \in X$.

The definitions of code (see [2], [4], [5] and [6]) and circular code (see [7] and [8]) that we recall in this section are well–known in the literature.

Definition 2.1. *Given an infinite word $s = s(1)s(2)\cdots s(i)\cdots$ and factors*
$$w = s(i,j),\ w_1 = s(i_1,j_1),\ w_2 = s(i_2,j_2),\ \ldots,\ w_n = s(i_n,j_n)$$
of s, we say that $w_1, w_2, \ldots, w_n$ is a tiling *(*tessellation*) of w if $j_1 + 1 = i_2$, $j_2 + 1 = i_3$, $\ldots$, $j_{n-1} + 1 = i_n$ and $i_1 \leq i \leq j \leq j_n$. We say that $w = s(i,j)$ is the* trivial tiling *of w.*

In Figure 1, x_1x_2 is a tiling of y_1. Also $x_1x_2x_3$ is a tiling of y_1 but it is not *minimal*, see [8] for a formal definition, as the *unnecessary* word x_3 is used. In the same figure x_3x_4 is a minimal tiling of y_2 and x_4x_5 is a minimal tiling of y_3.

In the following definition we recall the notion of equivalent tilings. For example, the tilings of y_{i_1} and y_{i_2} in Figure 2 are equivalent.

Definition 2.2. *Let $s = s(1)s(2)\cdots s(i)\cdots$ be an infinite word and let $s(i,j)$ and $s(i',j')$ be two occurrences of the same factor w of s. Then, given a tiling $s(i_1,j_1), s(i_2,j_2), \ldots, s(i_n,j_n)$ of $s(i,j)$ and a tiling $s(i'_1,j'_1), s(i'_2,j'_2), \ldots, s(i'_n,j'_n)$ of $s(i',j')$, we say that these two tilings are* equivalent *if there exist factors $w_1, w_2, \ldots, w_n$ of s such that $s(i_1,j_1) = s(i'_1,j'_1) = w_1$, $s(i_2,j_2) = s(i'_2,j'_2) = w_2$, $\ldots$, $s(i_n,j_n) = s(i'_n,j'_n) = w_n$ and there exist factors u, u', v, v' of s such that $w_1 w_2 \ldots w_n = us(i,j)u' = vs(i',j')v'$ with $|u| = |v|$.*

Definition 2.3. *A subset X of A^+ is said to be a* code *over A if for all $n, m \geq 1$ and $x_1, \ldots, x_n, x'_1, \ldots, x'_m \in X$ the condition*
$$x_1 \cdots x_n = x'_1 \cdots x'_m$$
implies
$$n = m \text{ and } x_i = x'_i \text{ for } i = 1, \ldots, n.$$

Definition 2.4. *Let $\{x_\alpha\}_{\alpha \geq 1}$ be a fixed infinite sequence of finite words from a set X and let x be an infinite sequence such that*
$$x = x_1 x_2 \cdots x_\alpha \cdots \in X^\omega.$$
We say that the infinite set of integers $\{1 = i_1, i_2, \ldots, i_\alpha, \ldots\}$ satisfying $i_1 < i_2 < \cdots < i_\alpha < \cdots$ is the natural tiling set *of x if $x_1 = x(i_1, i_2 - 1), x_2 = x(i_2, i_3 - 1), \ldots, x_\alpha = x(i_\alpha, i_{\alpha+1} - 1), \ldots$ and we denote it by $T(x)$. If $y = y_1 y_2 \cdots y_n \in X^n$ and $y_1 y_2 \cdots y_n = x(j_1, j_2 - 1)x(j_2, j_3 - 1)\cdots x(j_n, j_{n+1} - 1)$ is an occurrence of y in x, we say that the finite set of integers $\{j_1, j_2, \ldots, j_n, j_{n+1}\}$, with $j_1 < j_2 < \cdots < j_n < j_{n+1}$, is the* local tiling set *of this occurrence of y in x and, shortly, we denote it by $T(y)$.*

Definition 2.5. *Let X be a subset of A^+ and let k be a non-negative integer. We say that X has the* property $\mathcal{P}_k$ *if, for each infinite sequence $x \in X^\omega$ and local tiling set $T(y)$ of an occurrence of a factor y in x, we have*
$$Card(T(y) \setminus T(x)) \leq k.$$

Definition 2.6. *Let X be a subset of A^+. We say that X has the* property $\mathcal{P}$ *if X has the property $\mathcal{P}_k$ for some non-negative integer k.*

Proposition 2.1. *If A is an alphabet and X is a subset of A^+ having the property $\mathcal{P}$, then X is a code.*

Proof: See [8].

■

Definition 2.7. *A subset X of A^+ is a* circular code *over A if, for each $n, m \geq 1$ and for each $x_1, \ldots, x_n, x'_1, \ldots, x'_m$ in X, $p \in A^*$ and $s \in A^+$, the condition*

$$sx_2 \cdots x_n p = x'_1 \cdots x'_m$$

with $x_1 = ps$ implies $n = m$, $p = \epsilon$ and, for $i = 1, \ldots, n$, $x_i = x'_i$.

Proposition 2.2. *Given an alphabet A and a finite subset X of A^+ the following conditions are equivalent:*

i) X has the property $\mathcal{P}$;

ii) X is a circular code.

Proof: See [8].

■

Remark 2.1. *The letters (or nucleotides) define the genetic alphabet $\mathcal{A}_4$=$\{A, C, G, T\}$. The set of the 64 words of length three or trinucleotides is denoted by $\mathcal{A}_4^3$. Imagine writing an infinite sequence of trinucleotides x_i, with $x_i \in X$, where X is a subset of $\mathcal{A}_4^3$ and imagine shifting of one or two trinucleotides the reading frame. As we already remarked in [8], perhaps we are able to read as a prefix at least one trinucleotide of X in the shifted sequence (in this case X is not a comma free code). Perhaps we are able to read as a prefix even the product of two consecutive trinucleotides of X and so on. Perhaps we can factorize with trinucleotides of X the whole shifted sequence. Anyway, if X is circular, i.e. X has property $\mathcal{P}$, we are able at most to read four consecutive trinucleotides of X (this corresponds to the window of Arquès and Michel which has 13 nucleotides, see also [7]).*

Definition 2.8. *Given a code X over A we say that it is* prefix *(resp.* suffix*) if $u = v$ whenever $u, v \in X$ and u is a prefix (resp. suffix) of v. A code is* bifix *if it is both a prefix code and a suffix code.*

Now, we introduce the following definitions.

Definition 2.9. *A subset X of A^+ has property $\mathcal{L}$ if, for each $u, v \in X$ and for each $w \in A^+$, if w is a suffix of u and a prefix of v then $w = v$.*

Definition 2.10. *A subset X of A^+ has property $\mathcal{R}$ if, for each $u, v \in X$ and for each $w \in A^+$, if w is a suffix of u and a prefix of v then $u = w$.*

Definition 2.11. *A subset X of A^+ has property $\mathcal{LR}$ if, for each $u, v \in X$ and for each $w \in A^+$, if w is a suffix of u and a prefix of v then $u = w = v$.*

Remark 2.2. *In the previous definitions, u and v constitute an ordered pair of words and w is a non-empty word. So, the pair (u, v) has a strict overlap in the sense that there exist z and z' (possibly empty!) such that $u = zw$ and $v = wz'$. Note that the pair (u, v), where $u = a$ is a word of length 1 and $v = ab$ is a word of length 2, has a strict overlap on the non-empty word a, but the pair (v, u) has no strict overlaps! On the other hand, the pair (u, v), where $u = b$ is a word of length 1 and $v = ab$ is a word of length 2, has no strict overlaps but the pair (v, u) has a strict overlap on the non-empty word b! In Figure 3, 4, and 5 there are examples of strict overlaps. In particular, in Figure 3 the pair (u, v) has a strict overlap on the non-empty word w ($0 < |w| < |u|$, $0 < |w| < |v|$), in Figure 4 the pair (u, v) has a strict overlap on the non-empty word u ($0 < |u| < |v|$, i.e., u is a proper prefix of v) and in Figure 5 the pair (v, u) has a strict overlap on the non-empty word u ($0 < |u| < |v|$, i.e., u is a proper suffix of v).*

Let us also recall the following definition (see [2], for example).

Definition 2.12. *A subset X of A^+ is prefix set (resp, suffix set) if no $u \in X$ is a proper prefix (resp, proper suffix) of another $v \in X$. A subset X of A^+ is bifix set if X is both a prefix and a suffix set.*

It is well known that if a subset X of A^+ is a prefix (resp. suffix, bifix) set different from $\{\epsilon\}$ then X is a prefix code (resp. suffix code, bifix code). See [2].

3 Results

Proposition 3.1. *If a subset X of A^+ has property $\mathcal{L}$ then X is a prefix code.*

Proof: It is sufficient to prove that X is a prefix set. Let's argue by contradiction. Suppose that this is not the case. So, for some $u, v \in X$, u is a proper prefix of v. Since u is a suffix of u and u is a prefix of v we have that X does not verify property $\mathcal{L}$ (indeed, $u \neq v$). Contradiction.

■

Proposition 3.2. *If a subset X of A^+ has property $\mathcal{R}$ then X is a suffix code.*

Proof: Similar to the proof of the previous proposition.

■

Proposition 3.3. *If a subset X of A^+ has property $\mathcal{LR}$ then X is a bifix code.*

Proof: If X has property $\mathcal{LR}$ it has property $\mathcal{L}$ and also property $\mathcal{R}$. By previous propositions, it is a prefix and a suffix code. Hence it is a bifix code.

■

Definition 3.1. *A prefix code (resp. suffix code, resp bifix code) X having property $\mathcal{L}$ (resp. property $\mathcal{R}$, property $\mathcal{LR}$) is called an $\mathcal{L}$-code (resp. $\mathcal{R}$-code, $\mathcal{LR}$-code).*

Proposition 3.4. *The inclusion of the $\mathcal{L}$-codes (resp $\mathcal{R}$-codes, $\mathcal{LR}$-codes) in the class of prefix codes (resp suffix codes, bifix codes) is strict.*

Proof: The code ab, ba is prefix, suffix and bifix but it is not a $\mathcal{L}$-code, nor a $\mathcal{R}$-code nor a $\mathcal{LR}$-code.

■

Remark 3.1. *If u and v are the words of Figure 3 ($0 < |w| < |u|$, $0 < |w| < |v|$), then $\{u, v\}$ cannot be a $\mathcal{L}$-code. Also $\{u, v\}$ cannot be a $\mathcal{R}$-code and, a fortiori, $\{u, v\}$ cannot be a $\mathcal{LR}$-code. If u and v are the words of Figure 4 where u is a proper prefix of v (resp. Figure 5 where u is a proper suffix of v) then by Proposition 2 (resp. Proposition 3) $\{u, v\}$ cannot be a $\mathcal{L}$-code (resp. $\mathcal{R}$-code). On the other hand, if u and v are the words of Figure 5 (resp. Figure 4), the set $\{u, v\}$ could be a $\mathcal{L}$-code (resp. $\mathcal{R}$-code): for example $\{b, ab\}$ is a $\mathcal{L}$-code and $\{a, ab\}$ is a $\mathcal{R}$-code.*

Let us recall the following definition of [3], for example.

Definition 3.2. *A non empty word u is called a border of a word v if $v = uw = w'u$ for some suitable words w, w'. We call v bordered if it has a border which is shorter than v, otherwise v is called unbordered.*

So, a word v is *unbordered* if no proper non-empty prefix of v is also a proper suffix of v.

Proposition 3.5. *The elements of a $\mathcal{L}$-code (resp. $\mathcal{R}$-code, $\mathcal{LR}$-code) are unbordered words.*

Proof: Let's argue by contradiction. Suppose that a word v of an $\mathcal{L}$-code is a bordered word and suppose that u is one of its border. So u is both a (proper non-empty) prefix and a (proper non-empty suffix) of v and, by property $\mathcal{L}$, it must coincide with v. Contradiction. The proof for $\mathcal{R}$-codes and $\mathcal{LR}$-codes is similar.

■

Proposition 3.6. *Any pair of distinct unbordered words is a code.*

Proof: Let's argue by contradiction. Let u, v a pair of distinct unbordered words. Suppose that $\{u, v\}$ is not a code. So, for some n, m positive integers, for some $x_i \in \{u, v\}$ and $y_i \in \{u, v\}$ an equality $x_1 x_2 \cdots x_n = y_1 y_2 \cdots y_m$ holds. By the Defect Theorem (see [1] and also [2]), there is a word w and two positive integers h, k such that $u = w^h$ and $v = w^k$. Since u, v are distinct, we have that at least one of them is a bordered word. Contradiction.

■

Remark 3.2. *It is possible to avoid the use of Defect Theorem and prove the statement with an argument on the minimality of the above considered equality $x_1 x_2 \cdots x_n = y_1 y_2 \cdots y_m$. Indeed, without loss of generality, we have the following two cases:*

a) $(x_1, x_n) = (u, v)$ *and* $(y_1, y_m) = (v, u)$

b) $(x_1, x_n) = (u, u)$ *and* $(y_1, y_m) = (v, v)$.

In both cases, if, again without loss of generality, we suppose that u is shorter than v then we have that u is a prefix as well as a suffix of v. This ends this alternative proof.

Proposition 3.7. *The three words $\{a, b, ab\}$ are unbordered but do not constitute a code.*

Proof: Easy: $ab = (a)(b)$.

■

Remark 3.3. *Clearly $\{ab\}$ and $\{ba\}$ have property $\mathcal{L}$ but $\{ab, ba\}$ does not have property $\mathcal{L}$. Similarly for properties $\mathcal{R}$, and $\mathcal{LR}$. But, by the previous Proposition 7 and 8, we have the following: if $\{u\}$ and $\{v\}$ have property $\mathcal{L}$ (resp. $\mathcal{R}$, $\mathcal{LR}$) then $\{u, v\}$ is a code.*

Now, consider the set of the non–negative integers N as an alphabet and consider also the enlarged alphabet $N \cup \{a\}$ where a is a letter which does not belong to N. Consider also, for each $n \geq 0$, the infinite family $N^{<}a$ (resp. $N^{>}a$) of words of the free semigroup $(N \cup \{a\})^+$ having the form $012 \ldots (n-1)na$ (resp. $an(n-1) \ldots 210$).

Note that, for each $n \geq 0$, the word $012 \ldots (n-1)na$ (resp. $an(n-1) \ldots 210$) has length $n+2$, has a as a suffix (resp. prefix) and contains a prefix (resp. suffix) of length $n+1$ which is the juxtaposition of the first n non–negative integers in increasing (resp. decreasing) order. Finally, consider the two languages X and X' on $N \cup \{a\}$ defined as follows

$$X = N \setminus \{0\} \cup N^{<}a$$

and

$$X' = N \setminus \{0\} \cup N^{>}a.$$

We have that X and X' look as follows:

$$X = \{1, 2, \ldots, n, n+1, \ldots, 0a, 01a, \ldots 012 \ldots na, 012 \ldots n(n+1)a, \ldots\}$$
$$X' = \{1, 2, \ldots, n, n+1, \ldots, a0, a10, \ldots, an \ldots 210, a(n+1)n \ldots 210, \ldots\}.$$

Proposition 3.8. *The language X*
i) is infinite;
ii) has property $\mathcal{LR}$;
iii) is circular;
iv) does not have property $\mathcal{P}$.

Proof:

i) Clear.

ii) Suppose $u, v \in X$ and $w \in (N \cup \{a\})^+$. Suppose also that w is a suffix of u and a prefix of v. Let us distinguish two cases: a) $|w| = 1$ and b) $|w| \geq 2$.

a) We have $w \neq 0$ because u cannot end with 0 and $w \neq a$ because v cannot begin with a. So $w \in N \setminus \{0\}$. As the unique word of X which begins with w is w itself, we have $w = v$. Similarly, $w = u$.

b) The word w begins with 0, being a prefix of v, and ends with a, being a suffix of u. Now, as v has no proper prefix which ends with a, we have $w = v$. Similarly (as u has no proper suffix which begins with 0) $w = u$. As in all cases we have $u = w = v$, the language X has property $\mathcal{LR}$.

iii) Let's argue by contradiction. Suppose that X is not a circular code. Then there exist $x_1, \ldots, x_n, x'_1, \ldots, x'_m \in X$, $p \in (N \cup \{a\})^+$ and $s \in (N \cup \{a\})^+$, such that $sx_2 \cdots x_n p = x'_1 \cdots x'_m$, $x_1 = ps$ and the conditions of Definition 8 are not satisfied. If some x_i is in $N^{<}a$ then the product $x'_1 \ldots x'_m$ must contain at least one occurrence of a. This a must be the last letter of some x'_j. So $x'_j x'_{j+1} \ldots x'_m x'_1 x'_2 \ldots x'_{j-1} = x_i x_{i+1} \ldots x_n x_1 x_2 \ldots x_{i-1}$. So X is not a code. Similarly, if some x'_i is in $N^{<}a$ then X is not a code. So we can suppose that all the x_i's and all the x'_i's are in $N \setminus \{0\}$. Being all the elements of $N \setminus \{0\}$ of length 1, consequently also x_1 must have length 1. As $p \in (N \cup \{a\})^+$, $s \in (N \cup \{a\})^+$ and $x_1 = ps$, we are in contradiction. So X is a circular code.

iv) Put $n_a = 012 \ldots na$ and consider the infinite sequence $x = 0_a 1_a 2_a \ldots n_a \ldots$. It contains, for each n, an occurrence of $(1)(2) \ldots (n)$ and so, for each $n \geq 1$, the code X cannot have property $\mathcal{P}_{n+1}$. So X does not have property $\mathcal{P}$.

■

Proposition 3.9. *The language X':*
i) is infinite;
ii) has property $\mathcal{LR}$;
iii) is circular;
iv) does not have property $\mathcal{P}$.

Proof: ii) Similar to the previous proof.

■

Remark 3.4. *The study of the homogeneous codes X having property $\mathcal{L}$ (or property $\mathcal{R}$, or property $\mathcal{LR}$) seems to be interesting. In particular, this is especially true for the subsets of $\mathcal{A}_4^3$.*

4 Acknowledgements

We thank Dipartimento di Matematica "U. Dini" for giving us a friendly hospitality and Jacques Justin for very helpful conversations.

5 References

[1] J. Berstel, D. Perrin, J.-F. Perrot, A. Restivo, Sur le théorème du défaut, *J. Algebra*, 60, 1979, 169-180.

[2] J. Berstel, D. Perrin, Theory of codes, *Academic Press*, 1985.

[3] J.-P. Duval, T. Harju, D. Nowotka, Unbordered factors and Lyndon words, *Discrete Mathematics* 308 (2008) 2261–2264.

[4] J. Justin, G. Pirillo, On some factorizations of infinite words by elements of codes, *Inform. Process. Lett.*, 62(6), 289–294, 1997.

[5] M. Lothaire, Combinatorics on words, *Addison-Wesley*, 1983.

[6] G. Pirillo, Infinite words and biprefix codes, *Inform. Process Lett.*, 50, 1994, 293-295.

[7] G. Pirillo, A characterization for a set of trinucleotides to be a circular code in *Determinism, Holism, and Complexity* (Edited by C. Pellegrini, P. Cerrai, P. Freguglia, V. Benci and G. Israel), Kluwer (2003).

[8] G. Pirillo, A hierarchy for circular codes, *RAIRO-Theor. Inf. Appl.* 42 (2008) 717–728.

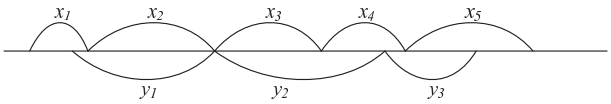

Figure 1: Examples of tilings.

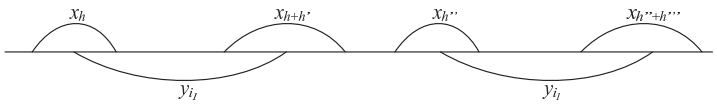

Figure 2: The equivalent tilings of y_{i_1} and y_{i_2}.

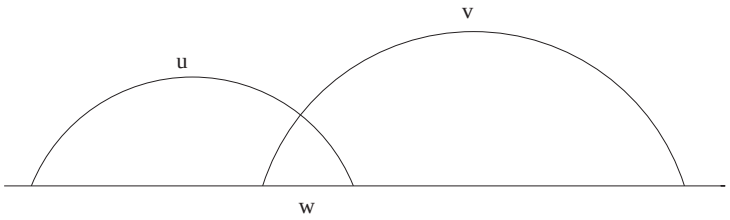

Figure 3: Example of overlap.

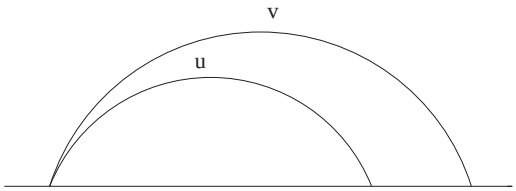

Figure 4: Example of overlap.

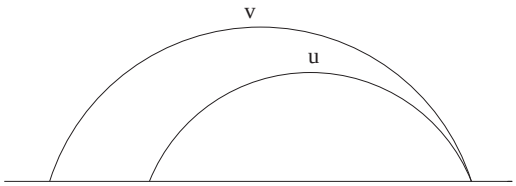

Figure 5: Example of overlap.