A Gray code for a regular language

Elena Barcucci
elena.barcucci@unifi.it

Antonio Bernini
antonio.bernini@unifi.it

Renzo Pinzani
renzo.pinzani@unifi.it

Dipartimento di Matematica e Informatica “U. Dini”,
Università degli Studi di Firenze, Viale G.B. Morgagni 65, 50134 Firenze, Italy.

Abstract

Given a sequence $\{a_m\}_{m\geq 0}$ of strictly increasing positive integers such that $a_0 = 1$, any non-negative integer N can be uniquely represented by $N = \sum_{i\geq 0} d_i a_i$, where d_i are non negative integers. The coefficients d_i form a string over a finite alphabet and all these strings form a language having different properties depending on the sequence. We investigate on the possibility of defining a Gray code over the language arising from particular choices of $\{a_m\}_{m\geq 0}$. We consider the sequence defined by a two termed linear recurrence with constant coefficients having some particular properties.

1 Introduction

Given a class of combinatorial objects, it is a common problem to list them in some specified order. A famous kind of list is the well-known Gray code [Gra53] where two successive objects differ according to some fixed constraints. Very often successive objects in a Gray code are required to differ as little as possible, depending on the nature of the objects we are dealing with. For instance, if the objects are strings (of the same length) they form a Gray code if two successive strings differ only in a few positions or, more precisely, if their Hamming distance d_H [Ham50] is bounded by some positive integer q ($d_H = i \leq q$). Gray codes have been constructed for several combinatorial structures (permutations, binary strings, Motzkin and Schröder words, derangements, involutions) and used in various technical applications such as circuit testing, signal encoding, data compression (see [BBPSV14, BBPSV15, BBPV15, BGPP07] and references therein). Even though the most well-known Gray codes have been studied in the context of the mentioned topics, here we only recall that recently [BBPSV14, BBPV17] Gray codes have been considered also in the framework of cross-bifix-free sets of strings (two strings are *bifix-free* if they can not be overlapped in any way, for details see for example [BPP12]). Moreover, in [BBBP17a], where cross-bifix-free sets of matrices are considered, the authors propose a Gray code where the matrices are listed in such a way that any two successive matrices differ in only one digit. Note that the problem of finding a Gray code for bidimensional structures (matrices) is also considered in [BBBP17b]; nevertheless the approach used in [BBBP17a] doesn't seem to be not useful.

In the present paper we consider a language of strings defined in [BR01] over a particular alphabet. These strings are the representations of integers by means of a system of numeration [Fra82] derived from certain sequences. Given a strictly increasing sequence of non-negative integers $1 = a_0 < a_1 < a_2 < ...$, it is possible to represent each non-negative integer N by recording the quotient d_n obtained by dividing N by the largest member a_n of the sequence that is less than or equal to N, then dividing the remainder r_n by a_{n-1} and recording the

In: L. Ferrari, M. Vamvakari (eds.): Proceedings of the GASCom 2018 Workshop, Athens, Greece, 18–20 June 2018, published at http://ceur-ws.org

quotient d_{n-1} and so on until you get to d_0, since dividing by 1 leaves a remainder of 0. The various quotients form a string which is an element of the language. In [BR01] sequences are defined by a two termed linear recurrence depending on two parameters k and h, and the arising language, strictly depending on k and h, is seen as a combinatorial interpretation of the sequence $a_i = ka_{i-1} + ha_{i-2}$. In this paper we aim at providing a Gray code for the language derived from particular conditions on k and h.

We point out that the considered language was introduced in order to provide a general solution to a previous issue appeared in [BSS93], where the authors asked for a combinatorial interpretation of the recurrence $f_{m+1} = 6f_m - f_{m-1}$, with $f_0 = 1$, $f_1 = 7$ (sequence M4423 of [SP96]). After some interesting answers (see [BBDD98, PP90, Sul98]), in their paper [BR01] the authors gave a general combinatorial interpretation for the recurrences of the form $a_m = ka_{m-1} + ha_{m-2}$, under some conditions on h and k which include the most frequently occurring two-termed recurrences. We also noted, in the previous paragraph, that this language is strictly linked to the system of numeration presented in [Fra82]. This is based on a very simple iterated division algorithm which is the main tool which produces the representation for any non-negative integer N, in terms of any sequence of the form $1 = u_0 < u_1 < u_2 < \dots$. Clearly, the particular choice of the sequence affects the representation of N, and in some particular cases interesting and useful representations can be obtained. In the references within [Fra82] several application of different systems of numeration can be found, ranging from compressing and partitioning large dictionaries, ranking permutations with repetitions, up to designing error-insensitive codes for data transmission.

The definition of the Gray code in the present paper could be presented independently of the definition of numeration systems and without showing which are the links between the considered language and the sequence. Nevertheless, in order to provide a self-contained paper and for the sake of completeness, we prefer to present concepts and preliminaries useful for our purpose. Section 2 and Section 3 are devoted to the presentation of some main notion about numeration systems and properties of the considered recurrence relations. In Section 4 we recall the definition of the language introduced in [BR01] and finally we define a Gray code for listing its elements.

2 Preliminaries

Given a sequence $\{a_m\}_{m\geq 0}$ of positive integers such that $a_0 = 1$ and $a_m < a_{m+1}$ for each $m \in \mathbb{N}$, let N be any non-negative integer. Consider the largest term a_n of the sequence such that $a_n \leq N$. More precisely, $a_n = \max\{a_m \mid a_m \leq N\}$ (for the particular case $N = 0$, see below). We divide N by a_n obtaining $N = d_n a_n + r_n$. Obviously, for the remainder r_n, it is clear that $r_n < a_n$. If we divide r_n by a_{n-1}, we get $r_n = d_{n-1}a_{n-1} + r_{n-1}$, with $r_{n-1} < a_{n-1}$. Then, iterating this procedure until the division by $a_0 = 1$ (where of course the remainder is 0), we have:

$$\begin{aligned}
N &= d_n a_n + r_n && 0 \leq r_n < a_n \ , \\
r_n &= d_{n-1}a_{n-1} + r_{n-1} && 0 \leq r_{n-1} < a_{n-1} \ , \\
r_{n-1} &= d_{n-2}a_{n-2} + r_{n-2} && 0 \leq r_{n-2} < a_{n-2} \ , \\
\cdots &= \dots\dots\dots\dots && \dots\dots\dots\dots \\
\cdots &= \dots\dots\dots\dots && \dots\dots\dots\dots \\
r_3 &= d_2 a_2 + r_2 && 0 \leq r_2 < a_2 \ , \\
r_2 &= d_1 a_1 + r_1 && 0 \leq r_1 < a_1 \ , \\
r_1 &= d_0 a_0 \ .
\end{aligned}$$

The above relations imply that:

$$N = d_n a_n + d_{n-1}a_{n-1} + d_{n-2}a_{n-2} + \dots\dots + d_1 a_1 + d_0 a_0 \ . \tag{1}$$

Expression (1) is the representation of N in the numeration system $S = \{a_0, a_1, a_2, \dots\dots\}$, and the string $d_n d_{n-1} \dots d_1 d_0$ is associated to the number N (in what follows the term “representation” equivalently refers to the expression (1) or to its associated string). The method presented here can be applied to every non-negative integer and in the case $N = 0$, clearly, all the coefficients d_i are 0 (in other words the representation of 0 is simply the string 0). Moreover, we have

$$r_i = d_{i-1}a_{i-1} + d_{i-2}a_{i-2} + \ldots\ldots + d_1a_1 + d_0a_0 < a_i \ , \tag{2}$$

for each $i \geq 0$.

It is possible to show [Fra82] that if $N = \sum_{i\geq 0}^{n} d_i a_i$ with

$$d_ia_i + d_{i-1}a_{i-1} + \ldots + d_1a_1 + d_0a_0 < a_{i+1} \tag{3}$$

for each $i \geq 0$, then the representation $N = \sum_{i\geq 0}^{n} d_i a_i$ is unique. For the sake of completeness, we recall the complete theorem:

Theorem 2.1. *Let $1 = a_0 < a_l < a_2 < \ldots$ be any finite or infinite sequence of integers. Any non-negative integer N has precisely one representation in the system $S = \{a_0, a_1, a_2, ...\}$ of the form $N = \sum_{i\geq 0}^{n} d_i a_i$ where the d_i are non-negative integers satisfying (3).*

As an example, consider the well-known sequence of Pell numbers (sequence M1413 in [SP96]) $p_m = 1, 2, 5, 12, 29, \ldots$ defined by $p_0 = 1$, $p_1 = 2$, $p_m = 2p_{m-1} + p_{m-2}$. The representation of $N = 16$ is associated to the string 1020.

3 The language and the alphabet

The next step is the definition of the language arising from the representations of non-negative numbers. Given $m > 0$, we consider all the integers $\ell \in \{0, 1, 2, \ldots, a_m - 1\}$. According to the scheme of the previous section, the representations of the integers j with $a_{m-1} \leq j < a_m$ is $j = d_{m-1}a_{m-1} + d_{m-2}a_{m-2} + \ldots + d_0a_0$ (so that the associated string is $d_{m-1}d_{m-2}\ldots d_0$), while, following the same scheme, the remaining integers have a representation with less than m digits. For example: the representation of $a_{m-1} - 1 = d_{m-2}a_{m-2} + \ldots + d_0a_0$ has $m - 1$ digits. For our purpose (the construction of a Gray code), we require that all the representations of the considered integers $\ell \in \{0, 1, 2, \ldots, a_m - 1\}$ have m digits, so we pad the string on the left with 0's until we have m digits: the representation of $a_{m-1} - 1$ becomes $a_{m-1} - 1 = 0a_{m-1} + d_{m-2}a_{m-2} + \ldots + d_0a_0$ (therefore, the associated string is $0d_{m-2}\ldots d_0$).

With this little adjustment, we now define the following sets:

$\mathscr{L}_0 = \{\varepsilon\}$,

$\mathscr{L}_m = \{d_{m-1}\ldots\ d_0|$ the string $d_{m-1}\ldots\ d_0$ is the representation of each $\ell < a_m$ in the numeration system $\{a_m\}_{m\geq 0}$ $\}$.

Finally, we denote by $\mathscr{L}$ the language obtained by taking the union of all the sets $\mathscr{L}_m$:

$\mathscr{L} = \bigcup_{m\geq 0} \mathscr{L}_m$.

We remark that each element of $\mathscr{L}_m$ has precisely m digits, so that some string $d_{m-1}\ldots\ d_0$ can admit a prefix constituted by a certain number of consecutive 0's. Moreover, each $\mathscr{L}_m$ contains precisely a_m elements (which are the representations of each $\ell \in \{0, 1, \ldots, a_m - 1\}$).

Referring to the sequence of Pell numbers $p_m = \{1, 2, 5, 12, 29, \ldots\}$ defined in Section 2, we have:

$$\begin{aligned}
\mathscr{L}_0 &= \{\varepsilon\} \\
\mathscr{L}_1 &= \{0, 1\} \\
\mathscr{L}_2 &= \{00, 01, 10, 11, 20\} \\
\mathscr{L}_3 &= \{000, 001, 010, 011, 020, 100, 101, 110, 111, 120, 200, 201\} \\
\mathscr{L}_4 &= \{0000, 0001, 0010, 0011, 0020, 0100, 0101, 0110, 0111, 0120, 0200, 0201, 1000, 1001, 1010, \\
&\qquad 1011, 1020, 1100, 1101, 1110, 1111, 1120, 1200, 1201, 2000, 2001, 2010, 2011, 2020\}
\end{aligned}$$

The strings in $\mathscr{L}_2$ are, respectively, the representations of the integers $\ell \in \{0, 1, 2, 3, 4\}$, being $a_2 = 5$. Note that $\mathscr{L}_2$ contains exactly $5 = a_2$ elements.

It is not difficult to realize that the alphabet of the language $\mathscr{L}$ strictly depends on the sequence $\{a_m\}_{m \geq 0}$. In general it is possible to set an upper bound for the digits d_i. From (3), we deduce $d_i a_i < a_{i+1} - \sum_{j=0}^{i-1} d_j a_j$, so that, since the numbers are all integers:

$$d_i a_i \leq a_{i+1} - 1 - \sum_{j=0}^{i-1} d_j a_j < a_{i+1} - 1 \ ,$$

leading to

$$d_i \leq \left\lfloor \frac{a_{i+1} - 1}{a_i} \right\rfloor \ . \tag{4}$$

Therefore, the alphabet for $\mathscr{L}_m$ is given by $\{0, 1, \ldots, s\}$ with $s = \max\limits_{i=0,1,\ldots,m-1} \left\{ \left\lfloor \frac{a_{i+1} - 1}{a_i} \right\rfloor \right\}$, and, denoting by Σ the alphabet for $\mathscr{L}$, we have $\Sigma = \{0, 1, \ldots, t\}$ with

$$t = \max_i \left\{ \left\lfloor \frac{a_{i+1} - 1}{a_i} \right\rfloor \right\} \ .$$

In this paper we focus our attention on sequences defined by linear recurrences of the form $a_m = k a_{m-1} + h a_{m-2}$, with suitable initial conditions and some restrictions on k and h. More precisely, we consider sequences of the form:

$$a_m = \begin{cases} 1 & \text{if } m = 0 \\ k & \text{if } m = 1 \\ k a_{m-1} + h a_{m-2} & \text{if } m \geq 2 \end{cases}$$

where $k \in \mathbb{N}^+$ and $h \in \mathbb{Z}$. Using standard techniques for solving recurrences (see for example [Aig07]) it is possible to show the the general term of the sequence is

$$a_m = \frac{1}{\sqrt{k^2 + 4h}} \left(\frac{k + \sqrt{k^2 + 4h}}{2} \right)^{m+1} - \frac{1}{\sqrt{k^2 + 4h}} \left(\frac{k - \sqrt{k^2 + 4h}}{2} \right)^{m+1}$$

and, following [BR01], we require the condition $k^2 + 4h \geq 0$ which assures that $a_m \geq 0$ and $a_{m+1} > a_m$ for each $m \geq 0$ (which are the hypothesis of Theorem 2.1).

Actually, here we further restrict to a more limiting case for k and h. More precisely we consider the case $k \geq h \geq 0$, which, of course, implies $k^2 + 4h \geq 0$. From [BR01] it is possible to deduce that in this case the alphabet Σ of the language $\mathscr{L}$ is $\Sigma = \{0, 1, \ldots, k\}$. The language $\mathscr{L}$ is the set of words $w \in \Sigma^*$ such that

1. $w = d_r d_{r-1} \dots d_1 d_0$ with $d_i \in \Sigma$;
2. if $d_i = k$, then $d_{i-1} < h$, for $i = 1, 2, \dots, r$;
3. $d_0 \neq k$.

In the above list, point 2. is due to the fact that, if $N < a_{i+1}$ is an integer such that $d_i = k$, then, if also $d_{i-1} = h$, it is $N \geq ka_i + ha_{i-1} = a_{i+1}$, which is in contrast with $N < a_{i+1}$. Moreover, since in the representation of N it is $r_1 < a_1 = k$, we have $d_0 < k$.

We recall [BR01] that the following unambiguous regular grammar generates $\mathscr{L}$:

$S \to T0|T1|\cdots|T(h-1)|Sh|\cdots|S(k-1)|\epsilon$,

$T \to T0|T1|\cdots|T(h-1)|Sh|\cdots|S(k)|\epsilon$.

4 A Gray code for $\mathscr{L}_n$

We introduce some notations (as in [BBPSV14]) in order to express the language $\mathscr{L}$ in an alternative recursive way.

- If α is a symbol and L is a list of words, $\alpha \cdot L$ is the list obtained by concatenating α to each string of L;
- if i and j are symbols, then $ij \cdot L$ is the list obtained by concatenating i to each string of $i \cdot L$ (or equivalently $ij \cdot L = i \cdot (j \cdot L)$);
- if L is a list of word, $\bar{L}$ is the list in the reverse order;
- if L is a list of word, $(L)^{\underline{i}}$ is L if i is even and $\bar{L}$ if i is odd;
- if L and M are two lists, $L \circ M$ is their concatenation;
- if $L_j, L_{j+1}, \dots, L_{j+r}$ are lists, $\bigcirc_{\ell=0}^{r} L_{j+\ell}$ is the list $L_j \circ \dots \circ L_{j+r}$.
- if L is a list of words, then $first(L)$ is the first element of L and $last(L)$ is the last element of L.

With the above notation it is not difficult to realize that $\mathscr{L}$ can be defined as follows:

$$\mathscr{L}_n = \begin{cases} \{\varepsilon\} & \text{if } n = 0 \\ \{0, 1, \dots, k-1\} & \text{if } n = 1 \\ 0 \cdot \mathscr{L}_{n-1} \cup 1 \cdot \mathscr{L}_{n-1} \cup \dots \cup (k-1) \cdot \mathscr{L}_{n-1} \cup \\ k0 \cdot \mathscr{L}_{n-2} \cup k1 \cdot \mathscr{L}_{n-2} \cup \dots \cup k(h-1) \cdot \mathscr{L}_{n-2} & \text{if } n \geq 2\ . \end{cases} \tag{5}$$

Note that from the recursive construction of $\mathscr{L}_n$ ($n \geq 2$), if an element $w \in \mathscr{L}_n$ starts with k, then k is followed by a symbol different from h, according to the definition of the language $\mathscr{L}$ in the previous section. If w starts with a symbol different from k, then it can be concatenated to any element of $\mathscr{L}_{n-1}$.

We propose the following definition providing a particular placement of the elements of $\mathscr{L}$ which reveals itself to be a Gray code.

Definition 4.1. Given $k \in \mathbb{N}$, $h \in \mathbb{Z}$, with $k \geq h \geq 0$, we define the string list $\mathcal{L}_n$ over the alphabet $\Sigma = \{0, 1, 2, \dots, k\}$:

$$\mathcal{L}_n = \begin{cases} \{\varepsilon\} & \text{if } n = 0 \\ \{0, 1, \dots, k-1\} & \text{if } n = 1 \\ \left(\bigcirc_{i=0}^{k-1}\ i \cdot (\mathcal{L}_{n-1})^{\underline{k+i}}\right) \circ \left(\bigcirc_{i=0}^{h-1}\ ki \cdot (\mathcal{L}_{n-2})^{\underline{k+i}}\right) & \text{if } n \geq 2\ . \end{cases} \tag{6}$$

We have:

Theorem 4.2. *The string list $\mathcal{L}_n$ is a Gray code with Hamming distance equal to one.*

Proof. We proceed by induction on n. If $n = 1$, then $\mathcal{L}_1$ is trivially seen to be a Gray code. Suppose that $\mathcal{L}_i$ is a Gray code for $i = 2, \ldots, n-1$ where $n \geq 2$ and consider the list $\mathcal{L}_n$. The sub-lists arising from each parenthesis in the third case of (6) are Gray codes since the lists $\mathcal{L}_{n-1}$ and $\mathcal{L}_{n-2}$ (which are Gray codes by the inductive hypothesis) are read alternatively from left to right and vice versa (depending on the parity of $k+1$). Therefore, we have to check only the Hamming distance between $last\left((k-1)\cdot(\mathcal{L}_{n-1})^{\underline{k+i}}\right)$ with $i = k-1$ and $first\left(k0\cdot(\mathcal{L}_{n-2})^{\underline{k+i}}\right)$ with $i = 0$.

In the case of k even we have:

$$last\left((k-1)\cdot(\mathcal{L}_{n-1})^{\underline{k+k-1}}\right) = last\left((k-1)\cdot\overline{\mathcal{L}_{n-1}}\right) = (k-1)last(\overline{\mathcal{L}_{n-1}}) = (k-1)first(\mathcal{L}_{n-1}) \ ,$$

and

$$first\left(k0\cdot(\mathcal{L}_{n-2})^{\underline{k+0}}\right) = first\left(k0\cdot\mathcal{L}_{n-2}\right) = k0first(\mathcal{L}_{n-2}) \ .$$

Easily, from (6), we have

$$first(\mathcal{L}_{n-1}) = 0first(\mathcal{L}_{n-2}) \ ,$$

then the only different digit between $(k-1)first(\mathcal{L}_{n-1})$ and $k0first(\mathcal{L}_{n-2})$ is the first one. So $\mathcal{L}_n$ is a Gray code.

The proof in the case of k odd can be conducted with similar arguments. □

5 Further developments

In the present paper we considered the recurrence relation defined by $a_m = ka_{m-1} + ha_{m-2}$ with $k \geq h \geq 0$, which is one of the two cases analysed in [BR01]. The other interesting case is $k \geq -h \geq 0$ which again leads to a strictly increasing sequence defining a language $\mathscr{L}'$ over an alphabet Σ' slightly different from Σ, with different properties. We think that also in this case it is possible to define a Gray code for listing the elements of $\mathscr{L}'$, with Hamming distance 1.

It could be interesting to investigate on the possibilities to characterize the language in terms of restricted words, following the line of [BBBP17b] or [Ber17] where sets of pattern avoiding words and recurrence relations are considered.

Acknowledgements

This work is partially supported by the GNCS 2018 project "Proprietà combinatorie e rilevamento di pattern in strutture discrete lineari e bidimensionali".

References

[Aig07] M. Aigner. *Discrete Mathematics.* American Mathematical Society, USA, 2007.

[BBBP17a] E. Barcucci, A. Bernini, S. Bilotta and R. Pinzani. Cross-bifix-free sets in two dimensions. *Theoretical Computer Science*, 664:29–38, 2017.

[BBBP17b] E. Barcucci, A. Bernini, S. Bilotta and R. Pinzani. Non-overlapping matrices. *Theoretical Computer Science*, 658:36–45, 2017.

[BBDD98] E. Barcucci, S. Brunetti, A. Del Lungo and F. Del Ristoro. A combinatorial interpretation of the sequence $f_{n+1} = 6f_n - f_{n-1}$. *Discrete Mathematics*, 190:235–240, 1998.

[BR01] E. Barcucci and S. Rinaldi. Some linear recurrences and their combinatorial interpretation by means of regular languages. *Theoretical Computer Science*, 255:679–686, 2001.

[Ber17] A. Bernini. Restricted binary strings and generalized Fibonacci numbers. *Lecture Notes in Computer Science*, 10248:32–43, 2017.

[BBPSV14] A. Bernini, S. Bilotta, R. Pinzani, A. Sabri and V. Vajnovszki. Prefix partitioned Gray code for particular cross-bifix-free sets. *Cryptography and Communication*, 6:359–369, 2014.

[BBPSV15] A. Bernini, S. Bilotta, R. Pinzani, A. Sabri and V. Vajnovszki. Gray code orders for q-ary words avoiding a given factor. *Acta Informatica*, 52:573–592, 2015.

[BBPV15] A. Bernini, S. Bilotta, R. Pinzani and V. Vajnovszki. A trace partitioned Gray code for q-ary generalized Fibonacci strings. *Journal of Discrete Mathematical Sciences and Cryptography*, 18:751–761, 2015.

[BBPV17] A. Bernini, S. Bilotta, R. Pinzani and V. Vajnovszki. A Gray code for cross-bifix-free sets. *Mathematical Structures in Computer Science*, 27:184–196, 2017.

[BGPP07] A. Bernini, E. Grazzini, E. Pergola and R. Pinzani. A general exhaustive generation algorithm for Gray structures. *Acta Informatica*, 44:361-376, 2007.

[BPP12] S. Bilotta, E. Pergola and R. Pinzani. A new approach to cross-bifix-free sets. *IEEE Transactions on Information Theory*, 58:4058–4063, 2012.

[BSS93] J. Bonin, L. Shapiro and R. Simion. Some q-analogues of the Schröder numbers arising from combinatorial statistics on lattice paths. *Journal of Statistical Planning and Inference*, 34:35–55, 1993.

[Fra82] A. S. Fraenkel. Systems of numeration. *American Mathematical Monthly*, 92:105–114, 1982.

[Gra53] F. Gray. *Pulse code communication.* US Patent 2(632), 058, 1953.

[Ham50] R. W. Hamming. Error detecting and error correcting codes. *The Bell System Technical Journal*, 29:147–160, 1950.

[PP90] E. Pergola and R. Pinzani. A combinatorial interpretation of the area of Schröder paths. *Electronic Journal of Combinatorics*, 6:R40, 1990.

[SP96] N. J. A. Sloane and S. Plouffe. *The Encyclopedia of Integer Sequences.* Academic Press, New York, 1996.

[Sul98] R. A. Sulanke. Bijective recurrences concerning Schröder paths. *Electronic Journal of Combinatorics*, 5:R47, 1998.