# Linear representation of categorical values

Arnaud Berny
research@courros.fr

## ABSTRACT

We propose a binary representation of categorical values using a linear map. This linear representation preserves the neighborhood structure of categorical values. In the context of evolutionary algorithms, it means that every categorical value can be reached in a single mutation. The linear representation is embedded into standard metaheuristics, applied to the problem of Sudoku puzzles, and compared to the more traditional direct binary encoding. It shows promising results in fixed-budget experiments and empirical cumulative distribution functions with high dimension instances, and also in fixed-target experiments with small dimension instances.

## CCS CONCEPTS

• **Computing methodologies** → **Randomized search**; **Discrete space search**; • **Mathematics of computing** → **Combinatorial optimization**; Coding theory.

## KEYWORDS

Combinatorial optimization, categorical values, binary representation, linear representation, Sudoku

**ACM Reference Format:**
Arnaud Berny. 2021. Linear representation of categorical values. In *2021 Genetic and Evolutionary Computation Conference Companion (GECCO '21 Companion), July 10–14, 2021, Lille, France.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3449726.3459513

## 1 INTRODUCTION

Representation is an important topic for evolutionary algorithms [4] and other metaheuristics, especially when applied to combinatorial optimization. It directly influences the range of problems which can be adressed by metaheuristics and the quality of their solutions. Many evolutionary algorithms have been designed with binary domains in mind. Although most of them can be adapted more or less easily to other domains, it is still desirable to be able to represent values from non binary domains in binary domains so as to leverage theoretical and practical knowledge of evolutionary algorithms in binary domains along with their implementations. In this paper, we are concerned with the binary representation of categorical values.

Often, categorical values are represented by means of direct binary encoding. As an example, let us address the problem of representing the four nucleobases $A$, $T$, $C$, and $G$ found in DNA. Using 2 bits, we can arbitrarily decide that $A = 00$, $T = 01$, $C = 10$, and $G = 11$. In the neighborhood system defined by 1-bit flips, it appears that each nucleobasis has 2 neighbors. In particular, it is not possible to go from $A$ to $G$ in a single bit flip. Direct binary encoding is inappropriate because resulting neighborhood systems among categorical values are not complete. Unary representation has the same limitation. In a set of categorical values, every element is the neighbor of every other element. In other words, categorical values are the vertices of a complete graph.

In this paper, we propose a binary representation of categorical values which is based on a linear map and which satisfies this requirement. The paper is organised as follows. In Sect. 2 we define binary representations for categorical values. In Sect. 3 we propose a linear representation for categorical values. In Sect. 4 we apply the linear representation to Sudoku puzzles. Sec. 5 concludes the paper.

## 2 REPRESENTATION

Let $V = \{v_1, v_2, \ldots, v_N\}$ be a set of $N \in \mathbb{N}$ categorical values and $n \in \mathbb{N}$ be the dimension of the binary domain used to represent them. A binary representation of $V$ is a surjective map $\phi : \{0,1\}^n \to V$, that is, for all $v \in V$, there exists $\boldsymbol{x} \in \{0,1\}^n$ such that $\phi(\boldsymbol{x}) = v$. The binary vector $\boldsymbol{x}$ is called a representative of $v$ which might have more than one representative. Such binary representations can be used, for example, to apply metaheuristics designed for binary spaces to the optimization of functions defined on categorical values.

Let $(\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n)$ be the canonical basis of $\{0,1\}^n$. For example, in $\{0,1\}^3$, $\boldsymbol{e}_1 = (1,0,0)^t$, where $t$ denotes transpose (we use column vectors). For all $\boldsymbol{x} \in \{0,1\}^n$, let $B(\boldsymbol{x}, 1)$ be the Hamming ball of radius 1 centered at $\boldsymbol{x}$, that is $B(\boldsymbol{x}, 1) = \{\boldsymbol{x}\} \cup \{\boldsymbol{x} + \boldsymbol{e}_i \mid i \in [1..n]\}$. Throughout this paper, we identify the set $\{0,1\}$ as the finite field $\mathbb{F}_2$. Thus, addition on $\{0,1\}$ or $\{0,1\}^n$ must be understood modulo 2 and is equivalent to the exclusive-or operator.

With mutation based metaheuristics or local search in mind, we would like to be able to reach any categorical value in a single bit mutation. We say that $\phi$ is locally bijective if, for all $\boldsymbol{x} \in \{0,1\}^n$, its restriction $\phi : B(\boldsymbol{x}, 1) \to V$ is bijective. In this case, necessarily, $n + 1 = N$.

## 3 LINEAR REPRESENTATION

We propose a linear representation wich is locally bijective. We suppose for now that $N = 2^k$, where $k \in \mathbb{N}$. The categorical values are first identified with $k$-bit binary vectors in an arbitrary manner. We are looking for a surjective linear representation, that is a $k \times n$ binary matrix of rank $k$. Let $\boldsymbol{x} \in \{0,1\}^n$ be the current search point and $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \in \{0,1\}^k$ its corresponding categorical value. The neighbors of $\boldsymbol{y}$ are $\boldsymbol{A}(\boldsymbol{x} + \boldsymbol{e}_i) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{A}\boldsymbol{e}_i = \boldsymbol{y} + \boldsymbol{A}\boldsymbol{e}_i$, where $i \in [1..n]$. Let $\boldsymbol{y'} \in \{0,1\}^k$ be any categorical value but $\boldsymbol{y}$. Then, $\boldsymbol{A}(\boldsymbol{x} + \boldsymbol{e}_i) = \boldsymbol{y'} \Leftrightarrow \boldsymbol{A}\boldsymbol{e}_i = \boldsymbol{y} + \boldsymbol{y'}$. The last equation has a unique solution if and only if the set $\{\boldsymbol{A}\boldsymbol{e}_i \mid i \in [1..n]\}$ is the set $\{0,1\}^k \setminus \{\boldsymbol{0}\}$


*GECCO '21 Companion, July 10–14, 2021, Lille, France*

ACM ISBN 978-1-4503-8351-6/21/07.
https://doi.org/10.1145/3449726.3459513

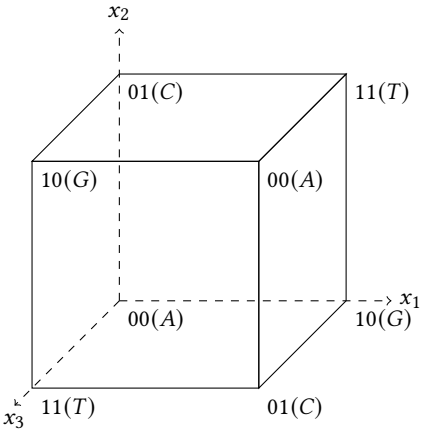


**Figure 1: Geometrical representation of a linear representation in the case of $N = 4$ categories, $k = 2$, and $n = 3$. Each vertex is labeled with a 2-bit string which, as a binary vector, is the image of its coordinates under the matrix $\boldsymbol{A}$. For illustration purpose, each 2-bit string is also arbitrarily identified as one of the four nucleobases found in DNA.**

and $n = N - 1 = 2^k - 1$, which means that the columns of $\boldsymbol{A}$ are made of all the vectors of $\{0,1\}^k$ but $\mathbf{0}$. We observe that $\boldsymbol{A}$ is precisely the parity-check matrix of the binary Hamming code [1]. It is remarkable that a requirement in the context of local search leads to a well known object of coding theory. We want to point out that we use $\boldsymbol{A}$ differently, though. Fig. 1 shows a geometric representation of $\boldsymbol{A}$ in the case $N = 4$:

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}.$$

Following [4], we can say that the linear representation has high locality; is uniformly redundant (each categorical value has exactly $2^{n-k}$ representatives); and is non synonymously redundant (for each categorical value, its representatives are spread all over the hypercube).

If $N$ is not a power of 2 then we let $k$ be the smallest natural such that $N < 2^k$ and repeat the construction of $\boldsymbol{A}$ with $n = 2^k - 1$. It is then necessary to map the output of $\boldsymbol{A}$ to $V$, for example with : $i \mapsto i \pmod N$. The resulting binary representation $\phi$ of $V$ is still locally surjective but not locally bijective.

## 4 EXPERIMENTS

We have applied standard metaheuristics designed for binary domains to Sudoku puzzles using linear representation and direct binary encoding. By counting the number of unsatisfied constraints, a Sudoku puzzle is turned into the minimization of a function $f : \{1, 2, ..., 9\}^d \to \mathbb{N}$, where $d$ is the number of unknowns. From the point of view of binary representations, Sudoku is a worst-case scenario since $N = 9$ is one past a power of 2. Direct representation requires $n = 4$ bits whereas linear representation requires $n = 15$ bits. Experiments[1] include the following metaheuristics: random local search, hill climbing, simulated annealing, GA, (1 + 1) EA, (10 + 1) EA, PBIL, MIMIC, UMDA, LTGA, and P3.

[1] All experiments have been produced with the HNCO framework [3].

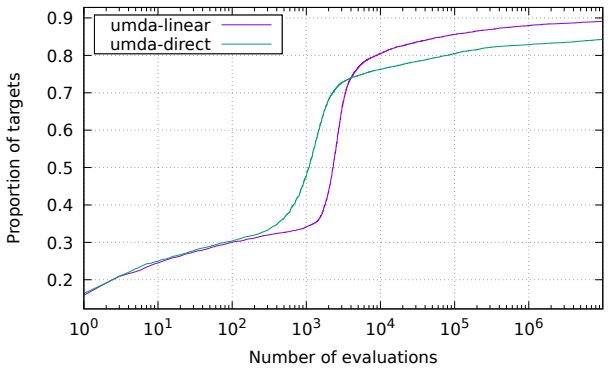


**Figure 2: ECDF's of UMDA with direct and linear representations (20 runs).**

To account for the dynamical behavior of metaheuristics, we have studied their empirical cumulative distribution functions (ECDF) [2] when applied to a fixed set of Sudoku instances of varying difficulty. Fig. 2 shows ECDF's of UMDA with direct and linear representations. Linear representation has shown a clear advantage over direct representation with all metaheuristics but MIMIC, LTGA, and P3. Only in the case of P3 has direct representation overtaken linear representation by a significant margin within the considered budget.

We have also studied the runtime of metaheuristics in fixed-target experiments. We have generated easy Sudoku instances, starting from complete boards and erasing a small number $r \in [1..10]$ of digits. For each dimension, 4 instances have been generated. Linear representation has surpassed direct representation with all metaheuristics but GA and MIMIC.

## 5 CONCLUSION

We have proposed a linear representation for categorical values in binary domains. Every value can be reached with a single mutation. This requirement has, in turn, lead to an unexpected connexion with coding theory. One drawback of linear representation is its size, which is linear in the number of categorical values but exponential in the size of direct binary encoding. This could explain some of its negative experimental results. Its advantage over direct binary encoding on Sudoku puzzles has to be confirmed in the context of other problems, preferably real-world ones. The influence of the number of categories and the number of categorical variables on the performance of metaheuristic-representation pairs are of particular interest.

## REFERENCES

[1] R. W. Hamming. Error detecting and error correcting codes. *The Bell System Technical Journal*, 29(2):147–160, 1950.
[2] Nikolaus Hansen, Anne Auger, Dimo Brockhoff, Dejan Tusar, and Tea Tusar. COCO: performance assessment. *CoRR*, abs/1605.03560, 2016.
[3] HNCO. https://github.com/courros/hnco. v0.16.
[4] Franz Rothlauf. *Representations for genetic and evolutionary algorithms.* Springer, 2006.