Topological methods for symbolic discrepancy

Valérie Berthé

In this lecture, we discuss the notion of bounded symbolic discrepancy for infinite words and subshifts, both for letters and factors, from a topological dynamics viewpoint. We focus on three families of words, namely hypercubic words, words generated by substitutions, and dendric words. Symbolic discrepancy measures the difference between the numbers of occurrences of a given word v in some word of length n minus n times the frequency μ_v of v when it exists (in other words, μ_v is the measure of the cylinder $[v]$ for some invariant measure μ). Bounded discrepancy thus provides particularly strong convergence properties of ergodic sums toward frequencies. More precisely, let $u \in \mathscr{A}^{\mathbb{Z}}$ be a bi-infinite word and assume that each factor v in its language admits a frequency μ_v in u. The *discrepancy* $\Delta_v(u)$ of u with respect to v is defined as

$$\Delta_v(u) = \sup_{n\in\mathbb{N}} \big| |u_{-n} \cdots u_0 \cdots u_n|_v - (2n+1)\mu_v \big|.$$

This notion extends to any minimal subshift (X,T) in a straightforward way. If $\Delta_v(u)$ is finite, the cylinder $[v]$ is said to be a bounded remainder set, according to the terminology developed in classical discrepancy theory. Bounded discrepancy is closely related to the notion of balance in word combinatorics.

To illustrate the relevance of the topological approach for symbolic discrepancy, let us start with a first classical remark. Let (X,T) be a minimal and uniquely ergodic subshift and let μ stand for its invariant measure. Given a factor v in its language, define $f_v = \chi_{[v]} - \mu([v]) \in C(X,\mathbb{R})$, where $\chi_{[v]}$ stands for the characteristic function of the cylinder $[v]$. Then, according to the Gottschalk–Hedlund theorem, v has bounded discrepancy in (X,T) if and only if the map f_v is a coboundary. Bounded discrepancy thus implies that $\mu([v])$ is an additive topological eigenvalue of (X,T).

Here, we deduce that for hypercubic words produced by d-to-1 cut-and-project schemes (with irrationality assumptions that yield minimality), letters have bounded

---

Valérie Berthé
IRIF, CNRS UMR 8243, Université Paris Diderot – Paris 7, France. e-mail: berthe@irif.fr

discrepancy, whereas factors of length at least 2 do not have bounded discrepancy for $d \geq 3$. Indeed, frequencies of factors of length at least 2 do not belong to the group of additive eigenvalues.

In the substitutive case, we stress the role played by the existence of coboundaries taking rational values and show simple criteria when frequencies take rational values for exhibiting unbounded discrepancy. For more precise results, see [1].

The third family we consider here is the family of dendric words, and we present results from [3]. Given a subshift over a finite alphabet, one can associate with every word in the associated language a bipartite graph, called extension graph, in which one puts edges between left and right letter extensions of this factor in the language. If, for every word in this language, the extension graph is a tree, then the subshift is a dendric subshift. Dendric subshifts are therefore defined in terms of combinatorial properties of their language. This class of linear factor complexity subshifts encompasses Sturmian subshifts, Arnoux–Rauzy subshifts, as well as subshifts generated by regular interval exchanges. We study the dimension group of dendric subshifts, providing necessary and sufficient conditions for two dendric subshifts to be (strongly) orbit equivalent. More precisely, let (X,T) be a minimal dendric subshift on the alphabet $\mathscr{A} = \{1,\cdots,d\}$ and let $\mathscr{M}(X,T)$ stand for its set of invariant measures. Then, its dimension group with ordered unit is isomorphic to

$$\Big(\mathbb{Z}^d, \big\{\boldsymbol{x} \in \mathbb{Z}^d \mid \langle \boldsymbol{x}, \boldsymbol{\mu} \rangle > 0 \text{ for all } \mu \in \mathscr{M}(X,T)\big\} \cup \big\{0\big\}, \vec{1}\Big)$$

where $\boldsymbol{\mu}$ denotes the vector $\big(\mu([1]),\cdots,\mu([d])\big)$. We deduce that, as soon as dendric words are balanced on letters, they are balanced on factors. The proof relies on the following property of dendric subshift from [3]: let X be a minimal dendric subshift defined on the alphabet $\mathscr{A}$. Then, for any w in its language, the set of left return words to w is a basis of the free group over $\mathscr{A}$.

References

1. Berthé, V., Cecchi Bernales, P.: Balances and coboundaries in symbolic systems. Theor. Comput. Sci. **777**, 93–110 (2019). arXiv:1810.07453
2. Berthé, V., Cecchi Bernales, P., Durand, F., Leroy, J., Perrin, D., Petite, S.: Dimension groups of dendric subshifts, preprint arXiv:1911.07700
3. Berthé, V., De Felice, C., Dolce, F., Leroy, J., Perrin, D., Reutenauer, C., Rindone, G.: Acyclic, connected and tree sets. Monatsh. Math. **176**, 521–550 (2015)