MUTUAL INFORMATION DECAY FOR FACTORS OF IID

BALÁZS GERENCSÉR AND VIKTOR HARANGI

Abstract. This paper is concerned with factor of i.i.d. processes on the d-regular tree for $d > 2$. We study the mutual information of the values on two given vertices. If the vertices are neighbors (i.e., their distance is 1), then a known inequality between the entropy of a vertex and the entropy of an edge provides an upper bound for the (normalized) mutual information. In this paper we obtain upper bounds for vertices at an arbitrary distance k, of order $(d-1)^{-k/2}$. Although these bounds are sharp, we also show that an interesting phenomenon occurs here: for any fixed process the rate of decay of the mutual information is much faster, essentially of order $(d-1)^{-k}$.

1. Introduction

For an integer $d > 2$ let T_d denote the d-regular tree: the (infinite) connected graph with no cycles and with each vertex having exactly d neighbors.

This paper deals with *factor of i.i.d. processes* on T_d. Loosely speaking, independent and identically distributed (say $[0,1]$ uniform) random labels are assigned to the vertices of T_d, then each vertex gets a new label that depends on the labeled rooted graph as seen from that vertex, all vertices "using the same rule".

For a formal definition, let $V(T_d)$ denote the vertex set and $\mathrm{Aut}(T_d)$ the automorphism group of T_d. Suppose that M is a measurable space. (In most cases M will be either a discrete set or $\mathbb{R}$.) A measurable function $F\colon [0,1]^{V(T_d)} \to M^{V(T_d)}$ is said to be an $\mathrm{Aut}(T_d)$-*factor* (or *factor* in short) if it is $\mathrm{Aut}(T_d)$-equivariant, that is, it commutes with the natural $\mathrm{Aut}(T_d)$-actions. Given an i.i.d. process $Z = (Z_v)_{v \in V(T_d)}$ on $[0,1]^{V(T_d)}$, applying F yields a factor of i.i.d. process $X = F(Z)$, which can be viewed as a collection $X = (X_v)_{v \in V(T_d)}$ of M-valued random variables. It follows immediately from the definition that the distribution of X is invariant under the action of $\mathrm{Aut}(T_d)$; in particular, each X_v has the same distribution.

One of the reasons why factor of i.i.d. processes have attracted a growing attention in recent years is that they give rise to some sort of randomized local algorithms that can be carried out on arbitrary regular graphs with "large essential girth", e.g. random regular graphs. (See [9, 10, 11, 12] how factors of i.i.d./local algorithms can be used to obtain large independent sets on large-girth graphs.) Factors of i.i.d. are also studied by ergodic theory (under the name of *factors of Bernoulli shifts*), see Section 2 for details.

The starting point of our investigations is the following entropy inequality which holds for any factor of i.i.d. process with a finite state space M:

$$H(\updownarrow) \geq \frac{2(d-1)}{d} H(\bullet). \tag{1}$$

2010 *Mathematics Subject Classification.* 37A35, 60K35, 37A50.

Key words and phrases. factor of IID, factor of Bernoulli shift, mutual information, entropy inequality.

The first author was supported by NKFIH (National Research, Development and Innovation Office) grant PD 121107. The second author was supported by Marie Skłodowska-Curie Individual Fellowship grant no. 661025 and the MTA Rényi Institute "Lendület" Groups and Graphs Research Group.

Here • represents a vertex, and $H(\bullet)$ is the (Shannon) entropy of the discrete random variable X_v for any vertex v. Similarly, $\updownarrow$ represents an edge, and $H(\updownarrow)$ stands for the entropy of the joint distribution (X_u, X_v) for any edge uv. (Because of the $\mathrm{Aut}(T_d)$-invariance the distribution of every edge/vertex is the same.) Entropy inequalities played a central role in a couple of remarkable results recently: the Rahman-Virág result [15] about the maximal size of a factor of i.i.d. independent set on T_d and the Backhausz-Szegedy result [3] on eigenvectors of random regular graphs.

The inequality (1) can also be expressed as an upper bound for the mutual information of two neighboring vertices u and v:

$$\frac{I(X_u, X_v)}{H(X_v)} \le \frac{2}{d}. \tag{2}$$

Recall that the mutual information $I(X_u, X_v)$ is defined as $H(X_u) + H(X_v) - H(X_u, X_v)$ and can be viewed as (the expected value of) the information gained about one of the random variables knowing the other one. In our case the random variables are identically distributed, therefore they have the same entropy $H(X_u) = H(X_v)$. Dividing the mutual information by this entropy gives some sort of *normalized mutual information* which measures the amount of shared information proportional to the total amount of information. This ratio is always between 0 and 1, and being close to 0 intuitively means that the random variables are "almost independent". (It is actually natural to normalize the mutual information this way: if we take a tuple of independent copies of a factor of i.i.d. process, which is also a factor of i.i.d., then both the entropy of a vertex and the mutual information get multiplied by the same number.)

A natural question arises: what can be said about the mutual information of two vertices u and v at distance k? One expects that the mutual information tends to 0 as the distance grows. But what is the rate of decay? We get very different answers depending on how the question is posed exactly.

First let us consider the problem for a fixed $k \ge 1$, that is, we look for a "universal" upper bound for the normalized mutual information $I(X_u, X_v)/H(X_v)$ that holds for any factor of i.i.d. process with a finite state space M. The following bounds are obtained.

Theorem 1. *Let $d > 2$ be an even integer and M a finite state space. For any $u, v \in V(T_d)$ at distance k and for any factor of i.i.d. process X on $M^{V(T_d)}$ we have*

$$\frac{I(X_u, X_v)}{H(X_v)} \le \begin{cases} \frac{2}{d(d-1)^l} & \text{if } k = 2l+1 \text{ is odd,} \\ \frac{1}{(d-1)^l} & \text{if } k = 2l \text{ is even,} \end{cases} \tag{3}$$

These bounds are the best possible in the sense that for any fixed k there exist factor of i.i.d. processes for which the normalized mutual information tends to the bound above.

The assumption of d being even is only technical. The statement is true for odd d as well but for the sake of simplicity we only prove it for even d in this paper; see Remark 3.2.

According to the above theorem, the normalized mutual information for distance k is (at most) of order $(1/\sqrt{d-1})^k$, and this is sharp. However, it turns out that there does not exist a single factor of i.i.d. process that would show the sharpness of the bound for all k at once. In fact, for any fixed process the mutual information decays at a much faster rate, basically of order $1/(d-1)^k$.

Theorem 2. *Let $d > 2$ be any integer and M a finite state space. If $X = (X_v)_{v \in V(T_d)}$ is a factor of i.i.d. process on $M^{V(T_d)}$, then*

$$I(X_u, X_v) \leq \frac{|M|(k+1)^2}{(d-1)^k}. \tag{4}$$

where $|M|$ denotes the cardinality of M (number of states).

This bound is essentially sharp, see Example 5.4.

Motivation. Our motivation to study this problem is multi-fold. On the one hand, many aspects of independence in factors of i.i.d. have been studied earlier (e.g. correlation for real-valued processes or triviality of various tail σ-algebras). Our goal was to get a quantitative result about how much independence these processes exhibit when M is finite. Mutual information has the advantage over correlation that the latter only detects linear dependence. On the other hand, we aimed to obtain new entropy inequalities. The edge-vertex inequality (and its "blow-ups") have a number of applications already. Theorem 1 is a generalization of this inequality, and as such we hope it proves to be a useful tool.

Proof methods. For even d the d-regular tree T_d can be considered as the Cayley graph of the free group F_r of rank $r = d/2$. The free group F_r acts on its Cayley graph T_d via automorphisms. Loosely speaking, F_r is a subgroup of $\mathrm{Aut}(T_d)$. There is a version of the edge-vertex entropy inequality (1) for F_r-factors of i.i.d. processes (see Section 2 for details). This (more general) version follows from the work of Lewis Bowen on the *f-invariant* [7]. The idea is to use this version for certain subgroups of F_r. To prove Theorem 1 for a given k we will need to find a subgroup (of maximal rank) with the property that it can be generated freely by elements of length k.

Theorem 2 will be deduced from the correlation decay result of Backhausz, Szegedy and Virág [4], which says that for a real-valued factor of i.i.d. process ($M = \mathbb{R}$) the correlation of two vertices u and v at distance k is (at most) of order $1/(\sqrt{d-1})^k$. In the case of a finite state space M, by assigning a real number to each state we can replace our original process with a real-valued one. Consequently, for any assignment $M \to \mathbb{R}$ the correlation bound tells us something about the joint distribution of X_u and X_v (for the original process). The idea is to try to find suitable assignments that yield a good bound on the mutual information of X_u and X_v.

Outline of the paper. The rest of the paper is structured as follows. In Section 2 we go through basic definitions and explain the more general entropy inequality we will need to prove the universal bound. The proofs of Theorem 1 and 2 are given in Section 3 and 4, respectively. Finally, in Section 5 we present examples showing that the above theorems are (essentially) sharp.

Acknowledgments. We are grateful to Ágnes Backhausz, Balázs Szegedy, Bálint Virág and Máté Vizer for fruitful discussions on the topic.

2. Preliminaries

2.1. **Factors of i.i.d.** Although the results of this paper concern $\mathrm{Aut}(T_d)$-factors, we will need to use the notion of factors in a more general setting. Suppose that a group Γ acts on a countable set S. Then Γ also acts on the space M^S for a set M: for any function $f : S \to M$ and for any $\gamma \in \Gamma$ let

$$(\gamma \cdot f)(s) := f(\gamma^{-1} \cdot s) \quad \forall s \in S. \tag{5}$$

First we define the notion of factor maps.

Definition 2.1. Let M_1, M_2 be measurable spaces and S_1, S_2 countable sets with a group Γ acting on both. A measurable mapping $F\colon M_1^{S_1} \to M_2^{S_2}$ is said to be a Γ-*factor* if it is Γ-equivariant, that is, it commutes with the Γ-actions.

By an *invariant process* on M^S we mean an M^S-valued random variable (or a collection of M-valued random variables) whose (joint) distribution is invariant under the Γ-action. For example, if Z_s, $s \in S_1$, are independent and identically distributed M_1-valued random variables, then we say that $Z = (Z_s)_{s\in S_1}$ is an i.i.d. process on $M_1^{S_1}$. Given a Γ-factor $F\colon M_1^{S_1} \to M_2^{S_2}$, $X \coloneqq F(Z)$ is a factor of the i.i.d. process Z. It can be regarded as a collection of M_2-valued random variables: $X = (X_s)_{s\in S_2}$.

In fact, all this can be viewed in the context of ergodic theory. An invariant process in the above sense gives rise to a dynamical system over Γ: the group Γ acts by measure-preserving transformations on the measurable space M^S equipped with a probability measure (the distribution of the invariant process). An i.i.d. process simply corresponds to a (generalized) Bernoulli shift. Therefore factor of i.i.d. processes are essentially factors of Bernoulli shifts. Classical ergodic theory ($\mathbb{Z}$-factors) have the largest literature and the most complete theory but Γ-factors have also been thoroughly investigated for general Γ.

For amenable group actions (the Kolmogorov-Sinai) entropy serves as a complete invariant (for isomorphism of Bernoulli shifts). As for the nonamenable case, Ornstein and Weiss asked whether all Bernoulli shifts are isomorphic over a nonamenable group [13]. This remained open until the breakthrough results of Lewis Bowen: he answered the question negatively by introducing the f-*invariant* for free group actions [6] and the Σ-*entropy* for actions of sofic groups [8]. In another paper he showed that the f-invariant is essentially a special case of the Σ-entropy which has the consequence that the f-invariant is non-negative for factors of the Bernoulli shift [7, Corollary 1.8]. We will need this fact in the form of an entropy inequality, see (6) below.

2.2. **Factors on T_d.** The main results of this paper (Theorem 1 and 2) are concerned with factor of i.i.d. processes on T_d. This corresponds to the case when Γ is the automorphism group $\mathrm{Aut}(T_d)$ of the d-regular infinite tree T_d and S is the vertex set $V(T_d)$.

When we say *factor of i.i.d. process*, we should also specify which i.i.d. process we have in mind (that is, specify M_1 and a probability distribution on it). By default we will work with the uniform $[0,1]$ measure (i.e., the Lebesgue measure on $[0,1]$). In fact, as far as the class of factor processes is concerned, it does not really matter which i.i.d. process we consider. For example, for $\{0,1\}$ with the uniform distribution we get the same class of factors as for the uniform $[0,1]$ measure. This follows from the fact that these two i.i.d. processes are $\mathrm{Aut}(T_d)$-factors of each other [5].

Note that a factor of i.i.d. process X on T_d is $\mathrm{Aut}(T_d)$-invariant. Therefore each X_v has the same distribution. Moreover, the joint distribution of X_u and X_v (and hence their correlation or mutual information) depends only on the distance between u and v.

As we mentioned in the introduction, factor of i.i.d. processes satisfy the edge-vertex entropy inequality (1). As we will see later, this inequality can be found implicitly in Lewis Bowen's work from 2009. Rahman and Virág proved it in a special setting for their result on factor of i.i.d. independent sets [15]. A full and concise proof was given by Backhausz and Szegedy in [2]; see also [14].

The proof in [2] actually works for a broader class of $\mathrm{Aut}(T_d)$-invariant processes that the authors coined *typical processes*. Loosely speaking, typical processes arise as limits of labelings of random d-regular graphs. Their significance lies in the fact that many questions about random regular graphs can be studied through typical processes.

2.3. **F_r-factors.** The other case that will be of particular interest for us is when $\Gamma = F_r$. We can set $S = \Gamma = F_r$ and consider the natural action of F_r on itself. Similarly as for $\mathrm{Aut}(T_d)$-factors, we use the uniform $[0, 1]$ measure for the i.i.d. process. Using other measures would result in the same class of factor processes.

This is actually a broader class than the class of $\mathrm{Aut}(T_d)$-factors (for $d = 2r$). If $d = 2r$, we can think of T_d as the Cayley graph of F_r with respect to a symmetric generating set $\{a_1^{\pm 1}, \ldots, a_r^{\pm 1}\}$. That is, $V(T_d) = F_r$ and a vertex g is incident to vertices of the form $ga_i^{\pm 1}$. Then F_r acts on $V(T_d) = F_r$ (from the left) via automorphisms of this Cayley graph. So, loosely speaking, $F_r \leq \mathrm{Aut}(T_d)$, and being $\mathrm{Aut}(T_d)$-equivariant is a stronger condition than being F_r-equivariant. In other words, every $\mathrm{Aut}(T_d)$-factor is an F_r-factor as well.

For a general F_r-factor of i.i.d. we only have F_r-invariance (but not necessarily $\mathrm{Aut}(T_d)$-invariance). It is still true that each X_g has the same distribution. As for the distribution of edges, however, $(X_g, X_{ga_i^{\pm 1}})$ might have different distributions for different $a_i^{\pm 1}$.

The following entropy inequality, which plays a central role in our proof of Theorem 1, easily follows from the fact that the f-invariant of a factor of a Bernoulli shift is non-negative [7].

Theorem 2.2. *Let $\Gamma = \langle a_1, \ldots, a_r \rangle$ be a free group of rank $r \geq 2$. If $X = (X_g)_{g \in \Gamma}$ is a Γ-factor of the i.i.d. process on $[0, 1]^\Gamma$, then for a fixed $g \in \Gamma$ we have*

$$\frac{1}{r} \sum_{i=1}^{r} H(X_g, X_{ga_i}) \geq \frac{2r-1}{r} H(X_g), \tag{6}$$

or equivalently:

$$\frac{1}{r} \sum_{i=1}^{r} \frac{I(X_g, X_{ga_i})}{H(X_g)} \leq \frac{1}{r}. \tag{7}$$

Remark 2.3. This is more general than the edge-vertex entropy inequality (1) for $\mathrm{Aut}(T_d)$-factors. Indeed, given an $\mathrm{Aut}(T_d)$-factor, it is also an F_r-factor, but with the extra property that the distributions of edges are the same.

3. The universal bound

Our goal is to apply the entropy inequality (6–7) for subgroups of F_r (which are themselves free groups). To obtain a result about vertices at distance k in T_d we need a subgroup that is generated by elements of length k. The higher the rank of our subgroup, the better inequality we get. Therefore we need to find as many elements of length k as possible such that they freely generate a subgroup. (When we have the maximal possible number of elements, the generated subgroup will have finite index.)

Lemma 3.1. *Let F_r be the free group of rank r. Fix r generating elements $a_1, \ldots, a_r$ and the corresponding word metric. Then for any positive integer k there exists a subgroup $H \leq F_r$ with the following properties:*

- *H is generated freely by elements of length k (in the word metric);*
- *the rank of H is $(2r-1)^l$ for even $k = 2l$ and $r(2r-1)^l$ for odd $k = 2l+1$;*
- *H has finite index.*

Before we prove this lemma, let us show how Theorem 1 follows.

Proof of Theorem 1. Let M be a finite set and F be any $\mathrm{Aut}(T_d)$-factor $[0,1]^{V(T_d)} \to M^{V(T_d)}$. For $d = 2r$ we can think of T_d as the Cayley graph of F_r with respect to a symmetric generating set $\{a_1^{\pm 1}, \ldots, a_r^{\pm 1}\}$. That is, $V(T_d) = F_r$ and a vertex g is incident to vertices of the form $ga_i^{\pm 1}$.

Now let H be a subgroup of F_r, and π_H denote the projection $M^{F_r} \to M^H$. We have the following situation:

$$[0,1]^H \longrightarrow [0,1]^{F_r} \xrightarrow{F} M^{F_r} \xrightarrow{\pi_H} M^H, \tag{8}$$

where the first mapping is as in the next claim.

Claim. *Let us equip the spaces $[0,1]^H$ and $[0,1]^{F_r}$ with the product of uniform $[0,1]$ measures. Then there exists a $[0,1]^H \to [0,1]^{F_r}$ mapping that is measure-preserving and H-equivariant.*

Proof. Fix a set T that contains exactly one element of each right H-coset. Let us start with independent uniform $[0,1]$ labels associated to the elements of H. For any $h \in H$ the label of h can be "decomposed" into countably many independent random bits (uniform 0-1). We can partition these bits into $|T|$ classes, each class containing countably many bits. (Note that this can be done even if $|T|$ is (countably) infinite so H does not need to have finite index.) This way each vertex ht, $t \in T$ inherits countably many bits from h. These bits can be "pieced together" into a real number between 0 and 1. Hence we get independent uniform $[0,1]$ labels for all elements in F_r.

The above procedure describes a measure-preserving $[0,1]^H \to [0,1]^{F_r}$ mapping. This mapping will commute with the H-action provided that we use the same partitioning of the bits for each h. □

It follows that all three mappings in (8) are H-equivariant meaning that their composition is actually an H-factor mapping.

Now let $X = (X_v)_{v \in V(T_d)}$ be a factor of i.i.d. process on T_d. In light of the above, if we "restrict" X to the vertices corresponding to the subgroup H, then we get an H-factor of i.i.d. process: $(X_h)_{h \in H}$. Now let k be a positive integer and choose H as in Lemma 3.1. The rank r' of H is $(2r-1)^l$ for even $k = 2l$ and $r(2r-1)^l$ for odd $k = 2l+1$. The lemma says that H has a free generating set containing elements of length k (w.r.t. the word metric of F_r). Let us apply (7) to H and this generating set. Then for any $h \in H$ and any generator s, the vertices h and hs have distance k (in the graph metric of T_d). Then, because of the $\mathrm{Aut}(T_d)$-invariance of X, the normalized mutual information $I(X_h, X_{hs})/H(X_h)$ is the same for all h and s. Therefore in our case the average on the left-hand side of (7) is simply equal to $I(X_u, X_v)/H(X_v)$ for any $u, v \in V(T_d)$ with $\mathrm{dist}(u,v) = k$, and Theorem 1 follows. (The sharpness will be shown in Section 5.) □

Remark 3.2. To obtain Theorem 1 for an arbitrary integer $d > 2$ one could prove a version of the entropy inequality (6) for the free product $\underbrace{\mathbb{Z}_2 * \cdots * \mathbb{Z}_2}_{d}$ instead of $F_r = \underbrace{\mathbb{Z} * \cdots * \mathbb{Z}}_{r}$.

Finally, we prove Lemma 3.1.

Proof of Lemma 3.1. Let A denote the set $\{a_1^{\pm 1}, \ldots, a_r^{\pm 1}\}$. We will refer to the elements of A as letters. Every element of F_r can be written as a product of these letters. If we perform all possible cancellations in such a "word", then we end up with the *reduced form* of the word. The norm of a group element w.r.t. the word metric is defined as the length of this reduced form.

We start with the odd case $k = 2l + 1$. A word is called a *palindrome* if it reads the same backward as forward. Let us consider the following set of generators:

$$S := \{s \in F_r : \text{the reduced form of } s \text{ is a palindrome and has length } 2l + 1\}.$$

That is, elements of S are in the form $b_1 \cdots b_l b_{l+1} b_l \cdots b_1$, where $b_i \in A$ and $b_{i+1} \neq b_i^{-1}$. The number of such elements is clearly $2r(2r-1)^l$.

Note that $S = S^{-1}$. Therefore there exists $S_0 \subset S$ such that $S = S_0 \cup S_0^{-1}$ and $|S_0| = |S|/2 = r(2r-1)^l$. We will see that S_0 is a free generating set of a subgroup $H \leq F_r$ that has all the required properties.

The following is obvious by induction.

Claim. *Let $s_1, \ldots, s_n \in S$ such that $s_{i+1} \neq s_i^{-1}$ for each i. Then the reduced form of the product $s_1 \cdots s_n$ has length at least $2l + n$ and its last $l + 1$ letters are the same as those of s_n.*

In particular, the product $s_1 \cdots s_n$ cannot be the unit element of F_r. Therefore S_0 freely generates some subgroup $H \leq F_r$, the rank of which is, obviously, $|S_0| = r(2r-1)^l$.

In fact, H has finite index. (We do not need this property in this paper.) This follows from the following observation. Let $T \subset F_r$ denote the set of elements of length at most l. Then it is easy to see that every element of F_r can be (uniquely) written in the form $s_1 \cdots s_n t$, where $t \in T$, $s_i \in S$ and $s_{i+1} \neq s_i^{-1}$.

The even case $k = 2l$ is slightly more complicated. We give a sketch of the proof. It is an easy exercise that for $l = 1$ the set

$$B := \left\{a_1^2\right\} \cup \{a_1 a_i : 2 \leq i \leq r\} \cup \{a_i a_1 : 2 \leq i \leq r\}$$

is a free generating set (of size $2r - 1$) of the subgroup consisting of all elements of even length. This subgroup has index 2.

For $l \geq 2$ we will need to "nest" the $2r - 1$ elements of B in palindrome-like words of length $2l$. First we define the mappings $\varphi_j \colon A \to A$: for $j \in \{0, 1, \ldots, r-1\}$ let φ_j shift the indices by j, that is, $\varphi(a_i^{\pm 1}) := a_{i+j}^{\pm 1}$. (The addition in the index is meant modulo r.) We will consider words of the following form:

$$b_1 b_2 \cdots b_l \varphi_j(b_l) \cdots \varphi_j(b_2) \varphi_j(b_1), \text{ where } b_i \in A; b_{i+1} \neq b_i^{-1}; j \in \{0, 1, \ldots, r-1\}.$$

All these words have length $2l$ but we will need only those for which at least one of the two letters in the middle is a_1 (i.e., $b_l \varphi_j(b_l) \in B$). To compute the number of such words first we choose the two letters in the middle ($|B| = 2r - 1$ possibilities), then we choose $b_{l-1}, \ldots, b_2, b_1$ one by one ($2r - 1$ possibilities for each). In total, $(2r-1)^l$ possibilities. We claim that the set S_0 of these $(2r-1)^l$ words of length $2l$ freely generates a finite-index subgroup.

The rigorous proof requires a straightforward but somewhat meticulous analysis that we omit here. The point is that given a product of elements from $S := S_0 \cup S_0^{-1}$, the length of the reduced form never decreases if an element is added to the end of the product. □

4. The rate of decay for a fixed process

We will need three ingredients to prove Theorem 2. The first one is a bound for the correlation of a pair of vertices for factor of i.i.d. processes on $\mathbb{R}^{V(T_d)}$, which was proved by Backhausz, Szegedy and Virág in [4]:

$$|\mathrm{corr}(X_u, X_v)| \leq \left(k + 1 - \frac{2k}{d}\right) \left(\frac{1}{\sqrt{d-1}}\right)^k, \text{ where } k = \mathrm{dist}(u, v), \tag{9}$$

that is, the rate of the correlation decay is essentially $1/(\sqrt{d-1})^k$. (Here it is assumed that $\operatorname{var} X_v < \infty$.)

Now suppose we have a finite state space M and a factor of i.i.d. process on $M^{V(T_d)}$. How can we make use of the above result in this case? Taking any function $f\colon M \to \mathbb{R}$ we can replace each X_v with $f(X_v)$ to get a factor of i.i.d. on $\mathbb{R}^{V(T_d)}$ so that (9) can be applied. The second ingredient is the next lemma from [1] which tells us that the same bound holds if we take different real-valued functions of X_u and X_v.

Lemma 4.1. *Let $(A, \mathcal{F})$ be an arbitrary measurable space. Suppose that the $(A, \mathcal{F})$-valued random variables X_1, X_2 are exchangeable (that is, (X_1, X_2) and (X_2, X_1) have the same joint distribution), and that there exists a constant $\alpha \geq 0$ with the property that for any measurable $f\colon A \to \mathbb{R}$ we have*

$$\left|\operatorname{corr}\big(f(X_1), f(X_2)\big)\right| \leq \alpha \text{ provided that } f(X_1) \text{ has finite variance.} \tag{10}$$

Then for any measurable functions $f_1, f_2\colon A \to \mathbb{R}$

$$\left|\operatorname{corr}\big(f_1(X_1), f_2(X_2)\big)\right| \leq \alpha \text{ provided that } f_1(X_1) \text{ and } f_2(X_2) \text{ have finite variances.} \tag{11}$$

Proof. The complete proof can be found in [1, Lemma 3.2] but we include a sketch here for the sake of completeness. After rescaling we might assume that $\operatorname{var}(f_1(X_1)) = \operatorname{var}(f_2(X_2)) = 1$. If we apply (10) to the function $f = f_1 + f_2$ and also to $f = f_1 - f_2$, we reach (11) after a short and simple calculation. Note that the exchangeability of X_1 and X_2 implies $\operatorname{cov}(f_1(X_1), f_2(X_2)) = \operatorname{cov}(f_1(X_2), f_2(X_1))$. □

The final ingredient is the following lemma linking correlation to mutual information.

Lemma 4.2. *Let X, Y be discrete random variables. Suppose that there exists a real number $\alpha \geq 0$ such that for any (real-valued) functions $f(X)$ and $g(Y)$ of X and Y it holds that $\left|\operatorname{corr}\big(f(X), g(Y)\big)\right| \leq \alpha$. Then we have*

$$I(X, Y) = H(X) - H(X|Y) \leq (m-1)\alpha^2,$$

where m denotes the number of values X can take.

Note that in some cases the conditional entropy $H(X|Y)$ can be defined even when Y is not discrete. The proof below works in those cases as well.

Proof. Let A be an event that depends on X, that is, $\mathbb{1}_A = f(X)$ for some function f. We denote the probability $\mathbb{P}(A)$ by p and we set

$$g_A(y) := \mathbb{P}(A|Y = y) - \mathbb{P}(A) = \mathbb{P}(A|Y = y) - p.$$

Clearly, $\mathbb{E} g_A(Y) = 0$, and it is also easy to see that

$$\operatorname{corr}\big(f(X), g_A(Y)\big) = \frac{\sqrt{\mathbb{E} g_A(Y)^2}}{\sqrt{p(1-p)}}.$$

It follows that

$$\mathbb{E} g_A(Y)^2 \leq \alpha^2 p(1-p). \tag{12}$$

Now let us assume that X takes the value x_i with probability p_i for $1 \leq i \leq m$. We will need to use the above inequality for each event $A_i = \mathbb{1}_{\{X = x_i\}}$, $1 \leq i \leq m$. We write g_i for the corresponding function g_{A_i}.

The mutual information of X conditioned on Y can be expressed as follows:

$$-H(X|Y) = \mathbb{E}\sum_{i=1}^{m}(p_i + g_i(Y)) \underbrace{\log(p_i + g_i(Y))}_{\log(p_i)+\log\left(1+\frac{g_i(Y)}{p_i}\right)} .$$

Now by using the inequality $\log(1+x) \le x$ we get that

$$-H(X|Y) \le \underbrace{\sum_{i=1}^{m} p_i \log(p_i)}_{-H(X)} + \sum_{i=1}^{m} \mathbb{E} g_i(Y)\log(p_i) + \sum_{i=1}^{m} \mathbb{E}\left(\Big(p_i + g_i(Y)\Big)\frac{g_i(Y)}{p_i}\right).$$

Using that $\mathbb{E} g_i(Y) = 0$ we conclude that

$$I(X,Y) = H(X) - H(X|Y) \le \sum_{i=1}^{m} \mathbb{E}\frac{g_i(Y)^2}{p_i} \overset{(12)}{\le} \alpha^2 \sum_{i=1}^{m}(1-p_i) = (m-1)\alpha^2.$$

□

Putting together the above ingredients we get Theorem 2.

Remark 4.3. Theorem 2 is actually true for the broader class of typical processes. This is simply because the correlation bound (9) holds for this class as well.

5. Examples

In this section we construct factor of i.i.d. processes showing that our bounds are (essentially) sharp.

5.1. **Sharpness of Theorem 1.** Let k be a fixed positive integer and $u, v \in V(T_d)$ vertices at distance k. We claim that there exist factor of i.i.d. processes X on T_d such that the normalized mutual information $I(X_u, X_v)/H(X_v)$ can be arbitrarily close to the upper bound

$$\beta_k := \begin{cases} \frac{2}{d(d-1)^l} & \text{if } k = 2l+1 \text{ is odd,} \\ \frac{1}{(d-1)^l} & \text{if } k = 2l \text{ is even.} \end{cases} \tag{13}$$

The idea is the following: given i.i.d. labels at each vertex, let the factor process "list" all the labels within some large distance R at any given vertex. When we look at the joint distribution of X_u and X_v we get a collection of i.i.d. labels with some labels listed twice. Hence the normalized mutual information is $|B_R(u) \cap B_R(v)|/|B_R(v)|$, where $B_R(v)$ denotes the ball of radius R around v. It is easy to see that this converges to β_k as $R \to \infty$.

For a rigorous argument we need to be more careful since listing the labels should be done in an $\mathrm{Aut}(T_d)$-invariant way. We first introduce two auxiliary lemmas and then precisely define our example.

Lemma 5.1. *For any positive integer L there exists a factor of i.i.d. 0-1 labeling of the vertices of T_d such that any ball of radius L contains a vertex with label 1 but any two vertices of label 1 have distance greater than L.*

Lemma 5.2. *For any positive integer L there exists a factor of i.i.d. coloring of the vertices of T_d such that finitely many colors are used and vertices of the same color have distance greater than L.*

Example 5.3. Given k and R, let $C = (C_w)_{w\in V(T_d)}$ be a factor of i.i.d. coloring provided by Lemma 5.2 for $L = 2R + k$. Given a positive integer N let Z_w, $w \in V(T_d)$ be i.i.d. uniform labels on $\{1, 2, \dots, N\}$. We set

$$X_v = \{(C_w, Z_w) \mid w \in B_R(v)\}.$$

Then for vertices u, v at distance k we have

$$\frac{I(X_u, X_v)}{H(X_v)} = \frac{|B_R(u) \cap B_R(v)|}{|B_R(v)|} + o_N(1). \tag{14}$$

Indeed, X_v can be viewed as the list of variables (C_w, Z_w), $w \in B_R(v)$ ordered by C_w (which are all different). This is now an $\mathrm{Aut}(T_d)$-invariant description. Conditioned on the coloring process C, the entropies are easy to compute:

$$H(X_v|C) = |B_R(v)| \log N \text{ and } H(X_u, X_v|C) = |B_R(u) \cup B_R(v)| \log N.$$

Since the contribution of the coloring to the entropies does not depend on N, it gets negligible when N is large enough, and (14) follows.

Finally, we prove the two lemmas we used.

Proof of Lemma 5.1. We describe the labeling as the output of a randomized local algorithm, which is easy to interpret as a factor of i.i.d. process.

In the beginning all labels are undefined. At every odd step every vertex with undefined label proposes to get a label 1 with probability 1/2. If there is no other proposition within distance L, the label is fixed, otherwise withdrawn. (Note that any undefined label gets fixed with probability at least some positive constant ε depending on L.) At even steps, undefined vertices check if a label 1 has appeared within distance L and set their own label 0 if this is the case.

It is easy to verify that the obtained process has all the required properties. □

Proof of Lemma 5.2. Lemma 5.1 is used to find vertices with color 1. A similar algorithm is applied for color 2, but now some vertices already have defined labels when launching the algorithm. We continue by adding more colors the same way.

After having added n colors this way, every ball of radius L around an uncolored vertex must contain vertices of each color $1, 2, \dots, n$. When n becomes equal to the number of vertices in a ball of radius L, this is not possible any longer, therefore we cannot have any more uncolored vertices at that point, meaning that we have colored all vertices in the required manner using at most n colors. □

5.2. **Sharpness of Theorem 2.** The next example shows that the bound obtained in Theorem 2 is essentially sharp. The idea is that we first take a linear factor of a standard normal i.i.d. process for which the correlation decay is close to the bound (9). Then we take the sign of the value of the process at every vertex. For this $\{\pm 1\}$-valued process the correlation decays at roughly the same rate. However, for symmetric binary variables the mutual information is essentially the square of the correlation. More precisely, for any $\varepsilon > 0$ we construct a factor of i.i.d. process (with two states) such that the mutual information for distance k is $\Omega\left(k^{2-\varepsilon}(d-1)^{-k}\right)$.

Example 5.4. Fix a parameter $\varepsilon > 0$. Let Z_w, $w \in V(T_d)$ be i.i.d. standard normal random variables. We first take a (linear) factor of the i.i.d. process Z:

$$Y_v := \sum_{w\in V(T_d)} \alpha_{\mathrm{dist}(v,w)} Z_w, \text{ where } \alpha_0 = 0 \text{ and } \alpha_k = \frac{k^{-\frac{1}{2}-\varepsilon}}{\sqrt{d-1}^k} \text{ for } k \geq 1;$$

then apply the sign function at each vertex:

$$X_v := \operatorname{sign}(Y_v).$$

Note that Y_v is well defined since the sum of the squares of the coefficients is finite. Therefore Y_v is a normal random variable with mean 0 and some positive and finite variance $\gamma = \gamma(\varepsilon)$. From this point on γ will denote a positive constant that depends only on ε (possibly a different constant at each occurrence).

Suppose that u and v have distance k. We denote the unique path connecting them by $u_0 = u, u_1, \ldots, u_{k-1}, u_k = v$. If we are at vertex u_j, $1 \leq j \leq k-1$, and move distance n away from the path, then we get to a vertex w for which $\operatorname{dist}(u,w) = j+n$ and $\operatorname{dist}(v,w) = k-j+n$. The number of such vertices is clearly $(d-2)(d-1)^{n-1}$. Thus

$$\operatorname{cov}(Y_u, Y_v) = \sum_{w \in V(T_d)} \alpha_{\operatorname{dist}(u,w)} \alpha_{\operatorname{dist}(v,w)} \geq \frac{\gamma}{\sqrt{d-1}^k} \sum_{j=1}^{k-1} \sum_{n=1}^{\infty} (j+n)^{-\frac{1}{2}-\varepsilon} (k-j+n)^{-\frac{1}{2}-\varepsilon}.$$

We ignore the terms for which $j+n<k$ and rearrange the rest of the sum grouping the terms based on the value $m := j+n$. For a given $m \geq k$ and $j \in \{1, \ldots, k-1\}$ we have $n = m-j$ and hence $k-j+n = k+m-2j$. Therefore the average of $k-j+n$ for a given m as j runs through $1, \ldots, k-1$ is exactly m, and consequently the convexity of $x^{-\frac{1}{2}-\varepsilon}$ implies that

$$\sum_{j=1}^{k-1} (k+m-2j)^{-\frac{1}{2}-\varepsilon} \geq (k-1) m^{-\frac{1}{2}-\varepsilon}.$$

It follows that

$$\operatorname{cov}(Y_u, Y_v) \geq \frac{\gamma(k-1)}{\sqrt{d-1}^k} \sum_{m=k}^{\infty} m^{-1-2\varepsilon} \geq \frac{\gamma(k-1)}{\sqrt{d-1}^k} \underbrace{\int_k^\infty x^{-1-2\varepsilon} \, \mathrm{d}x}_{k^{-2\varepsilon}/(2\varepsilon)} \geq \frac{\gamma k^{1-2\varepsilon}}{\sqrt{d-1}^k},$$

and the same is true for $\operatorname{corr}(Y_u, Y_v)$ (again with a different γ). Note that there exist constants $0 < \gamma < \tilde{\gamma}$ such that for any W, W' jointly normal random variables we have

$$\gamma \big| \operatorname{corr}(W, W') \big| \leq \big| \operatorname{corr}(\operatorname{sign}(W), \operatorname{sign}(W')) \big| \leq \tilde{\gamma} \big| \operatorname{corr}(W, W') \big|.$$

This means that we get the same correlation (up to a constant factor) after taking the sign of Y:

$$\operatorname{corr}(X_u, X_v) \geq \frac{\gamma k^{1-2\varepsilon}}{\sqrt{d-1}^k}.$$

Now working with symmetric binary variables, elementary computations show that when $P(X_u = X_v)$ is close to 1/2, we have

$$\gamma \left| P(X_u = X_v) - \frac{1}{2} \right| \leq \big| \operatorname{corr}(X_u, X_v) \big| \leq \tilde{\gamma} \left| P(X_u = X_v) - \frac{1}{2} \right|,$$

and

$$\gamma \left| P(X_u = X_v) - \frac{1}{2} \right|^2 \leq \big| I(X_u, X_v) \big| \leq \tilde{\gamma} \left| P(X_u = X_v) - \frac{1}{2} \right|^2.$$

It follows that

$$I(X_u, X_v) \geq \frac{\gamma k^{2-4\varepsilon}}{(d-1)^k},$$

which indeed confirms that the bound in Theorem 2 is essentially sharp.

References

[1] Ágnes Backhausz, Balázs Gerencsér, Viktor Harangi, and Máté Vizer. Correlation bound for distant parts of factor of iid processes. *Combin. Probab. Comput.*, to appear.

[2] Ágnes Backhausz and Balázs Szegedy. On large girth regular graphs and random processes on trees. arXiv:1406.4420, 2014.

[3] Ágnes Backhausz and Balázs Szegedy. On the almost eigenvectors of random regular graphs. arXiv:1607.04785, 2016.

[4] Ágnes Backhausz, Balázs Szegedy, and Bálint Virág. Ramanujan graphings and correlation decay in local algorithms. *Random Structures Algorithms*, 47(3):424–435, 2015.

[5] Karen Ball. Factors of independent and identically distributed processes with non-amenable group actions. *Ergodic Theory Dyn. Syst.*, 25(3):711–730, 2005.

[6] Lewis Bowen. A measure-conjugacy invariant for free group actions. *Ann. Math. (2)*, 171(2):1387–1400, 2010.

[7] Lewis Bowen. The ergodic theory of free group actions: entropy and the f-invariant. *Groups Geom. Dyn.*, 4(3):419–432, 2010.

[8] Lewis Bowen. Measure conjugacy invariants for actions of countable sofic groups. *J. Am. Math. Soc.*, 23(1):217–245, 2010.

[9] Endre Csóka. Independent sets and cuts in large-girth regular graphs. arXiv:1602.02747, 2016.

[10] Endre Csóka, Balázs Gerencsér, Viktor Harangi, and Bálint Virág. Invariant Gaussian processes and independent sets on regular graphs of large girth. *Random Structures Algorithms*, 47(2):284–303, 2015.

[11] Viktor Harangi and Bálint Virág. Independence ratio and random eigenvectors in transitive graphs. *Ann. Probab.*, 43(5):2810–2840, 2015.

[12] Carlos Hoppen and Nicholas Wormald. Local algorithms, regular graphs of large girth, and random regular graphs. *To appear in Combinatorica.* arXiv:1308.0266.

[13] Donald S Ornstein and Benjamin Weiss. Entropy and isomorphism theorems for actions of amenable groups. *J. Analyse Math*, 48:1–141, 1987.

[14] Mustazee Rahman. Factor of iid percolation on trees. arXiv:1410.3745, 2014.

[15] Mustazee Rahman and Bálint Virág. Local algorithms for independent sets are half-optimal. *Annals of Prob.*, to appear.

MTA Alfréd Rényi Institute of Mathematics H-1053 Budapest, Reáltanoda utca 13-15; and Eötvös Loránd University, Department of Probability and Statistics H-1117 Budapest, Pázmány Péter sétány 1/c

E-mail address: gerencser.balazs@renyi.mta.hu

MTA Alfréd Rényi Institute of Mathematics H-1053 Budapest, Reáltanoda utca 13-15

E-mail address: harangi@renyi.hu