Comparing Entropy Rates on Finite and Infinite Rooted Trees

Thomas Hirschler and Wolfgang Woess

Abstract—We consider denumerable stochastic processes with (or without) memory. Their evolution is encoded by a finite or infinite rooted tree. The main goal is to compare the entropy rates of a given base process and a second one, to be considered as a perturbation of the former. The processes are described by probability measures on the boundary of the given tree and by corresponding forward transition probabilities at the inner nodes. The comparison is in terms of Kullback–Leibler divergence. We elaborate and extend ideas and results of Öcherer and Amjad. Our extensions involve length functions on the edges of the tree as well as nodes with countably many successors. In particular, in Section V, we consider trees with infinite nonbacktracking paths and random perturbations of a given process.

Index Terms—Tree, forward Markov chain, entropy rate, Kullback–Leibler divergence, perturbed stochastic process.

I. INTRODUCTION

ÖCHERER and Amjad [1] consider probability distributions on the set of leaves of a finite, rooted tree. They derive a bound on the difference of the entropies between two such distributions, where the second one is “memory-less” in terms of an encoding of the tree in terms of strings over a finite alphabet. The bound is in terms of the Kullback-Leibler divergence of the first with respect to the second distribution, and of the lengths (heights) of the leaves; see (1) below.

In this note, we generalise these inequalities, thereby also streamlining the mathematical substance. Our generalisations are qualitative rather than being quantitative in the sense of sharp numerical bounds. They consist in (1) comparing the entropies of two arbitrary distributions on the leaves in terms of the Kullback-Leibler divergence, (2) involving more general length functions than those where each edge of the tree has length 1, and (3) allowing internal nodes with a countable infinity of forward neighbours. Finally, (4), we also consider trees which have infinitely long non-backtracking paths, corresponding to boundary points at infinity, and (5) consider the entropy rate of a random perturbation of a given stochastic process.

In the main parts of this paper – except for § V – the boundary ∂T of our tree T with root o consists of leaves (vertices with no forward neighbour) and is at most countably infinite.

It is equipped with two probability measures P and Q, where Q plays the role of a fixed reference measure and P should be thought of as a perturbation of Q which corresponds to some experimental situation: for example, the perturbation errors present in P may be inherent in the simulation of a stochastic process whose distribution is Q.

The internal vertices (nodes) x of T are equipped with a function ℓ(x) > 0 with the interpretation that each outgoing edge at x has length ℓ(x). The absolute value |x|_ℓ of any node is then the sum of the lengths of the edges on the unique geodesic path from o to x. The habitual choice is ℓ = ℓ_o ≡ 1. Thus we have the expected length ℓ( P) = ∑_o∈T |o|_P P(o), and analogously ℓ(Q). Let H(P) and H(Q) be the entropies of the two measures, and D(P∥Q) the Kullback-Leibler divergence of P with respect to Q. The aim is to estimate the difference between the entropy rates H(P)/ℓ( P) and H(Q)/ℓ(Q) in terms of D(P∥Q)/ℓ( P).

The probability measure P gives rise to a forward Markov chain on the internal nodes of T, whose transition probabilities are p(y|x) = P(∂T_y)/P(∂T_x), when y is a forward neighbour of x, and where ∂T_x denotes the set of leaves v for which x lies on the geodesic path from o to v. The analogous transition probabilities associated with Q are denoted q(y|x). A basic tool for achieving the desired estimate is the “Leaf-Average Node-Sum Interchange Theorem” (LANSIT) of Rueppel and Massey [11]. It relates H(P)/ℓ(P) with the entropies H(p(·|x)) and D(P∥Q)/ℓ(P) with the divergences D(p(·|x)∥q(·|x)).

The main concern of [1] regards the situation when the tree is finite, ℓ = ℓ_o, and all probability measures q(·|x) = q(·) coincide under a suitable labelling of the sets of forward neighbours of the internal nodes x. In this case, the LANSIT yields H(P)/ℓ_o(Q) = H(q), and the estimate of [1] can be read as

\[ \left| \frac{H(P)}{\ell_o(P)} - H(q) \right| < M\phi(\delta) + C\delta \sqrt{\frac{D(P\parallel Q)}{\ell_o(P)}}, \quad (1) \]

where δ is an arbitrarily small positive constant, M is the (constant) number of forward neighbours of the internal nodes, \( \phi(\delta) = \delta \log_2(1/\delta) \), and C ≈ 1.25.

We start in § II by explaining our notation, which comprises some variations with respect to previous work [1], [11]. We review the proof of the LANSIT in presence of of a general length function as above. As mentioned, our tree is allowed to have nodes with infinitely many forward neighbours. Thus, our entropies and divergences can be infinite series, and we require the latter to be absolutely convergent. Furthermore, we have local entropies H(p(·|x)) and H(q(·|x)) at the possibly infinitely many internal nodes, and we require the defining series

Manuscript received April 19, 2017; accepted December 10, 2017. Date of publication December 27, 2017; date of current version July 12, 2018.

The authors are with the Institut für Diskrete Mathematik, Technische Universität Graz, A-8010 Graz, Austria (e-mail: thirschler@tugraz.at; woess@tugraz.at).

Copyright (c) 2017 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

Communicated by P. Harremoës, Associate Editor for Probability and Statistics.

Digital Object Identifier 10.1109/TIT.2017.2787712
to be uniformly summable. See §III, which consists mainly in a “guided tour” through the needed summability properties and simpler conditions under which they hold. Some further basics are also presented in §III.

Theorem 1 in §IV contains our generalisation of (1). We consider arbitrary probability measures \( P \) and \( Q \) that are supported by \( \partial T \) and satisfy the mentioned summability conditions. Consequently, our estimate of \( H(P)/\ell(P) - H(Q)/\ell(Q) \) is a sum of three terms, where the first can be made arbitrarily small, the central one comprises \( \sqrt{D(P||Q)/\ell(P)} \), and the last one is a multiple of the variational difference between the node-average probability measures induced by \( P \) and \( Q \) on the set of internal vertices. The pre-factor of this term vanishes precisely when \( H(q(x)/|x|)/\ell(x) \) is constant (or at least asymptotically constant) on the internal nodes. After a discussion of various aspects and variations of this estimate, in §V we turn our attention to trees which have infinite non-backtracking paths (“geodesics”). In this case, the boundary of the tree consists of the space of “ends” (limit points of infinite paths) and possibly also leaves. Applying Theorem 1 yields estimates for the entropy rate and variants thereof, see Theorem 2. In particular, coming back to the natural length function \( \ell_x \), if \( Q \) is the probability distribution on the trajectory space of a stochastic process with asymptotically constant, finite forward entropies \( H(q(x)/|x|) \) on the associated infinite tree and \( P \) is the probability governing a perturbation of that process, then finiteness of \( D(P||Q) \) implies that the perturbed process has the same entropy rate as the non-perturbed one. Note again that we allow the state space of the processes to be countably infinite, as long as \( H(q(x)/|x|) \) is finite. Moreover, equality of the entropy rates also holds when \( D(P_n||Q_n)/n \to 0 \), where \( P_n \) and \( Q_n \) are the distributions of the respective processes up to time \( n \). A final application is given in Theorem 3, where we consider random perturbations of a given stochastic process.

The proofs of all theorems are postponed to the Appendix. We remark that in a somewhat different vein, there has been work on the statistics of ergodic Markov chains. Ciucuera and Girardin [2] and Ciucuera et al. [3], as well as Yari and Nikooravesh [14], consider the approximation of the entropy rate via the natural estimator (in terms of relative frequencies of transitions between two states, resp. visits in one state) of the chain at time \( n \), as \( n \to \infty \). Constantness of local entropies of the chain is not needed here. Note that this is not comparison with one perturbed chain, but with a sequence of random perturbations: the estimated transition probabilities and invariant distribution are updated at each time step. In [2], the state space is finite, and in [14], at each state there are only finitely many positive outgoing transition probabilities. Both correspond to situations where the corresponding infinite tree is locally finite. In [3], this finite range assumption is replaced by a quasi-power property.

II. SETTING, NOTATION, AND PRELIMINARY FACTS

A. Trees, Leaves, and Lengths

A rooted tree is a connected graph \( T \) without circles, with one vertex \( o \) designated as the root. We shall tacitly identify \( T \) with the set of vertices of the tree, and write \( E(T) \) for the set of edges, if needed.

A geodesic or geodesic path in \( T \) is a sequence \( \pi = [x_0, x_1, \ldots, x_n] \) of successive neighbours in \( T \) such that all \( x_i \) are distinct. For any pair of vertices \( x, y \) there is precisely one geodesic \( \pi(x, y) \) from \( x \) to \( y \). Analogously, an infinite geodesic has the form \( \pi = [x_0, x_1, x_2, \ldots] \) where \( x_{k-1} \) and \( x_k \) are neighbours for each \( k \in \mathbb{N} \), and there is no backtracking, i.e., \( x_{k+1} \neq x_k \).

While we do not necessarily assume \( T \) to be finite, we shall mostly assume that it contains no infinite geodesic, with the exception of §V.

Every vertex \( x \in T \setminus \{o\} \) has a unique predecessor \( x^- \), the neighbour of \( x \) on the geodesic \( \pi(o, x) \). In this case, we say that \( x \) is a successor of \( x^- \), and write \( N(x) \) for the set of successors of \( x \in T \). The forward degree \( \deg^+(x) \) of \( x \in T \) is the number of successors of \( x \). In our setting, it will be natural to assume that \( \deg^+(x) \neq 1 \) for all \( x \in T \). The set of leaves or boundary \( \partial T \) of \( T \) consists of all vertices which have no successor. By our assumption, it is non-empty, and every vertex lies on a geodesic from \( o \) to some leaf. The interior of \( T \) is \( \text{int} T = T \setminus \partial T \).

We can describe each edge of \( T \) as \( e = [x^-, x] \), where \( x \in T \setminus \{o\} \). We now assign a length \( \ell(e) > 0 \) to each edge \( e \).

We restrict attention to length functions where \( \ell(e) \) depends only on \( x^- \) for \( e = [x^-, x] \).

That is, at each vertex the outgoing edges have the same length, and for \( e \) as above one can write \( \ell(e) = \ell(x^-) \). We remark here that this restriction is not completely satisfactory: in general, each edge might have a label from some alphabet, and one would like the cost (length) of the edge to depend on that label. However, in this case, one of the tools, the LANSIT lemma 1 fails in general. But with the interpretation that the cost of an edge in the next step along \( T \) depends on the last used symbol of the alphabet, one does have a reasonable model. For example, it is realistic in a Markovian setting of the evolution of a language that the transitions and their cost in the next step depend on the last input symbol.

Back to our length function, the length of a path is defined as the sum of the lengths of its edges. For a vertex \( x \), its length (distance to \( o \)) \( |x| \) is the length of \( \pi(o, x) \). The standard length function is induced by \( \ell_x(e) = 1 \). The associated path length is the number of edges. For \( x \in T \), we then just write \( |x| \) for the resulting length of \( \pi(o, x) \). This is the ordinary height of \( x \).

B. Forward Markov Chains and Leaf Distributions

On a rooted tree, a forward Markov chain is given by transition probabilities

\[
p(y|x), \quad x \in \text{int} T, \quad y \in N(x), \quad \text{with} \quad \sum_{y \in N(x)} p(y|x) = 1.
\]

We can interpret this via a random process: a particle starts at \( o \). If at some time it is at a vertex \( x \in \text{int} T \) then the particle moves to a random successor of \( x \) chosen according to the probability distribution \( p(x) \). When the particle reaches a leaf, it stops there (so that we might add the trivial transition probabilities \( p(o|o) = 1 \) for \( o \in \partial T \)).
We now define a measure (function) on $\partial T$, as follows.
\[
\mathbb{P}(v) = \prod_{x \in \pi(o,v) \setminus \{o\}} p(x|\pi(x), \quad v \in \partial T. \quad (2)
\]

This is the hitting distribution on that set: for a leaf $v$, one has that $\mathbb{P}(v)$ is the probability that the forward Markov chain terminates at $v$. Conversely, given any probability measure $\mathbb{P}$ on $\partial T$, it induces a forward Markov chain as follows.

**Definition 1:** For $x \in T$, the cone or branch at $x$ is the sub-tree with root $x$ given by
\[
T_x = \{ y \in T : x \in \pi(o,y) \}
\]
Its boundary $\partial T_x$ consists of all $v \in \partial T$ with $x \in \pi(o,v)$. Then we define
\[
p(x|v-) = \frac{\mathbb{P}(\partial T_x)}{\mathbb{P}(\partial T_{x-})},
\]
where $\mathbb{P}(\partial T_x) = \sum_{v \in \partial T_x} \mathbb{P}(v)$. Thus, we have a one-to-one correspondence between probability distributions on $\partial T$ and forward Markov chains on the rooted tree $T$. See [11] for an introduction to these ideas. More generally, we can formulate the following.

**Definition 2:** A cross section or cut is a subset $S \subset T$ such that for every leaf $v \in \partial T$, the geodesic $\pi(o,v)$ meets $S$ in precisely one point. We write $T_x^S$ for the sub-tree with root $o$ consisting of all vertices that lie on some geodesic $\pi(o,s)$ with $s \in S$.

We have $\partial T^S = S$ for any section. We define $\mathbb{P}_S$ on $S$ in the same way as $\mathbb{P}$ was defined on $\partial T$ in (2), so that $\mathbb{P}_S(x) = \mathbb{P}(\partial T_x)$. We get the hitting distribution of the forward Markov chain on $S$, a probability distribution.

Given our length function $\ell$, the expected length of a section $S$ with respect to $\mathbb{P}_S$ is
\[
\ell(\mathbb{P}_S) = \sum_{x \in S} |x| \mathbb{P}_S(x), \quad \text{in particular,} \quad \ell(\mathbb{P}) = \ell(\mathbb{P}_{\partial T}).
\]
In §II – §IV, we shall always assume that $\ell(\mathbb{P}) < \infty$. At last, given the probability distribution $\mathbb{P}$ on $\partial T$, we define a new (node average) measure $\mu_{\mathbb{P}}$ on $\mathbb{P}$ by
\[
\mu_{\mathbb{P}}(x) = \frac{\ell(x)}{\ell(\mathbb{P})} \mathbb{P}(\partial T_x). \quad (3)
\]
It is a probability measure: we use the fact that $v \in \partial T_x \iff x \in \pi(o,v) \setminus \{o\}$ for any leaf $v$ and internal node $x$, whence
\[
\sum_{x \in \mathbb{P}} \ell(x) \mathbb{P}(\partial T_x) = \sum_{x \in \mathbb{P}} \ell(x) \sum_{v \in \partial T_x} \mathbb{P}(v)
= \sum_{v \in \partial T} \mathbb{P}(v) \sum_{x \in \pi(o,v)} \ell(x) = \ell(\mathbb{P}).
\]

**C. Gradient, Laplacian, and Leaf-Node Interchange**

**Definition 3:** The gradient $\nabla$ and Laplacian $\Delta = \Delta_{\mathbb{P}}$ on $T$ are as follows. For a function $f : T \to \mathbb{R}$
\[
\nabla f (y) = \frac{f(y) - f(y^-)}{\ell(y)}, \quad y \in T \setminus \{o\}, \quad \text{and} \quad \Delta f(x) = \sum_{y \in N(x)} \nabla f(y) p(y|x).
\]

Some remarks are due: here, our notation differs from [1] as well as [11], who write $\Delta$ for our $\nabla$. In general, one should see $\nabla f$ as a function on $E(T)$, where $f$ could be interpreted as a voltage function For $e = [y^-, y]$, the difference $f(y) - f(y^-)$ is the voltage drop, and thinking of $\ell(e)$ as the resistance of the edge, $\nabla f(e) = (f(y) - f(y^-))/\ell(e)$ can be interpreted as Ohm’s law. With our assumption, $\ell(e) = \ell(y^-)$, and we can adopt the above notation. Regarding the Laplacian
\[
\Delta f(x) = \frac{1}{\ell(x)} \sum_{y \in N(x)} (f(y) - f(x)) p(y|x),
\]
as long as the forward degrees are finite, it is well defined for any function $f$. In presence of infinitely many successors, we require that the series converges absolutely.

**Assumption 1:** For every $x \in \mathbb{P}$,
\[
|\Delta_{\mathbb{P}}| f(x) := \frac{1}{\ell(x)} \sum_{y \in N(x)} |f(y) - f(x)| p(y|x) < \infty.
\]

Our notation is compatible with the one of reversible Markov chain theory, see e.g. [13, Ch. 4], although the present forward Markov chains are not reversible.

For a probability distribution $\nu$ on any finite or countable set $X$, and any function $f : X \to \mathbb{R}$, we write
\[
\mathbb{E}_{\nu}(f) = \sum_{x \in X} f(x)v(x),
\]
whenever $f$ is $\nu$-integrable, that is $\mathbb{E}_{\nu}(|f|) < \infty$. In this setting, the “Leaf-average node-sum interchange theorem” (LAN-SIT) of [11] is the following.

**Lemma 1:** Let $\mathbb{P}$ be a probability distribution on $\partial T$ and $p(\cdot|\cdot)$ the associated forward Markov chain. Let $f : T \to \mathbb{R}$ be a function which satisfies $\mathbb{E}_{\mu_{\mathbb{P}}}(|\Delta_{\mathbb{P}}f|) < \infty$. Then $f$ is $\mathbb{P}$-integrable on $\partial T$, and
\[
\frac{1}{\ell(\mathbb{P})} \mathbb{E}_{\mathbb{P}}(f - f(o)) = \mathbb{E}_{\mu_{\mathbb{P}}}(\Delta_{\mathbb{P}}f).
\]

This is a slight generalisation of [11, Th. 1], where only $\ell = \ell_1$ is considered. However, in the proof of [11], it is suggested in the last lines of p. 347 that already the condition $\mathbb{E}_{\mu_{\mathbb{P}}}(|\Delta_{\mathbb{P}}f|) < \infty$ is sufficient for the conclusion. But this appears to be problematic, since it does not allow to separate in two parts the sum as displayed in the middle of the proof on that page. For safety’s sake, we present an updated proof.

**Proof of Lemma 1:** We start with the right hand side:
\[
\mathbb{E}_{\mu_{\mathbb{P}}}(\Delta_{\mathbb{P}}f) \overset{(a)}{=} \frac{1}{\ell(\mathbb{P})} \sum_{x \in \partial T} \sum_{y \in N(x)} (f(y) - f(x)) p(y|x) \mathbb{P}(\partial T_x)
\overset{(b)}{=} \frac{1}{\ell(\mathbb{P})} \sum_{y \in N(x)} \sum_{o \in \partial T_x} (f(y) - f(y^-)) \mathbb{P}(\partial T_y)
\overset{(c)}{=} \frac{1}{\ell(\mathbb{P})} \sum_{o \in \partial T} \sum_{v \in \partial T_x} (f(y) - f(y^-)) \mathbb{P}(v)
\overset{(d)}{=} \frac{1}{\ell(\mathbb{P})} \sum_{v \in \partial T} \sum_{x \in \pi(o,v)} (f(y) - f(y^-)) \mathbb{P}(v)
\overset{(e)}{=} \frac{1}{\ell(\mathbb{P})} \sum_{v \in \partial T} (f(v) - f(o)) \mathbb{P}(v).
By assumption, the right hand side of (a) is absolutely convergent, so that the sum can be rearranged in order to get (b) and thus (c), which are also absolutely convergent. This allows to exchange the order of summation, leading to (d). □

III. ENTROPIES AND TIGHTNESS

A. Entropies and Relative Entropies

The entropy of our probability measure \( P \) on \( \partial T \) is

\[
H(P) = \sum_{v \in \partial T} P(v) \log_2 \frac{1}{P(v)}.
\]

When we have infinite forward degrees in \( T \) then \( \partial T \) is infinite. In this case, we assume that \( H(P) < \infty \). At each inner node \( x \) of \( T \), we have the outgoing probability distribution \( p(\cdot|x) \) on the set of successors of \( x \). Its entropy is defined analogously,

\[
H_P(x) = H(p(\cdot|x)) = \sum_{y \in N(x)} p(y|x) \log_2 \frac{1}{p(y|x)}.
\]

Lemma 2: If, as assumed, \( H(P) < \infty \) then \( H_P(x) \) is finite for each \( x \), and

\[
\frac{H(P)}{\ell(P)} = \mathbb{E}_{\mu_P} \left( \frac{H_P(x)}{\ell(x)} \right).
\]

On the right hand side, \( x \) is the variable with respect to which the expectation is taken.

Proof: As in [11], we set \( f(x) = -\log_2 P(\partial T_x) \). Then \( \Delta \mathbb{P} f \geq 0 \), so that Lemma 1 applies even when the involved expectations are infinite. Thus, the stated formula holds, and when \( H(P) < \infty \), then also \( H_P(x) \) is finite for all \( x \). □

Note in particular the following special choice of the length function.

If \( \ell(x) = H_P(x) \) then \( \ell(P) = H(P) \).

(4)

Now let us take a second probability measure \( Q \) on \( \partial T \), and denote the transition probabilities of the associated forward Markov chain by \( q(y|x) \), when \( y^{-} = x \). The well-known Kullback-Leibler divergence or relative entropy of \( P \) with respect to \( Q \) is

\[
D(P\|Q) = \sum_{v \in \partial T} P(v) \log_2 \frac{P(v)}{Q(v)}.
\]

For our purposes, it will always suffice to consider the non-degenerate situation when \( P(v), Q(v) > 0 \) for each leaf \( v \). In the case when \( \partial T \) is infinite, we assume again that the above series converges absolutely. Similarly, we have the local version

\[
D_{p,q}(x) = D(p(\cdot|x)\|q(\cdot|x)), \quad x \in \text{int} T.
\]

We need again absolute convergence of its defining series. The following generalises the corresponding result of [1] to the case where \( T \) has vertices with infinite forward degree.

Proposition 1: If \( H(P) < \infty \) and the defining series of \( D(P\|Q) \) is absolutely convergent, then the same holds for the defining series of each \( D_{p,q}(x) \), and

\[
\frac{D(P\|Q)}{\ell(P)} = \mathbb{E}_{\mu_P} \left( \frac{D_{p,q}(x)}{\ell(x)} \right).
\]

Again, \( x \) is the variable with respect to which the expectation is taken on the right hand side.

Proof: For the function \( f(x) = \log_2(P(\partial T_x)/Q(\partial T_x)) \), we compute

\[
\ell(P) \mathbb{E}_{\mu_P}(|\Delta \mathbb{P}| f) = \sum_{x \in \partial T} P(\partial T_x) \sum_{y \in N(x)} p(y|x) \left| \log_2 \frac{p(y|x)}{q(y|x)} \right|
\]

\[
= \sum_{y \in N(x) \setminus \{x\}} \sum_{v \in \partial T} P(v) \left| \log_2 \frac{p(y|x)}{q(y|x)} \right|
\]

\[
\leq \sum_{y \in N(x) \setminus \{x\}} \sum_{v \in \partial T} \left( \log_2 \frac{1}{p(y|x)} + \log_2 \frac{1}{q(y|x)} \right)
\]

\[
= \sum_{y \in N(x) \setminus \{x\}} \sum_{v \in \partial T} \left( \log_2 \frac{P(v)}{Q(v)} \right)
\]

\[
= 2H(P) + \sum_{v \in \partial T} P(v) \left| \log_2 \frac{P(v)}{Q(v)} \right|
\]

which is finite by assumption. In particular, we get for each internal node \( x \) that \( |\Delta \mathbb{P}| f(x) < \infty \), that is, \( D_{p,q}(x) \) converges absolutely. Therefore we can apply Lemma 1 to \( f(x) \) to obtain the proposed identity. □

In the following two lemmas we state without proofs some elementary properties of the variational distance between two probability distributions, and the defining function of the entropy.

Lemma 3: For any two probability distributions \( v_1, v_2 \) on a finite or countable set \( X \), let

\[
||v_1 - v_2||_1 = \sum_{x \in X} |v_1(x) - v_2(x)|
\]

Then we have the following.

(i) \[
\sum_{x : v_1(x) > v_2(x)} (v_1(x) - v_2(x)) = \sum_{x : v_1(x) < v_2(x)} (v_2(x) - v_1(x)) = ||v_1 - v_2||_1 / 2.
\]

(ii) \[
||v_1 - v_2||_2^2 \leq 2 \ln 2 \cdot \mathbb{D}(v_1 \| v_2).
\]

The latter is Pinsker’s inequality. The optimal constant \( 2 \ln 2 \) is due to Csiszár [5], Kemperman [9] and Kullback [10] in independent work of the 1960ies. There are also converse inequalities; see [6] and [12] and the references given there. The following is elementary.

Lemma 4: The function \( \phi(t) = t \log_2(1/t), t \in [0, 1] \), with the convention \( \phi(0) = 0 \), has the following properties.

(i) \( \phi \) is strictly concave and assumes its maximal value \( 1/(e \ln 2) \) at \( t = 1/e \).

(ii) For \( 0 \leq t < t + \delta \leq 1 \),

\[
|\phi(t + \delta) - \phi(t)| \leq \max\{\phi(\delta), \phi(1 - \delta)\}.
\]

(iii) For \( 0 \leq \delta \leq 1/2 \), one has \( \max\{\phi(\delta), \phi(1 - \delta)\} = \phi(\delta) \).
(iv) For $t, u \in [0, 1/(2\varepsilon)]$, we have
\[ \varphi(t + u) \leq 2(\varphi(t) + \varphi(u)). \]

B. Tightness and Uniform Summability

Returning to the setting of a countable tree, suppose that for each internal node $x$, we have a real valued function $y \mapsto g(y|x)$, defined on the successors $y$ of $x$. We are interested in the series $\sum_{y \in N(x)} g(y|x)$. Then we say that the family of functions, resp., the associated series are tight or uniformly (absolutely) summable, if for every $\varepsilon > 0$ there is an integer $M_\varepsilon \geq 1$ such that for each $x \in \text{int } T$, there is a subset $N_\varepsilon(x) \subset N(x)$ such that
\[ |N_\varepsilon(x)| \leq M_\varepsilon \quad \text{and} \quad \sum_{y \in N(x) \setminus N_\varepsilon(x)} |g(y|x)| < \varepsilon. \]
(The terminology “tight” is commonly used in presence of a family of probability distributions.) As long as the tree is finite, (5) is always valid. In this case, we need no $\varepsilon$, i.e., we can replace $N_\varepsilon(x)$ by all of $N(x)$ and $M_\varepsilon$ by
\[ M = \max\{\deg^+(x) : x \in \text{int } T\}. \]

When there are only finitely many types of different functions $g(\cdot|x)$ up to bijections between the sets $N(x)$, $x \in \text{int } T$, then (5) is equivalent with absolute convergence of the involved series.

Assumption 2: It will be required that all the sums
\[ H_p(x) = \sum_y p(y|x) \log \frac{1}{p(y|x)} \quad \text{and} \quad H_q(x) = \sum_y q(y|x) \log \frac{1}{q(y|x)} \]
are uniformly summable.

Uniform summability implies with a short computation that $H_q(x) \leq \log_2 M_\varepsilon + 1/(\varepsilon \ln 2) + \varepsilon$. In particular, if $\inf\{\ell(x) : x \in \text{int } T\} > 0$ then $A = \sup\{H_q(x)/\ell(x) : x \in \text{int } T\} < \infty$. This will be needed further below.

We next present sufficient conditions for uniform summability which are easy to grasp. Enumerate each set of forward neighbours $N(x) = \{y_n(x) : n \in \mathbb{N}, n \leq \deg^+(x)\}$ (finite or countable) and set
\[ p_s(n) = \begin{cases} 
  p(y_n(x)|x) & \text{for } n \leq \deg^+(x) \quad \text{and} \quad \text{set } \deg^+(x) \text{ (finite)}; \\
  0 & \text{for } n > \deg^+(x) \quad \text{(the latter when } \deg^+(x) \text{ is finite)}. 
\end{cases} \]

Then tightness of the family of distributions $p_s(\cdot|x)$, $x \in \text{int } T$, means that we can do the above enumerations in such a way that
\[ \tau(k) = \sup\{\tau_x(k) : x \in \text{int } T\} \to 0 \quad \text{as } k \to \infty, \]
where $\tau_x(k) = \sum_{n=k}^\infty p_s(n)$.

In this case, $p(n) = \tau(n) - \tau(n+1)$ defines a probability distribution on $\mathbb{N}$, which stochastically dominates each distribution $p_s$, that is, its tails $\tau(k)$ majorise the tails of all $p_s$.

Lemma 5: Suppose that the distributions $p_s = p(\cdot|x)$ satisfy (7) and that the dominating distribution $p$ has finite mean,
\[ \sum_{n=1}^\infty np(n) < \infty. \]

Then the entropies $H_p(x), x \in \text{int } T$, are uniformly summable. If in addition the two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on $\partial T$ are such that
\[ \frac{1}{c} \mathbb{Q} \leq \mathbb{P} \leq c \mathbb{Q} \]
for some finite $c > 0$, then also the entropies $H_p(x), x \in \text{int } T$, are uniformly summable.

Proof: By Lemma 4, the function $\varphi(t)$ is monotone increasing in the interval $[0, 1/e]$. If $n \geq 2$ and $p_s(n) \leq 2^{-n}$ then $\varphi(p_s(n)) \leq n2^{-n}$. Otherwise, $\varphi(p_s(n)) \leq np_s(n)$. Choose $m \geq 2$. We have
\[ \sum_{n=m}^\infty np_s(n) = m \tau_m(m) + \sum_{k=m+1}^\infty \tau_s(k) \leq m \tau_m(m) + \sum_{k=m+1}^\infty \tau(k) = \sum_{n=m}^\infty np_s(n), \]
Consequently,
\[ \sum_{n=m}^\infty \varphi(p_s(n)) \leq \sum_{n=m}^\infty n(2^{-n} + p(n)), \]
which tends to 0 as $m \to \infty$. This proves uniform summability. The second part is obvious. \hfill \Box

IV. A COMPARISON THEOREM—DISCUSSION AND VARIATIONS

Here is our first main result. Recall that in this section, the underlying tree may have nodes with infinitely many neighbours, but the boundary consists of at most countably many leaves.

Theorem 1: Let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures supported by all leaves in $\partial T$, and let $p(\cdot|x)$ and $q(\cdot|x)$ be the transition kernels of the respective associated forward Markov chains. Suppose that $H(\mathbb{P}), H(\mathbb{Q})$ and $D(\mathbb{P}\|\mathbb{Q})$ are finite and that the defining series of $H_p(x)$ and $H_q(x), x \in \text{int } T$, are uniformly summable (without loss of generality with the same value $M_\varepsilon$ and sets $N_\varepsilon(x)$ for every $\varepsilon > 0$).

Then, for any choice of $\varepsilon > 0$ and $0 < \delta < 1/2$,
\[ \frac{H(\mathbb{P})}{\ell(\mathbb{P})} - \frac{H(\mathbb{Q})}{\ell(\mathbb{Q})} \leq L \left( 2\varepsilon + M_\varepsilon \varphi(\delta) \right) \]
\[ + \frac{C \sqrt{L}}{\delta} \sqrt{D(\mathbb{P}\|\mathbb{Q})} + A - a \|\mathbb{P} - \mathbb{Q}\|_1, \]
where
\[ L = \frac{\ell^2(\mathbb{P})}{\ell(\mathbb{P})}, \quad C = \frac{2\sqrt{2}}{e\sqrt{\ln 2}} \approx 1.25, \]
\[ A = \sup \left\{ \frac{H_q(x)}{\ell(x)} : x \in \text{int } T \right\}, \quad \text{and} \]
\[ a = \inf \left\{ \frac{H_q(x)}{\ell(x)} : x \in \text{int } T \right\}. \]
When $T$ has bounded forward degrees, we can set $\varepsilon = 0$ and replace the $M_\varepsilon$ of (5) by the $M$ of (6).

The proof of the theorem is deferred to the Appendix. Right now, we present a detailed discussion of various aspects and special cases of this inequality. As outlined in the Introduction, $\mathbb{Q}$ is to be considered as the basic reference measure, while $\mathbb{P}$ may vary; it should be viewed as a perturbation of $\mathbb{Q}$.

We say that two internal nodes $x$ and $x'$ have the same $\mathbb{Q}$-type, if there is a bijection $\psi : N(x) \rightarrow N(x')$ such that 

$$q(\psi y | x') = q( y | x) \text{ for all } y \in N(x).$$

(9)

A. Identical $\mathbb{Q}$-Types

The simplest, but most significant situation is the one where all internal nodes have the same $\mathbb{Q}$-type. In this case, $H_q(x) = H_{\mathbb{Q}}$ is constant, and uniform summability of $H_q(x)$ reduces to finiteness of $H_q$. If in addition $\ell = \ell_2$, then $A = a$, and by Lemma 2, one has $H(\mathbb{Q}) = \ell_2(\mathbb{Q}) \cdot H_q$. The inequality of Theorem 1 becomes

$$\left| \frac{H(\mathbb{P})}{\ell_2(\mathbb{P})} - H_q \right| \leq 2\varepsilon + M_\varepsilon \phi(\delta) + \frac{C}{\delta} \sqrt{\frac{D(\mathbb{P} \| q^{\mathbb{Q}^n})}{\ell_2(\mathbb{P})}}.$$ 

(10)

When $T$ is finite, so that we can set $\varepsilon = 0$ and replace $M_\varepsilon$ by $M$, this is the main result of [1]. The inequality (10) provides an extension to the case when all $N(x)$ are infinite and there is one $\mathbb{Q}$-type with $H_q < \infty$. Note that when in addition the measures $\mathbb{P}$ and $\mathbb{Q}$ are comparable as in (8) then the uniform summability conditions of Theorem 1 hold.

Suppose furthermore that $T$ has height $h$, that is, $|v|_2 = n$ for each leaf of $T$. Then with suitable labelling, $\mathbb{Q} = q^{\mathbb{Q}^n} = q \otimes \cdots \otimes q$ is the $n$-fold product measure of $q(x) := q(x | o)$, $x \in N(o)$, or in other words, it is the joint distribution of $n$ independent random variables $X_1, \ldots, X_n$ with common distribution $q(\cdot)$, and (10) becomes

$$\left| \frac{H(\mathbb{P})}{n} - H_q \right| \leq 2\varepsilon + M_\varepsilon \phi(\delta) + \frac{C}{\delta} \sqrt{\frac{D(\mathbb{P} \| q^{\mathbb{Q}^n})}{n}}.$$ 

(11)

Assume that for fixed $n$, we have a sequence $\mathbb{P}^{(k)}$, $k = 1, 2, \ldots$, each of which is a perturbation of the joint distribution $\mathbb{Q}$ of $(X_1, \ldots, X_n)$ such that $D(\mathbb{P}^{(k)} \| q^{\mathbb{Q}^n}) \rightarrow 0$ as $k \rightarrow \infty$. Then, under the stated conditions of uniform summability – in particular, when $1 / c \cdot q^{\mathbb{Q}^n} \leq \mathbb{P}^{(k)} \leq c \cdot q^{\mathbb{Q}^n}$ –

$$\lim_{k \rightarrow \infty} \frac{H(\mathbb{P}^{(k)})}{n} = H_q.$$ 

We shall come back to this further, considering the situation where $n \rightarrow \infty$.

B. Variable $\mathbb{Q}$-Types

If $H_q(x)$ is not constant, then we may use the length function $\ell_q(\mathbb{P}) = H_q(x)$ instead of $\ell_2$. Again, $A = a$, and (4) – applied to $\mathbb{Q}$ instead of $\mathbb{P}$ – leads to

$$\left| \frac{H(\mathbb{P})}{\ell_q(\mathbb{P})} - 1 \right| \leq 2\varepsilon + M_\varepsilon \phi(\delta) + \frac{C}{\delta} \sqrt{\frac{D(\mathbb{P} \| q^{\mathbb{Q}^n})}{\ell_q(\mathbb{P})}}.$$ 

(12)

If $H_q(x)$ varies between two positive bounds, then $\ell_q(\mathbb{P}) / \ell_q(\mathbb{P})$ varies between the same bounds, i.e., the two expected lengths are of the same order of magnitude.

Otherwise, e.g. if we want to stick to $\ell_2$, we need to discuss the quantity $\frac{D(\mathbb{P} \| q^{\mathbb{Q}^n})}{\ell_2(\mathbb{P})}$, since it remains present in the estimate of Theorem 1. We can use the following variant.

**Lemma 6:** Let $S$ be a cross section of $T$ (Def. 2), and let

$$a_S^* = \inf \left\{ \frac{H_q(x)}{\ell(x)} : x \in N \cap \mathbb{T}^S \right\}$$

and

$$A_S^* = \sup \left\{ \frac{H_q(x)}{\ell(x)} : x \in N \cap \mathbb{T}^S \right\}.$$

Then the term $\frac{D(\mathbb{P} \| q^{\mathbb{Q}^n})}{\ell_2(\mathbb{P})}$ in the estimate of Theorem 1 can be replaced with

$$\frac{A_S^* - a_S^*}{2} \cdot \frac{\| \mu_\mathbb{P} - \mu_\mathbb{Q} \|_1}{1 + (A - a) \max \left\{ \frac{\ell(\mathbb{T}^S)}{\ell} , \frac{\ell(\mathbb{T}^S)}{\ell} \right\}}.$$ 

This is useful, if $H_q(x) / \ell(x)$ is “almost constant” outside a small sub-tree. By this we mean that there is a cross section $S$ such that $A_S^* - a_S^*$ is small, and at the same time, $T$ is much larger than $\mathbb{T}^S$, so that also $\ell(\mathbb{T}^S) / \ell$ and $\ell(\mathbb{T}^S) / \ell$ are small. See below in §V. The proof of Lemma 6 is also deferred to the Appendix.

We conclude this section with an example which shows that the term $\frac{D(\mathbb{P} \| q^{\mathbb{Q}^n})}{\ell_2(\mathbb{P})}$ is indispensable in the general estimate of Theorem 1.

**Example 1:** We consider a tree $T = T(n)$ of height $n + 1$ whose root has 2 forward neighbours, so that the two branches $T_1$ and $T_2$ of height $n$ which are attached to $o$ have all inner nodes with forward degrees $d_1$ in $T_1$ and $d_2$ in $T_2$, respectively. Then $d_2 = d_2(T_1)\cup d_2(T_2)$. For $0 < \theta < 1$, we define

$$\mathbb{P}(\theta | o) = \mathbb{P}(\theta | o) = \left\{ \begin{array}{ll} \theta/d_1, & \text{if } v \in d_1(T_1), \\ (1 - \theta)/d_2, & \text{if } v \in d_2(T_2). \end{array} \right.$$ 

With $\ell = \ell_2$, we have $\ell(\mathbb{P}(\theta)) = n + 1$. It is easy to compute

$$H(\mathbb{P}(\theta)) = H(\theta, 1 - \theta) + n \left( \theta \log_2 d_1 + (1 - \theta) \log_2 d_2 \right),$$

where $H(\theta, 1 - \theta)$ is the entropy of the Bernoulli distribution with parameter $\theta$.

Now we take $\mathbb{P}_n = \mathbb{P}_n(\theta)$, where $\theta \neq 1/2$, and $\mathbb{Q}_n = \mathbb{P}_n,1/2$. Then it is also easy to compute

$$\mathbb{D}(\mathbb{P}_n \| \mathbb{Q}_n) = 1 - H(\theta, 1 - \theta)$$

As $n \rightarrow \infty$, we have that $\mathbb{D}(\mathbb{P}_n \| \mathbb{Q}_n)/n \rightarrow 0$, while

$$\left| \frac{H(\mathbb{P}_n)}{n} - \frac{H(\mathbb{Q}_n)}{n} \right| \rightarrow \left| (\theta - \frac{1}{2}) \log_2 d_1 + (1 - \theta - \frac{1}{2}) \log_2 d_2 \right|,$$

and the latter limit is non-zero for appropriate choices of $\theta$, $d_1$ and $d_2$. Thus, there can be no improved general estimate that can remove the term $\mathbb{D}(\mathbb{P} \| \mathbb{Q}^n)$, or replace it by something of significantly smaller order of magnitude. □

The last example can be seen in terms of two stationary stochastic processes on denumerable state spaces. We first toss a coin to decide which of the two processes to run, and then “forget” about the other one. The reference measure $\mathbb{Q}$ corresponds to a fair coin toss, while $\mathbb{P}$ corresponds to a biased one. Then one cannot compare their entropy rates just in terms
of their Kullback-Leibler divergence, as one sees from the above example.

V. TREES WITH INFINITE GEODESICS

Now, and only in this last section before the Appendix, we consider the situation where the tree does have infinite geodesics (non-backtracking paths). In their presence, let us describe the boundary \( \partial T \) of \( T \). It consists of eventual leaves (vertices \( v \neq o \) with no forward neighbour) plus the ends of \( T \). Each end \( w \) is represented by an infinite geodesic \( \pi(o,w) \) starting at \( o \). We can put a metric on \( \partial T = \text{int} T \cup \partial T \); for any pair of distinct points \( v, w \in \partial T \), their confluent \( v \wedge w \) is the last common vertex on the geodesics \( \pi(o,v) \) and \( \pi(o,w) \). Then the metric is

\[
d(v, w) = \begin{cases} 
0, & \text{if } v = w, \\
2^{-|v \wedge w|}, & \text{if } v \neq w.
\end{cases}
\]

Thus, a sequence \( (x_n) \) of vertices of \( T \) converges to an end \( w \), if \( |x_n \wedge w| \to \infty \). For more details in a context close to the present note, see e.g. [13, Ch. 9.B]. Note that in many typical examples, such as the infinite binary tree, \( \partial T \) is a Cantor set, whence uncountable.

For \( x \in \text{int} T \), the branch \( T_x \) and its boundary \( \partial T_x \) are defined as before. The boundary now may contain ends as well as leaves, and the sets of all \( \partial T_x \), \( x \in \text{int} T \), are the open (and also closed!) balls for the metric on \( \partial T \). Any probability measure \( P \) on \( \partial T \) is given via consistent definition of the values \( P(\partial T_x), x \in \text{int} T \). That is, we need to have \( P(\partial T) = 1 \) and

\[
\sum_{y \in N(x)} P(\partial T_y) = P(\partial T_x) \text{ for every } x \in \text{int} T.
\]

Example 2: The most typical class of examples arises from a stochastic process \( (X_n)_{n \geq 1} \) on a finite or countable state space \( \mathcal{S} \). In this case, for \( n \in \mathbb{N} \), let

\[
p_n(s_1, \ldots, s_n) = P[X_1 = s_1, \ldots, X_n = s_n] \quad (s_i \in \mathcal{S})
\]

be the joint distribution of \( (X_1, \ldots, X_n) \). Then we define (the vertex set of) the tree as

\[
T = \{s_1s_2 \cdots s_n : n \geq 0, \ s_i \in \mathcal{S}, \ p_n(s_1, \ldots, s_n) > 0\},
\]

For \( n = 0 \), this includes the empty word \( o \), for which \( p_0(o) = 1 \). The predecessor vertex of \( x = s_1 \cdots s_n \ (n \geq 1) \) is \( x^- = s_1 \cdots s_{n-1} \), and

\[
p(x | x^-) = \frac{P[X_1 = s_1, \ldots, X_{n-1} = s_{n-1}, X_n = s_n]}{P[X_1 = s_1, \ldots, X_{n-1} = s_{n-1}]}.
\]

while for \( x = s \in N(o) \), we have \( p(x | o) = P[X_1 = x] \). In this case, \( T \) has no leaves, the end space

\[
\partial T = \{s_1s_2s_3 \cdots : p_n(s_1, \ldots, s_n) > 0 \text{ for each } n\}
\]

is the space of active trajectories of our process, and \( P \) is the distribution of that process.\( \square \)

In general, one cannot speak directly about the entropy of \( P \) on \( \partial T \). Take any section \( S \) of \( T \) and consider \( H(P_S) \), as in Definition 2 and the subsequent lines. We restrict attention to

the following sequence of sections with respect to the standard length function:

\[
S(n) = \{x \in T : |x| = n\} \cup \{v \in \partial T : |v| < n\}.
\]

The second part appears when \( T \) has leaves. Write

\[
P_n = P_{S(n)} \text{ and } Q_n = Q_{S(n)}
\]

and assume that \( H(P_n) \) and \( D(P_n \| Q_n) \) are finite (absolutely convergent) for each \( n \). Then one can speak about the quantity

\[
h = h(P, \ell) = \lim_{n \to \infty} \frac{H(P_n)}{\ell(P_n)}
\]

whenever the limit exists. This is a variant of the entropy rate. In particular, in the situation of Example 2, for the standard length function, \( h(P) = h(P, \ell_2) \) is indeed the classical entropy rate of the process \( (X_n) \) with distribution \( P \).

In any case, the relative entropy is well defined; see [7, §7.1]:

\[
D(P \| Q) = \lim_{n \to \infty} D(P_n \| Q_n) \tag{14}
\]

Indeed, one can apply the well-known log-sum inequality – see [4, Th. 2.7.1] – to verify that \( D(P_n \| Q_n) \) is increasing with \( n \). Of course, the limit may be infinite. For the proof of the next theorem, see once more the Appendix.

Theorem 2: Let \( T \) be a countable tree with infinite geodesics and \( P \) and \( Q \) two probability measures giving positive mass to the set of ends of \( T \) and satisfying the tightness requirement of Assumption 2 for the local entropies at the inner vertices. Assume furthermore that

(i) The length function \( \ell \) is comparable with \( \ell_2 \), that is

\[
0 < c_1 \leq \ell(x) \leq c_2 < \infty \text{ for suitable } c_1, c_2 \text{ and all } x \in \text{int} T,
\]

(ii) the limit \( h = \lim_{|x| \to \infty} H_q(x)/\ell(x) \) exists and is finite, and

(iii) for the relative entropies, \( \lim_{n \to \infty} D(P_n \| Q_n)/n = 0 \).

Then

\[
h(P, \ell) = h(Q, \ell) = h.
\]

Note that condition (iii) holds in the particular case when \( D(P \| Q) < \infty \).

Application 1: We take up Example 2, with the difference that we now write \( Q \) for the distribution on the trajectory space of our random process \( (X_n) \). We have identified that space with the space of ends of a rooted tree \( T \) which has no leaves. Take the natural length function \( \ell_2 \) and assume that \( H_q(x) = h \) is constant. Then the standard entropy rate \( h(Q) \) exists and coincides with \( h \).

Now consider a perturbation of \( (X_n) \), that is, another stochastic process \( (\tilde{X}_n) \) with overall distribution \( P \). In this situation, Theorem 2 yields that if \( D(P_n \| Q_n)/n \to 0 \), and in particular if \( D(P \| Q) < \infty \), then also the entropy rate of \( (\tilde{X}_n) \) exists and is equal to the same \( h \).

Typical cases where this can be used are as follows.

(a) As in (11), let \( (X_n) \) be a sequence of i.i.d. discrete random variables with common distribution \( q \). Then \( Q = q^\otimes N \) is the infinite product measure of countably many copies of \( q \), and the entropy rate is \( h(Q) = H(q) \). If we have a
perturbed process \((\tilde{X}_n)\) whose overall distribution \(\mathbb{P}\) satisfies \(D(\mathbb{P}_n\|q^{\otimes n})/n \to 0\) then also \(h(\mathbb{P}) = H(q)\). (When \(q()\) is finitely supported, this is covered by [1].)

The following specific example is related to the famous paper of Kakutani [8] on absolute continuity versus orthogonality of infinite product measures. We let \(q()\) be uniform distribution on \([1, \ldots, M]\) and \(Q = q^{\otimes \mathbb{N}}\). The associated tree is the \(M\)-ary tree, where each vertex has \(M\) successors. For \(a \in (0, 1)\), we define \(p_a()\) on \([1, \ldots, M]\) by

\[
p_a(1) = \frac{1 + a}{M}, \quad p_a(2) = \frac{1 - a}{M}, \quad \text{and} \quad p_a(s) = \frac{1}{M} \quad \text{for} \ s = 3, \ldots, M.
\]

Then

\[
D(p_a \| q) = H(q) - H(p_a) = \frac{f(a)}{M},
\]

where

\[
f(a) = (1 + a) \log_2(1 + a) + (1 - a) \log_2(1 - a).
\]

Note that \(f(a)/a^2 \to \ln 2\) as \(a \to 0\). Now take a sequence \((a_n)\) and the perturbed measure \(\mathbb{P} = \prod_{n=1}^{\infty} p_{a_n}\). Then

\[
D(\mathbb{P}_n\|Q_n) = H(Q_n) - H(\mathbb{P}_n) = \frac{f(a_1) + \cdots + f(a_n)}{M}
\]

If \(a_n \to 0\) then \(D(\mathbb{P}_n\|Q_n)/n \to 0\), whence \(h(\mathbb{P}) = H(Q) = \log_2 M\). In particular \(a_n = n^{-\beta}\) with \(\beta > 0\) then one sees that \(D(\mathbb{P}\|Q) < \infty\) precisely when \(\beta \geq 1/2\).

(b) Next, let \((X_n)\) be a Markov chain on a finite or countable state space \(\mathcal{S}\) with transition matrix \(\mathcal{Q} = (q(s'|s))_{s,s' \in \mathcal{S}}\) and initial distribution \(\nu\) on \(\mathcal{S}\). Denote once more by \(\mathcal{Q}\) – instead of \(\mathbb{P}\) as in Example 2 – the overall probability distribution of the process on its trajectory space, which we interpret as the boundary at infinity of the tree \(T\). In this case, as in (2) but with \(q_s\) instead of \(p_n\), we have

\[
q_n(s_1, \ldots, s_n) = q(s_n|s_{n-1}) \cdots q(s_2|s_1)\nu(s_1).
\]

In particular, for a node \(x = s_1 \cdots s_{n-1} s_n\) of the tree, with \(n \geq 2\), we have \(q(x|x^{-}) = q(s_n|s_{n-1})\).

Now suppose that the transition matrix is such that all its rows \(q(\cdot|x)\) have the same finite entropy \(h\). Then \(H(\mathcal{Q}) = h\). Again, if one has a perturbation \((\tilde{X}_n)\) of \((X_n)\) with overall distribution \(\mathbb{P}\) such that the tightness requirements (2) are met and \(D(\mathbb{P}_n\|Q_n)/n \to 0\), then also \(h(\mathbb{P}) = h\).

(c) A specific example concerning the last situation is that of a finite or infinite oriented graph with vertex set \(\mathcal{S}\) which is \(\mathbb{d}\)-regular, that is, every vertex has \(d\) outgoing edges. For our Markov chain, we take simple random walk which at each vertex moves with equal probability to one of those neighbours. Then \(h = \log_2 d\) and all of the above applies.

There are many further examples of Markov chains where \(H(q(\cdot|x))\) is constant, in particular those where all rows \(q(\cdot|x)\) of the transition matrix are permutations of each other.

If more generally \(H_q(x)\) is not constant, then one can use \(\ell(x) = H_q(x)\) as in (12), and when \(D(\mathbb{P}\|Q) < \infty\), or just \(D(\mathbb{P}_n\|Q_n)/n \to 0\), one gets that

\[
a \leq \liminf_{n \to \infty} \frac{H(\tilde{X}_1, \ldots, \tilde{X}_n)}{n} \leq \limsup_{n \to \infty} \frac{H(\tilde{X}_1, \ldots, \tilde{X}_n)}{n} \leq A,
\]

where \(a\) and \(A\) are as in Theorem 1. In this situation, it may also be of interest to note that if \(h(\mathbb{P}) = \lim_{n \to \infty} H(\tilde{X}_1, \ldots, \tilde{X}_n)/n\) exists, then

\[
h(\mathbb{P}) = \lim_{n \to \infty} \frac{\ell(q(\mathbb{P}_n))}{n},
\]

which follows once more from (12).

Remark 1: All results stated so far adapt easily to the situation where the reference measure \(Q\) is supported by the whole of \(\partial T\), while the probability measure \(\mathbb{P}\) is supported by a possibly strict subset. In this situation, one may have \(p(x|x^{-}) = 0\) for some edges of \(T\), and when \(\partial T\) consists only of leaves as in \[II - IV\], the node-average measure \(\mu_\mathbb{P}\) will only live on the induced sub-tree of \(T\) which is spanned by all vertices that can be reached from the root. In the presence of infinite geodesics, the situation is analogous, considering the measure \(\mu_{\mathbb{P}_n}\) on \(T^{S(n)}\).

The following is our final application.

Random Perturbations: We start once more with the setting of Example 2, that is, the infinite tree is as in (13), and its boundary is equipped with our reference probability measure \(Q\). At the internal vertices we have the transition probabilities \(q(\cdot|x)\) and a length function \(\ell(x)\).

\(Q\) describes the basic stochastic process which then is randomly perturbed: there is a second collection of transition probabilities \(q'(\cdot|x)\) at each internal vertex, supported by some or all forward neighbours of \(x\). We assume that both families of probability distributions \(q(\cdot|x)\) and \(q'(\cdot|x)\), \(x \in T\), are tight and that their entropies are uniformly summable; see Lemma 5 for a sufficient condition.

Now consider a sequence of random variables \(\delta_n, n \geq 0\), taking their values in the interval \([0, 1]\), defined on a separate probability space. Then a perturbed random process is given via the forward transition probabilities

\[
p(y|x) = (1 - \delta_n)q(y|x) + \delta_n q'(y|x), \quad \text{if} \ y \in S(n).
\]

Note that this defines random transition probabilities, so that one gets a randomised process: at time \(n\) – when \(x \in S(n)\) – the moving particle chooses to proceed according to the “wrong” forward transition probabilities \(q'(\cdot|x)\) with probability \(\delta_n\), while it follows the “correct” transition rule with the complementary probability. Write \(\mathbb{P}\) for the random probability measure on \(\partial T\) induced by \(p(\cdot|\cdot)\), and as before \(\mathbb{P}_n = \mathbb{P}_{S(n)}\).

The proof of the following is again contained in the Appendix.

**Theorem 3:** In the setting of (15) with tight families of forward transition probabilities \(q(\cdot)\) and \(q'(\cdot)\) as well as uniformly summable local entropies, assume that

(i) The length function \(\ell\) is comparable with \(\ell'\),

(ii) the limit \(h = \lim_{|x| \to \infty} H_q(x)/\ell(x)\) exists and is finite, and

(iii) there is \(D < \infty\) such that \(D_{q',q}(x) < D\) for all \(x \in T\).

Under these conditions,

\[
\text{if} \ \lim_{n \to \infty} E(\delta_n) = 0 \ \text{then} \ \lim_{n \to \infty} \frac{H(\mathbb{P}_n)}{\ell(\mathbb{P}_n)} = h \ \text{in probability.}
\]

Moreover, if \(\delta_n \to 0\) almost surely, then \(h(\mathbb{P}, \ell) = h\) almost surely, i.e., for almost every realisation of the sequence \((\delta_n)\).
Example 3: Suppose that $\delta_n$ is a Bernoulli random variable with $\mathbb{P}[\delta_n = 1] = 1/n^\beta$ and $\mathbb{P}[\delta_n = 0] = 1 - 1/n^\beta$, where $\beta > 0$. Then $E(\delta_n) \to 0$, so that we get convergence in probability in Theorem 3. If $\beta > 1$ then by the Borel-Cantelli Lemma we have with probability 1 that $\delta_n = 0$ for all but a (random) finite number of $n$. In this case, we get that $h(\mathbb{P}, \ell) = h(\mathbb{Q}, \ell)$ almost surely.

The general picture that we have in mind regarding random perturbations with $\delta_n \to 0$ is that $\mathbb{Q}$ relates to a given process which one wants to simulate. The simulation is then a process with overall distribution $\mathbb{P}$, and at each step, the error in the simulation is described via $q(\cdot|\cdot)$ and $\delta_n$ as in $(15)$. Then $\delta_n \to 0$ means that in the course of time, there is a learning effect, so that the simulation gets better and better.

APPENDIX

In this appendix, we present the proof details of Theorem 1, Lemma 6, as well as theorems 2 and 3. For the first proof, we need the following, always under the tightness assumptions of Theorem 1.

Lemma 7: For all $\epsilon > 0$, $0 < \delta \leq 1/2$ and $x \in \text{int} T$,

$$|H_p(x) - H_q(x)| \leq 2\epsilon + M_\epsilon \varphi(\delta) + \frac{2}{\epsilon \ln 2 \delta} \|p(\cdot|x) - q(\cdot|x)\|_1.$$

Proof: If $y \in N(x)$ is such that $|p(y|x) - q(y|x)| < \delta$, then Lemma 4 yields

$$|\varphi(p(y|x)) - \varphi(q(y|x))| \leq \varphi(\delta).$$

In view of this,

$$|H_p(x) - H_q(x)| \leq \sum_{y \in N(x)} |\varphi(p(y|x)) - \varphi(q(y|x))|$$

$$= \text{Sum}_1 + \text{Sum}_2$$

with

$$\text{Sum}_1 = \sum_{y:|p(y|x) - q(y|x)| < \delta} |\varphi(p(y|x)) - \varphi(q(y|x))|$$

$$\leq \sum_{y:|p(y|x) - q(y|x)| < \delta} \left( |\varphi(p(y|x))| + |\varphi(q(y|x))| \right)$$

$$+ \sum_{y \in N(x)} \varphi(\delta) < 2\epsilon + M_\epsilon \varphi(\delta).$$

Next,

$$\text{Sum}_2 = \sum_{y:|p(y|x) - q(y|x)| \geq \delta} |\varphi(p(y|x)) - \varphi(q(y|x))|$$

$$\leq \sum_{y:|p(y|x) - q(y|x)| \geq \delta} \frac{2 \max \varphi}{\delta} \delta$$

$$\leq \frac{2 \max \varphi}{\delta} \sum_{y:|p(y|x) - q(y|x)| \geq \delta} |p(y|x) - q(y|x)|$$

$$\leq \frac{2 \max \varphi}{\delta} \|p(\cdot|x) - q(\cdot|x)\|_1,$$

completing the proof.

Proof of Theorem 1: We start by observing that

$$\sum_{x \in \text{int} T} \frac{\mu_p(x)}{\ell(x)} = \frac{\ell_1(\mathbb{P})}{\ell(\mathbb{P})},$$

(16) Using Lemma 2, we write

$$\left| \frac{H_p(\mathbb{P})}{\ell(\mathbb{P})} - \frac{H_q(\mathbb{Q})}{\ell(\mathbb{Q})} \right| = \left| \sum_{x \in \text{int} T} \left( \frac{\mu_p(x)}{\ell(x)} \frac{H_p(x)}{\ell(x)} - \frac{\mu_q(x)}{\ell(x)} \frac{H_q(x)}{\ell(x)} \right) \right|$$

$$\leq \text{Sum}_1 + |\text{Sum}_2|,$$

where

$$\text{Sum}_1 = \sum_{x \in \text{int} T} \frac{\mu_p(x)}{\ell(x)} \left| H_p(x) - H_q(x) \right|$$

and

$$\text{Sum}_2 = \sum_{x \in \text{int} T} \frac{H_q(x)}{\ell(x)} \left( \mu_p(x) - \mu_q(x) \right).$$

For the following sequence of estimates to bound $\text{Sum}_1$, in (a) we use Lemma 7, in (b) equation $(16)$ and Lemma $3(ii)$, in (c) the Cauchy-Schwarz inequality, and in (d) once more $(16)$, as well as Proposition 1,

$$\text{Sum}_1 \leq \sum_{x \in \text{int} T} \frac{\mu_p(x)}{\ell(x)} \left( 2\epsilon + M_\epsilon \varphi(\delta) \right)$$

$$+ \sum_{x \in \text{int} T} \frac{\mu_p(x)}{\ell(x)} \left( 2\epsilon + M_\epsilon \varphi(\delta) \right)$$

$$+ \frac{C}{\delta} \sum_{x \in \text{int} T} \frac{\mu_p(x)}{\ell(x)} \frac{D_{p,q}(x)}{\ell(x)}$$

$$\leq \frac{\epsilon}{\delta} \frac{\ell_1(\mathbb{P})}{\ell(\mathbb{P})} \left( 2\epsilon + M_\epsilon \varphi(\delta) \right)$$

$$+ \frac{C}{\delta} \sqrt{\sum_{x \in \text{int} T} \frac{\mu_p(x)}{\ell(x)} \frac{\mu_p(x)}{\ell(x)} \frac{D_{p,q}(x)}{\ell(x)}}.$$

Next, by Lemma $3(i)$,

$$\sum_{x \in \text{int} T} \frac{H_q(x)}{\ell(x)} \left( \mu_p(x) - \mu_q(x) \right)$$

$$\leq \sum_{x: \mu_p(x) > \mu_q(x)} A(\mu_p(x) - \mu_q(x))$$

$$+ \sum_{x: \mu_p(x) < \mu_q(x)} a(\mu_p(x) - \mu_q(x))$$

$$= A - a \left\| \mu_p - \mu_q \right\|_1.$$

Exchanging the roles of $\mu_p$ and $\mu_q$, we get the same upper bound. This proves that $|\text{Sum}_2| \leq \frac{\epsilon}{2\delta} \left\| \mu_p - \mu_q \right\|_1$, concluding the proof.

At this point we remark that one might want to refrain from deriving an upper bound on $\text{Sum}_1$ that involves Pinsker’s inequality in step (b) of the above proof: instead one could work directly with $\text{Sum}_1$. However, recall from above that there are converses to Pinsker’s inequality, so that one cannot expect big improvements from such a change of proof strategy.

We next turn to the lemma which allows us to use asymptotic constantness of the length-normalised local entropies $H_q(x)/\ell(x)$ in Theorem 2.
Proof of Lemma 6: We need to modify the estimate of the term $|\text{Sum}_y|$ in the proof of Theorem 1. Consider the sets \( \text{Pos} = \{ x \in \text{int} T : \mu_P(x) > \mu_Q(x) \} \) and \( \text{Neg} = \{ x \in \text{int} T : \mu_P(x) > \mu_Q(x) \} \), as well as \( \text{Pos}_S = \text{Pos} \cap \text{int} T^3 \) and \( \text{Neg}_S = \text{Neg} \cap \text{int} T^3 \). We split
\[
\sum_{x \in \text{int} T} \frac{H_Q(x)}{\ell(x)} (\mu_P(x) - \mu_Q(x)) \leq \sum_{x \in \text{Pos}_S} A(\mu_P(x) - \mu_Q(x)) + \sum_{x \in \text{Neg}_S} a(\mu_P(x) - \mu_Q(x)) \\
+ \sum_{x \in \text{Pos}_S \cap \text{Neg}_S} A^*_S(\mu_P(x) - \mu_Q(x)) \\
+ \sum_{x \in \text{Neg}_S} a^*_S(\mu_P(x) - \mu_Q(x)) = A_S^* \sum_{x \in \text{Pos}} (\mu_P(x) - \mu_Q(x)) + a_S^* \sum_{x \in \text{Neg}} (\mu_P(x) - \mu_Q(x)) \\
+ (A - A_S^*) \mu_P(\text{Pos}_S) + (a_S^* - a) \mu_Q(\text{Neg}_S).
\]
Now, by the definition (3) of \( \mu_P \),
\[
\mu_P(\text{Pos}_S) = \frac{\ell(P)}{\ell(P)} \mu_P(\text{Pos}_S) \leq \frac{\ell(P)}{\ell(P)} \mu_P(\text{Pos}_S),
\]
and analogously \( \mu_Q(\text{Neg}_S) \leq \frac{\ell(Q)}{\ell(Q)} \). This and an exchange of the roles of \( P \) and \( Q \) lead to the stated upper bound. □

At last, we come to the proofs of the two theorems of Section V.

Proof of Theorem 2: Write \( \tilde{c}_nT \) for the space of ends of \( T \). By assumption, \( \tilde{c}_n \) for \( n \to \infty \). Therefore
\[
\ell\left(\frac{\ell(P)}{\ell(P)}\right) \geq n \tilde{c}_n \to \infty, \quad n \to \infty. \quad (17)
\]
The same applies to \( Q \). Next, we use Lemma 6. Start with an arbitrary \( \epsilon > 0 \). By assumption (ii) there is an index \( k \) such that \( A_S^*(k) - A_S^*(k) < \epsilon \). Take \( n > k \) and consider the sub-tree \( T^{S(n)} \) with \( P \) and \( Q \). Note that \( L = \ell\left(\frac{\ell(P)}{\ell(P)}\right) \leq c_2 \) by assumption (i). Next,
\[
\delta_n := \left(\frac{D}{\ell(P)}\right)^{1/4} \to 0 \quad n \to \infty
\]
by assumption (i) and (iii) together with (17).

Now the upper bound of Theorem 1, combined with the variant provided by Lemma 6, yields
\[
\left| \frac{H(P_n)}{\ell(P_n)} - \frac{H(Q_n)}{\ell(Q_n)} \right| < c_2 (2\epsilon + M_c \varphi(\delta_n)) + C \sqrt{c_2} \delta_n \\
+ \epsilon + (A - a) \max \left\{ \frac{\ell(S(k))}{\ell(P_n)}, \frac{\ell(Q(S(k)))}{\ell(Q_n)} \right\}.
\]
If \( n \to \infty \), again using assumption (i) together with (17) to control the last maximum, one sees that
\[
\limsup_{n \to \infty} \left| \frac{H(P_n)}{\ell(P_n)} - \frac{H(Q_n)}{\ell(Q_n)} \right| \leq (2c_2 + 1)\epsilon,
\]
because all other terms tend to 0. We are left with showing that \( H(Q_n)/\ell(Q_n) \to h \). This is standard: with \( \epsilon \) and \( k \) as above,
\[
\left| \frac{H(Q_n)}{\ell(Q_n)} - h \right| \leq \sum_{x \in \text{int} T^{S(n)}} \left| \frac{H_Q(x)}{\ell(x)} - h \right| \mu_Q(x) \\
< \sum_{x \in \text{int} T^{S(k)}} (A + h) \mu_Q(x) + \sum_{x \in \text{int} T^{S(n)} \setminus T^{S(k)}} \epsilon \mu_Q(x) \\
\leq (A + h) \frac{\ell(Q(k))}{\ell(Q_n)} + \epsilon.
\]
This becomes smaller than \( 2\epsilon \) when \( n \) is large enough. □

Finally, we prove the result on random perturbations.

Proof of Theorem 3: First of all, we need to verify that the (random) local entropies \( H_P(x) \) are uniformly summable. By tightness of \( q(x) \) and \( q'(x) \), there is \( M \) such that for each \( x \) there is a set \( N(x) \subseteq N(x) \) with \( |N(x)| \leq M \) such that
\[
\max \{q(y|x), q'(y|x)\} \leq 1/(2\epsilon) \quad \text{for all} \quad y \in N(x) \setminus N(x).
\]
Using Lemma 4(iv),
\[
\sum_{y \in N(x) \setminus N(x)} \varphi(p(y|x)) \leq 2 \sum_{y \in N(x) \setminus N(x)} \left( \varphi(q(y|x)) + \varphi(q'(y|x)) \right)
\]
As \( x \) varies, the right hand sides are uniformly summable along with \( H_q(x) \) and \( H_q'(x) \).

At this point, with the same line of reasoning as in the proof of Theorem 2, what remains is to show that \( D(P_n \| Q_n)/\ell(P_n) \to 0 \) in probability, resp. almost surely. For this purpose, we first take \( x \in S(n) \), and use convexity, see [4, Th. 2.7.2]:
\[
D(x) = D(x) + \delta_n q(x) \leq D(x) + \delta_n q(x) \leq \delta_n D(x).
\]
Now we combine the last inequality with Proposition 1 and the Definition (3) of \( \mu_P \),
\[
\frac{D(P_n \| Q_n)}{\ell(P_n)} \leq \frac{D}{\ell(P_n)} \sum_{k=0}^{n} \frac{\mu_P(x)}{\ell(x)} \delta_k D(x) \leq \frac{D}{\ell(P_n)} \sum_{k=0}^{n} \delta_k \leq \frac{D}{c_1} \sum_{k=0}^{n} \delta_k.
\]
The last inequality follows from assumption (i). If \( \mathbb{E}(\delta_n) \to 0 \), resp. \( \delta_n \to 0 \) almost surely then
\[
\frac{D}{n} \sum_{k=0}^{n} \delta_k \to 0
\]
in probability, resp. almost surely. □
REFERENCES

[1] G. Böcherer and R. A. Amjad, “Informational divergence and entropy rate on rooted trees with probabilities,” in Proc. IEEE Int. Symp. Inf. Theory (ISIT), Jun./Jul. 2014, pp. 176–180, doi: 10.1109/ISIT.2014.6874818.

[2] G. Ciuperca and V. V. Girardin, “On the estimation of the entropy rate of finite Markov chains,” in Proc. 11th Int. Symp. ASMDA, Brest, France, 2005, pp. 1110–1117.

[3] G. Ciuperca, V. Girardin, and L. Lhote, “Computation and estimation of generalized entropy rates for denumerable Markov chains,” IEEE Trans. Inf. Theory, vol. 57, no. 7, pp. 4026–4034, Jul. 2011.

[4] T. M. Cover and J. A. Thomas, Elements of Information Theory, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.

[5] I. Csiszár, “Information-type measures of difference of probability distributions and indirect observations,” Studia Sci. Math. Hungarica, vol. 2, pp. 299–318, 1967.

[6] I. Csiszár and Z. Talata, “Context tree estimation for not necessarily finite memory processes, via BIC and MDL,” IEEE Trans. Inf. Theory, vol. 52, no. 3, pp. 1007–1016, Mar. 2006.

[7] R. M. Gray, Entropy and Information Theory, 2nd ed. New York, NY, USA: Springer, 2011.

[8] S. Kakutani, “On equivalence of infinite product measures,” Ann. Math., vol. 49, no. 1, pp. 214–224, 1948.

[9] J. H. B. Kemperman, “On the optimum rate of transmitting information,” Ann. Math. Statist., vol. 40, no. 6, pp. 2156–2177, 1969.

[10] S. Kullback, “A lower bound for discrimination information in terms of variation (corresp.),” IEEE Trans. Inf. Theory, vol. IT-13, no. 1, pp. 126–127, Jan. 1967.

[11] R. A. Rueppel and J. I. Massey, et al., “Leaf-average node-sum interchanges in rooted trees with applications,” in Communications and Cryptography: Two Sides of One Tapestry, R. E. Blahut, Ed. Dordrecht, The Netherlands: Kluwer, 1994, pp. 343–355.

[12] I. Sason and S. Verdú, “f-Divergence inequalities,” IEEE Trans. Inf. Theory, vol. 62, no. 11, pp. 5973–6006, Nov. 2016.

[13] W. Woess, Denumerable Markov Chains: Generating Functions, Boundary Theory, Random Walks on Trees, Zürich, Switzerland: European Mathematical Society, 2009.

[14] G. H. Yari and Z. Nikooravesh, “Estimation of the entropy rate of Ergodic Markov chains,” J. Iranian Statist. Soc., vol. 11, no. 1, pp. 75–83, 2012.

Thomas Hirschler received his diploma degree in 1993 and his PhD (Dr.techn.) in mechanical engineering in 1997, both at Graz University of Technology, Austria, and worked as an assistant professor at the Institute of Fluid Mechanics and Heat Transfer. From 2003 to 2008 he studied Mathematics at the University of Graz. In 2013, he moved to the Institute of Mathematical Structure Theory, now Institute of Discrete Mathematics, at Graz University of Technology. His mathematical interests are in probability theory and in random processes on graphs.

Wolfgang Woess received his diploma ( ≡ master) degree in Mathematics from Technische Universität Wien (Vienna) in 1978 and his PhD in mathematics (Dr. rer. nat.) from the University of Salzburg, Austria, in 1980. From 1982 to 1988 he served as an assistant professor at Montanuniversität Leoben, Austria, interrupted by a PostDoc year at University of Rome-I in 1984/85. In 1986, he obtained his “Habilitation”. In 1988, he moved to an associate professor position at the University of Milan, Italy, where he became a full professor of probability in 1995. Since 1999, he is a full professor of mathematics at Graz University of Technology.

His teaching includes calculus for electrical engineers, various courses for students of mathematics, as well as stochastics + information theory for mathematicians and computer scientists.

His research interests are in random walks on infinite graphs and groups, which is also the title of a research monograph which he published in 2000. Further themes of research are Markov chains and discrete potential theory, random processes on mixed discrete/non-discrete state spaces, as well as infinite graphs and geometric group theory, and formal languages.

The present paper was originally motivated by a talk by G. Böcherer at the Signal Processing and Speech Communication Laboratory of Graz University of Technology in 2013.