ISRAEL JOURNAL OF MATHEMATICS, 75 (1991), 289–328

CODING MARKOV CHAINS FROM THE PAST

BY

JAMES GARY PROPP*

Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139-4307, USA

ABSTRACT

It is shown that a mixing Markov chain is a unilateral or one-sided factor of every ergodic process of equal or greater entropy. This extends the work of Sinai, who showed that the result holds for independent processes, and the work of Ornstein and Weiss, who showed that the result holds for mixing Markov chains in which all transition probabilities are positive. The proof exploits the Rothstein-Burton joinings-space formulation of Ornstein's isomorphism theory, and uses a random coding argument.

1. Introduction

Let $(X, T, \mathcal{A}, \mu)$ be a dynamical system, with $\mathcal{A}$ a σ-algebra of subsets of X, μ a non-atomic probability measure defined on $\mathcal{A}$, and T a bimeasurable map of X onto itself that preserves μ. Taking a generating partition P and viewing the components of P as corresponding to "symbols", we may regard X as the set of doubly-infinite sequences of elements of the finite alphabet P, with T the shift-map; the symbolic stochastic process $(X, T, \mathcal{A}, \mu, P)$ may thus be abbreviated P, μ.

A factor map φ from the process P, m_P to the process Q, m_Q is said to be unilateral if $\varphi^{-1}(Q)$ is $\bigvee_{n=0}^{\infty} T^{-n}P$-measurable. Sinai showed in 1964 [15] that if the independent process Q, m_Q has entropy less than or equal to that of the ergodic process P, m_P then Q, m_Q is a unilateral factor of P, m_P. In 1975, Ornstein and Weiss [9] gave a different demonstration of this result, and also claimed, with a brief indication of proof, that the hypothesis on Q, m_Q could be

*Partially supported by an NSF Graduate Fellowship, an NSF Postdoctoral Fellowship, and NSF Grant #DMS 84-03182 during the writing of this article.
Received February 1, 1990 and in revised form May 13, 1991

weakened to the assumption that Q, m_Q is a Markov chain *with all transition probabilities positive*. Note that such a chain is mixing.

Here we define a class of processes called "unilaterally finitely determined" or UFD processes, containing *all* mixing Markov chains with finitely many states (not just those with positive transition probabilities), and we show that every UFD process is a unilateral factor of every ergodic process of equal or greater entropy. The methods of proof also give an analogous result for non-mixing Markov chains.

The main theorem of this article can alternatively be construed as a statement about non-invertible measure-preserving transformations; in this setting, it says that the one-sided mixing Markov processes are "universal factors" just as in the two-sided theory.

Our style of proof is heavily influenced by the work of Arthur Rothstein and Robert Burton [13] on the "joinings" viewpoint in ergodic theory. Every factor map $\varphi : X \to Y$ from one system to another determines a joining of the two systems supported on the graph of the factor map; given such a "graph joining", it is easy to recover the map that gave rise to it. There is a canonical topology on the space of joinings, and the way Rothstein and Burton construct graph joinings is by taking limits of joinings that are not themselves graph joinings but become increasingly concentrated in the Y-direction. Imitating their style of proof, we will employ a Baire category argument to show that under suitable hypotheses, the set of joinings that correspond to unilateral factor maps is not only non-empty but actually *dense* in a certain natural subspace of the set of joinings (namely the space of "pre-unilateral" ergodic joinings).

For a more discursive exposition of the methods used here, see this author's [11], an updated edition of which is now available. The major substantive difference between the theorem proved there and the one proved here is that in the earlier work, the process P, m_P was required to be an i.i.d. process. Here that restriction is removed, at the expense of introducing an extra level of complexity in the proof of the Copying Lemma (see section 3).

The rest of the article is organized as follows. In the remainder of this section, we introduce terminology and notation. Section 2 sets up the framework in which the proof takes place. Sections 3, 4, and 5 are given to the proof of the theorem. Section 6 contains miscellaneous remarks on extensions of the theorem and related results.

For the most part we adhere to the conventions of [1] (see chapters 1, 8, and 10) and [6] (see chapters 1, 2, and 3).

Let $(X, T, \mathcal{A}, \mu)$ be a dynamical system. We may assume $(X, \mathcal{A}, \mu)$ is measure-

theoretically equivalent (modulo sets of measure 0) to the interval $[0,1]$ with measure defined on all Borel sets.

Suppose now that $P = \{P_i : 1 \le i \le r\}$ is a finite (labeled) partition on $(X, T, \mathcal{A}, \mu)$. P and μ determine the distribution vector

$$\mathrm{dist}_\mu(P) = (\mu(P_1), \mu(P_2), \ldots, \mu(P_r)) ;$$

we put the L^1-norm on the set of such vectors, so that for example if P and Q are partitions of spaces (X, μ) and (Y, ν) with $\#(P) = \#(Q) = r$,

$$|\mathrm{dist}_\mu(P) - \mathrm{dist}_\nu(Q)| = \sum_{i=1}^{r} |\mu(P_i) - \nu(Q_i)| .$$

We will sometimes write $P_{(i)}$ instead of P_i, to clearly distinguish the label i from the time-indices introduced below. If P and Q are partitions of a space X equipped with two measures μ and ν, and if P refines Q, then

$$|\mathrm{dist}_\mu(P) - \mathrm{dist}_\nu(P)| \ge |\mathrm{dist}_\mu(Q) - \mathrm{dist}_\nu(Q)| .$$

If P and Q are partitions of the same space satisfying $\#(P) = \#(Q) = r$, define their symmetric difference as the set

$$P \triangle Q = \bigcup_{i=1}^{r} P_i \triangle Q_i .$$

If P and Q are partitions of different spaces $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$, let $\overline{P} = \{\overline{P}_i = P_i \times Y\}$, $\overline{Q} = \{\overline{Q}_i = X \times Q_i\}$; if $\#(P) = \#(Q)$ (and μ, ν are non-atomic measures), then we have

$$\inf_\pi \pi(\overline{P} \triangle \overline{Q}) = \frac{1}{2} \left| \mathrm{dist}_\mu(P) - \mathrm{dist}_\nu(Q) \right| ,$$

where the infimum is taken over all measures π on $X \times Y$ with $\pi(\overline{P}_i) = \mu(P_i)$, $\pi(\overline{Q}_i) = \nu(Q_i)$ for all i.

If A is a measurable subset of X, we let $P \mid A$ denote the restriction of the partition P to the set A (namely $\{P_i \cap A\}$), and if $\mu(A) > 0$ we let $\mu \mid A$ or μ_A denote the conditioning of the measure μ on the set A. More generally, given any measurable partition of the space, we may disintegrate the measure μ into measures μ_α where α ranges over the atoms of the measurable partition. (Here, as hereafter, an "atom" of a measurable partition or σ-algebra is a measurable set none of whose non-empty proper subsets are measurable.)

For $x \in X$, we let $P(x)$ denote the component of P that contains x. We define

$$P^n = T^{-n}P = \{T^{-n}P_i\}$$

("the partition P shifted to time n"); it has the property that $P^n(x) = P(T^n x)$. We use the convenient abbreviation

$$P_a^b = \bigvee_{n=a}^{b} P^n .$$

It is convenient to shift back and forth between regarding $P_a^b(x)$ as an atom (i.e. component) of the partition P_a^b and as a string of $b-a+1$ symbols in the P-alphabet; when we wish to take the latter view, we will call $P_a^b(x)$ a name instead of an atom. It is also sometimes useful to think of P_a^b as a finite σ-algebra.

When a is $-\infty$ or b is ∞, we use the same definition of P_a^b as above, only now it is to be understood that the right hand side denotes a σ-algebra, not a partition. Two σ-algebras of special importance are $P_{-\infty}^{0}$ (the **past** of P up to time 0) and P_1^∞ (the **future** of P from time 1 onward).

Let A be a measurable set, P, Q partitions of the space into measurable sets, and $\mathcal{B}, \mathcal{C}$ sub-σ-algebras of $\mathcal{A}$. We write: $A \in P$ if A is a component of P; $A \subset P$ ("A is P-measurable") if A is a union of components of P; $P \subset Q$ if every component of P is Q-measurable; $P \subset \mathcal{B}$ if every component of P is $\mathcal{B}$-measurable; and $\mathcal{B} \subset \mathcal{C}$ if every $\mathcal{B}$-measurable set is $\mathcal{C}$-measurable. As usual, all statements made about measurability of sets and functions are to be interpreted modulo sets of measure 0.

If the set A is a subset of a single component of the partition P, we write $P(A)$ to denote the component of P containing A.

A **joining** of two systems $(X, T, \mathcal{A}, m_P)$, $(Y, S, \mathcal{B}, m_Q)$ is a dynamical system $(X \times Y, T \times S, \mathcal{A} \times \mathcal{B}, \pi)$ where the projection of π on $\mathcal{A}$ (the "first marginal") is m_P, the projection of π on $\mathcal{B}$ (the "second marginal") is m_Q, and π is invariant under the product action $T \times S$ on $X \times Y$. (If the last condition fails we call the system $(X \times Y, T \times S, \mathcal{A} \times \mathcal{B}, \pi)$ a **non-stationary joining**; if only the first marginal is what it is supposed to be, we call the system a **half-joining**.) Let $\overline{\mathcal{A}}$, $\overline{\mathcal{B}}$ be the lifts of $\mathcal{A}$, $\mathcal{B}$ to $X \times Y$.

Every factor map φ determines a joining π that is concentrated on the graph $\{(x, \varphi(x)) : x \in X\}$, for it suffices to define π on rectangles, and we can do this by putting $\pi(A \times B) = m_P\{x \in A : \varphi(x) \in B\}$. Such measures π are characterized by the property that $\overline{\mathcal{B}} \subset \overline{\mathcal{A}}$ modulo π; that is, every $\overline{\mathcal{B}}$-measurable

set differs from some $\overline{\mathcal{A}}$-measurable set by a π-null set. We call such joinings **graph joinings.**

One very important example of a joining is the independent (or direct product) joining $(X \times Y, T \times S, \mathcal{A} \times \mathcal{B}, \mu \times \nu)$, which gives measure $\mu(A)\nu(B)$ to each rectangle $A \times B$. More generally, two systems with a common factor have a "conditionally independent joining" over that factor (see [3], pages 110-115). If $(X_i, T_i, \mathcal{A}_i, \mu_i)$ (for $i = 1, 2$) are systems with a common factor $(Y, S, \mathcal{B}, \nu)$, the conditionally independent joining $\pi = \mu_1 \times_\nu \mu_2$ has the property that $\overline{\mathcal{A}_1} \perp \overline{\mathcal{A}_2} \mid \overline{\mathcal{B}}$ (modulo π).

All the dynamical systems considered here are symbolic dynamical systems, for which it is clear what the transformation is (namely, the left-shift); we use the same symbol T to denote the left-shift in all cases.

We will not restate the definition and basic properties of entropy, but we will remind the reader of the key fact that if $\mathcal{B}_1 \subset \mathcal{B}_2 \subset \mathcal{B}_3 \subset \cdots$ are nested σ-algebras with limit $\mathcal{B}$, then

$$H(P \mid \mathcal{B}_n) \searrow H(P \mid \mathcal{B}) .$$

The **conditional mutual information** between a partition P and a σ-algebra $\mathcal{B}$ relative to another σ-algebra $\mathcal{C}$ is

$$I(P; \mathcal{B} \mid \mathcal{C}) = H(P \mid \mathcal{C}) - H(P \mid \mathcal{B} \vee \mathcal{C}) ;$$

it vanishes if and only if P and $\mathcal{B}$, when restricted to atoms of $\mathcal{C}$, are independent for almost all atoms. Either $\mathcal{B}$ or $\mathcal{C}$ may be a partition rather than a σ-algebra; if $\mathcal{B}$ is a partition Q, then we have the symmetry relation $I(P; Q \mid \mathcal{C}) = I(Q; P \mid \mathcal{C})$. If $\mathcal{C}$ is trivial, we drop the word "conditional" and write $I(P; \mathcal{B})$.

All of these information-theoretic quantities implicitly depend on the probability measure being used; when we wish to stress this, we will use notations such as $H_\mu(P \mid \mathcal{B})$ or $I_\nu(P; Q)$. In the case of conditioned measures μ_A, it is inconvenient to write H_{μ_A}, so we will often write H_A instead, to indicate that we are conditioning μ on the set A.

If $\mathcal{B}_1$, $\mathcal{B}_2$, and $\mathcal{C}$ are σ-algebras on a space, we write $\mathcal{B}_1 \perp \mathcal{B}_2 \mid \mathcal{C}$ to signify that on almost every atom of $\mathcal{C}$, the σ-algebras $\mathcal{B}_1$ and $\mathcal{B}_2$ are independent (with respect to the implicit measure). We assume that the reader is comfortable with assertions like the following: for all σ-algebras $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$, if $\mathcal{A} \perp \mathcal{C} \mid \mathcal{D}$ and $\mathcal{A} \perp \mathcal{B} \mid \mathcal{C} \vee \mathcal{D}$ then $\mathcal{A} \perp \mathcal{B} \vee \mathcal{C} \mid \mathcal{D}$, and conversely.

2. Preliminaries

Suppose P, m_P and Q, m_Q are symbolic processes on measure spaces X and Y. A **unilateral factor map** from P, m_P to Q, m_Q is a measure-preserving map $\varphi : X \to Y$ that commutes with the shift and has the property that $\varphi^{-1}(Q)$ is $P_{-\infty}^{0}$-measurable. A **unilateral graph joining** of P, m_P and Q, m_Q is an invariant measure on $X \times Y$ that projects to m_P on X and m_Q on Y, with the property that $\overline{Q}$ is $\overline{P}_{-\infty}^{0}$-measurable.

CLAIM 1: *The following conditions on processes P, m_P and Q, m_Q are equivalent: (a) there exists a unilateral factor map φ from P, m_P to Q, m_Q; (b) there exists a unilateral graph joining μ of P, m_P and Q, m_Q; (c) there exists a partition $\tilde{Q} \subset P_{-\infty}^{0}$ in the past of the P-process such that the induced sub-process $\tilde{Q}, \tilde{m}_Q$ has the same law as Q, m_Q.*

Proof: (a) $\Rightarrow$ (b): Define μ on $X \times Y$ by putting $\mu(A \times B) = m_P(A \cap \varphi^{-1}(B))$ for every measurable rectangle $A \times B$ and extending to the full σ-algebra.

(b) $\Rightarrow$ (c): For all $1 \leq i \leq \#(Q)$, $\overline{Q}_{(i)}$ is $\overline{P}_{-\infty}^{0}$-measurable modulo μ, so there exists a set $\tilde{Q}_{(i)}$ in $P_{-\infty}^{0}$ such that the corresponding set in $\overline{P}_{-\infty}^{0}$ differs from $\overline{Q}_{(i)}$ by a set of μ-measure 0.

(c) $\Rightarrow$ (a): There exists a $P_{-\infty}^{0}$-measurable function $f : X \to \{1, \ldots, \#(Q)\}$ such that for almost all (x, y) in $X \times Y$ (modulo μ), $y \in \overline{Q}_{(f(x))}$. Let $\varphi(x)$ denote the point $(\ldots, f(T^{-1}x), f(x), f(Tx), \ldots)$ in the Q-process. □

Note that the condition $\overline{Q} \subset \overline{P}_{-\infty}^{0}$ is equivalent to the seemingly stronger condition $\overline{Q}_{-\infty}^{0} \subset \overline{P}_{-\infty}^{0}$.

Henceforth, we shall leave the over-bars in $\overline{P}$, $\overline{Q}$, $\overline{\mathcal{A}}$, $\overline{\mathcal{B}}$ tacit.

We call a joining μ of P, m_P and Q, m_Q a **pre-unilateral joining** if

$$P_1^{\infty} \perp Q_{-\infty}^{0} \mid P_{-\infty}^{0}$$

modulo μ. Colloquially, this amounts to the assertion that the future of P contains no information about the past of Q that is not contained in the past of P, and reciprocally, the past of Q contains no information about the future of P that is not contained in the past of P. If we wish to emphasize the asymmetric roles played by the two processes, we will say that μ is "pre-unilateral from $P_{-\infty}^{0}$ to $Q_{-\infty}^{0}$".

CLAIM 2: *A joining is a unilateral graph joining if and only if it is both a graph joining and a pre-unilateral joining.*

Proof: Suppose μ is a joining of P, m_P and Q, m_Q. If μ is a unilateral graph joining, then $Q \subset P_{-\infty}^{0}$, so that a fortiori $Q \subset P_{-\infty}^{\infty}$; thus μ is a graph joining. Also,

since $Q \subset P_{-\infty}^{0}$ we also have $Q_{-\infty}^{0} \subset P_{-\infty}^{0}$, which implies $P_1^{\infty} \perp Q_{-\infty}^{0} \mid P_{-\infty}^{0}$; thus μ is also a pre-unilateral joining. Conversely, suppose μ is a graph joining that happens to be pre-unilateral. Since μ is a graph-joining,

$$Q \subset P_1^{\infty} \vee P_{-\infty}^{0} \ (= P_{-\infty}^{\infty}) \ .$$

Also, since μ is pre-unilateral, $Q_{-\infty}^{0} \perp P_1^{\infty} \mid P_{-\infty}^{0}$, so that

$$Q \ \perp \ P_1^{\infty} \ \mid \ P_{-\infty}^{0} \ .$$

These two conditions together imply

$$Q \subset P_{-\infty}^{0} \ ,$$

so that μ is a unilateral graph joining. □

CLAIM 3: *A joining μ is pre-unilateral if and only if it satisfies the seemingly weaker condition $P^1 \perp Q_{-\infty}^{0} \mid P_{-\infty}^{0}$.*

Proof: One direction is trivial. To prove the other, assume that

$$P^1 \perp Q_{-\infty}^{0} \mid P_{-\infty}^{0};$$

we will use induction to show that $P_1^n \perp Q_{-\infty}^{0} \mid P_{-\infty}^{0}$ for all n. Suppose that

$$P_1^k \ \perp \ Q_{-\infty}^{0} \ \mid \ P_{-\infty}^{0} \ . \tag{1}$$

Since $P^1 \perp Q_{-\infty}^{0} \mid P_{-\infty}^{0}$, stationarity implies $P^{k+1} \perp Q_{-\infty}^{k} \mid P_{-\infty}^{k}$, i.e.,

$$P^{k+1} \ \perp \ Q_{-\infty}^{0} \vee Q_1^k \ \mid \ P_{-\infty}^{0} \vee P_1^k \ .$$

It follows a fortiori that

$$P^{k+1} \ \perp \ Q_{-\infty}^{0} \ \mid \ P_{-\infty}^{0} \vee P_1^k \ . \tag{2}$$

Combining (1) and (2), we get

$$P_1^k \vee P^{k+1} \ \perp \ Q_{-\infty}^{0} \ \mid \ P_{-\infty}^{0} \ ,$$

so that $P_1^{k+1} \perp Q_{-\infty}^{0} \mid P_{-\infty}^{0}$. Hence by induction $P_1^n \perp Q_{-\infty}^{0} \mid P_{-\infty}^{0}$ for all n. Now send $n \to \infty$. □

Remark: The condition $P^1 \perp Q_{-\infty}^0 \mid P_{-\infty}^0$ is equivalent to either of the conditions

$$I(P^1; Q_{-\infty}^0 \mid P_{-\infty}^0) = 0$$

and

$$H(P^1 \mid P_{-\infty}^0 \vee Q_{-\infty}^0) = H(P^1 \mid P_{-\infty}^0) = h(P) .$$

CLAIM 4: *If $Q \subset P_{-\infty}^\infty$ (modulo μ) and $h(Q) = 0$, then $Q \subset P_{-\infty}^0$ (modulo μ). Therefore every zero-entropy factor of a process is a unilateral factor.*

Proof: Since Q is a zero-entropy process, $H(Q^1 \mid Q_{-\infty}^0) = 0$, and $Q^1 \subset Q_{-\infty}^0$ modulo μ. The chain of inequalities

$$\begin{aligned} h(P \vee Q) &= H((P \vee Q)^1 \mid (P \vee Q)_{-\infty}^0) \\ &\leq H(P^1 \mid (P \vee Q)_{-\infty}^0) + H(Q^1 \mid (P \vee Q)_{-\infty}^0) \\ &= H(P^1 \mid (P \vee Q)_{-\infty}^0) \\ &\quad \text{(since } \quad Q^1 \subset Q_{-\infty}^0 \subset (P \vee Q)_{-\infty}^0) \\ &\leq H(P^1 \mid P_{-\infty}^0) \\ &= h(P) \\ &\leq h(P \vee Q) \end{aligned}$$

implies that $H(P^1 \mid (P \vee Q)_{-\infty}^0) = H(P^1 \mid P_{-\infty}^0)$, so that $P^1 \perp Q_{-\infty}^0 \mid P_{-\infty}^0$; hence (by Claim 3) μ is a pre-unilateral joining. Since μ is also a graph joining, Claim 2 implies that μ is a unilateral graph joining. The second sentence of Claim 4 follows from the first, using the equivalence between joinings and factor maps set forth in Claim 1. □

If μ and ν are probability measures on a σ-algebra $\mathcal{A}$ with countable generating sub-algebra $\{A_1, A_2, \dots\}$, then $\mu = \nu$ if and only if $\mu(A_k) = \nu(A_k)$ for all k. Moreover, every countably-additive non-negative set-function on $\{A_k\}$ that assigns the value 1 to the space as a whole extends to a probability measure on all of $\mathcal{A}$ (see "The Extension Theorem," pages 219-225 in [14]). Define

$$d(\mu, \nu) = \sum_{k=1}^{\infty} \frac{|\mu(A_k) - \nu(A_k)|}{2^k} .$$

This gives a metric topology on the space of probability measures on $\mathcal{A}$, with the property that $\mu_n \to \nu$ if and only if $\mu_n(A_k) \to \nu(A_k)$ for all k. For our purposes, the collection $\{A_k\}$ will be the algebra of cylinder sets in some process,

and we will call the topology determined by $\{A_k\}$ (via the function $d(\cdot,\cdot)$) the **distribution topology** on the set of measures. This topology is generated by balls of the form

$$\{\mu : |\mathrm{dist}_\mu R - \mathrm{dist}_{\mu_0} R| < \epsilon\}$$

where $\epsilon > 0$ and R is some partition into cylinder sets.

When $\{A_k\}$ is the cylinder algebra of a sequence space with finitely many symbols, the compactness of the sequence space under the usual product topology guarantees that no infinite disjoint union of non-empty cylinder sets can be a cylinder set, so that every finitely additive set function on $\{A_k\}$ is automatically countably additive. Hence we obtain a one-to-one correspondence between process-measures on the σ-algebra generated by $\{A_k\}$ and finitely additive set functions on $\{A_k\}$ taking values in $[0,1]$.

CLAIM 5: *In the distribution topology, the set of invariant probability measures is a compact separable metric space.*

Proof: Under the injective mapping $\mu \mapsto (\mu(A_1), \mu(A_2), \dots)$, we can realize the space as a subset of the compact separable space $[0,1] \times [0,1] \times \cdots$, determined by various constraints. One of these constraints is that the measure of the whole space must be 1. Another constraint is that every set must get measure ≥ 0. Still other constraints are consistency conditions that a set-function must satisfy if it is to qualify as a measure; if we want μ to be countably additive on $\mathcal{A}$, it is necessary and sufficient that its restriction to $\{A_k\}$ be finitely additive. Lastly, there are the stationarity constraints $\mu(A_k) = \mu(A_{k'})$ (where $A_{k'} = TA_k$). Since each of these constraints involves only a finite number of the coordinates $\mu(A_k)$, they jointly determine a closed subset of $[0,1] \times [0,1] \times \cdots$. □

A special case of the above arises from looking at the set of joinings of two dynamical systems P, m_P and Q, m_Q. In this case, we assume that the countable generating sub-algebra is the algebra of $P \vee Q$ cylinder sets; a measure μ on the product space $X \times Y$ will be a joining if (in addition to the constraints mentioned in the proof of Claim 5) it satisfies $\mu(A \times Y) = m_P(A)$ and $\mu(X \times B) = m_Q(B)$ for all P-cylinders A and all Q-cylinders B. As in the proof of Claim 5, each such constraint involves only a finite number of cylinders (one, in fact). So, for the same reason as before, we have

CLAIM 6: *In the distribution topology, the set of joinings of two stochastic processes is a compact separable metric space.* □

Remark: It turns out that this topology is actually independent of which generators P, Q one uses for the two processes; see [11] or [13].

CLAIM 7: *If R and R' are cylinder partitions, then $H_\mu(R \mid R')$ is a continuous function of μ.*

Proof: The quantity is a continuous function of the distribution vector $\mathrm{dist}_\mu(R \vee R')$, which is itself a continuous function of μ. □

CLAIM 8: *If R is a cylinder partition and $\mathcal{B}$ is an increasing limit of cylinder sub-algebras of $\mathcal{A}$, then $H_\mu(R \mid \mathcal{B})$ is an upper semi-continuous function of μ.*

Proof: Fix μ and fix $\epsilon > 0$. There exists a cylinder partition $R' \subset \mathcal{B}$ such that $H_\mu(R \mid R') < H_\mu(R \mid \mathcal{B}) + \epsilon/2$. Hence (by Claim 7) for all ν sufficiently close to μ in distribution, $H_\nu(R \mid \mathcal{B}) \leq H_\nu(R \mid R') < H_\mu(R \mid R') + \epsilon/2 < H_\mu(R \mid \mathcal{B}) + \epsilon$. That is, $H_\nu(R \mid \mathcal{B}) < H_\mu(R \mid \mathcal{B}) + \epsilon$ for all ν sufficiently near μ. Hence $H(R \mid \mathcal{B})$ varies upper semi-continuously with the measure. Putting it another way: for all $\alpha > 0$, the set $\{\mu : H_\mu(R \mid \mathcal{B}) < \alpha\}$ is open while the set $\{\mu : H_\mu(R \mid \mathcal{B}) \geq \alpha\}$ is closed. □

CLAIM 9: *The set of pre-unilateral joinings of two processes is a closed subset of the space of joinings.*

Proof: Recall that pre-unilaterality is equivalent to

$$H(P^1 \mid P_{-\infty}^0 \vee Q_{-\infty}^0) = H(P^1 \mid P_{-\infty}^0) = h(P)$$

(see the remark following the proof of Claim 3). Since $H(P^1 \mid P_{-\infty}^0 \vee Q_{-\infty}^0)$ is automatically less than or equal to $H(P^1 \mid P_{-\infty}^0)$, pre-unilaterality is equivalent to

$$H(P^1 \mid P_{-\infty}^0 \vee Q_{-\infty}^0) \geq h(P) .$$

But $H_\mu(P^1 \mid P_{-\infty}^0 \vee Q_{-\infty}^0)$ is an upper semi-continuous function of μ, so the set of μ's for which the preceding inequality holds must be closed. □

Heuristically, one may think of the set of pre-unilateral joinings as the distribution-closure of the set of unilateral graph-joinings. More specifically, we will show that every ergodic pre-unilateral joining can be approximated arbitrarily well by unilateral graph-joinings. Anticipating our special interest in ergodic joinings, let us prove

CLAIM 10: *A joining μ is pre-unilateral if and only if almost all of its ergodic components are.*

Proof: Let Z denote the measure space μ lives on. Without loss of generality, we may put $Z = \bigcup_\alpha Z_\alpha$ and $\mu = \int \lambda_\alpha \, d\alpha$, where the λ_α's are ergodic measures supported on the respective Z_α's ($\lambda_\alpha(Z_\alpha) = 1$). Think of λ as the measurable partition of Z into its ergodic components; since P, m_P is ergodic, almost every λ_α projects to m_P on $P_{-\infty}^{\infty}$. The ergodic theorem implies that for almost all $z \in Z$, the infinite name $(P \vee Q)_{-\infty}^{0}(z)$ manifests (via frequency statistics) the distribution $\text{dist}_{\lambda(z)} (P \vee Q)_0^{r-1}$ for all r, where $\lambda(z)$ is the λ_α with $z \in Z_\alpha$. That is, $(P \vee Q)_{-\infty}^{0}(z)$ almost surely determines the statistics of $\lambda(z)$ on cylinder sets. Also, no distinct λ_α's can have the same statistics on all cylinder sets, since these statistics uniquely characterize each λ_α (by the uniqueness part of the extension theorem). Hence (up to measure zero) λ is $P_{-\infty}^{0} \vee Q_{-\infty}^{0}$-measurable.

Now we may write

$$\begin{aligned} H_\mu(P^1 \mid P_{-\infty}^{0} \vee Q_{-\infty}^{0}) &= H_\mu(P^1 \mid P_{-\infty}^{0} \vee Q_{-\infty}^{0} \vee \lambda) \\ &= \int H_{\lambda_\alpha}(P^1 \mid P_{-\infty}^{0} \vee Q_{-\infty}^{0}) \, d\alpha \\ &\leq \int H_{\lambda_\alpha}(P^1 \mid P_{-\infty}^{0}) \, d\alpha \\ &= \int h(P) \, d\alpha \\ &= h(P) \, . \end{aligned}$$

Hence the equality $H_\mu(P^1 \mid P_{-\infty}^{0} \vee Q_{-\infty}^{0}) = h(P)$ holds if and only if

$$H_{\lambda_\alpha}(P^1 \mid P_{-\infty}^{0} \vee Q_{-\infty}^{0}) = h(P)$$

for almost all α. □

If μ is a joining of Q, m_Q and $\tilde{Q}, \tilde{m}_Q$ with $\#(Q) = \#(\tilde{Q})$ such that $\mu(Q \triangle \tilde{Q}) \leq \epsilon$, we say that the joining μ is ϵ-tight. We say that the ergodic process Q, m_Q is **unilaterally finitely determined** (or UFD) if for all $\epsilon > 0$ there exists a $\delta > 0$ and a distribution neighborhood of Q, m_Q in the space of processes on $\#(Q)$ symbols, such that for all processes $\tilde{Q}, \tilde{m}_Q$ in that neighborhood of Q, m_Q that satisfy $|h(\tilde{Q}) - h(Q)| < \delta$, there exists an ϵ-tight pre-unilateral joining of Q, m_Q and $\tilde{Q}, \tilde{m}_Q$. (Here pre-unilaterality means $\tilde{Q}_1^{\infty} \perp Q_{-\infty}^{0} | \tilde{Q}_{-\infty}^{0}$.) Note that if the pre-unilaterality requirement is dropped, we obtain the standard notion of a finitely determined process.

At a certain point in this article we shall also have to consider non-stationary joinings of two processes Q, m_Q and $\tilde{Q}, \tilde{m}_Q$. If μ is such a non-stationary joining (that is, μ projects to m_Q and $\tilde{m}_Q$ but is not necessarily invariant under the shift), we say that it is ϵ-tight if and only if $\pi(P^n \triangle Q^n) \leq \epsilon$ for all n. In the case that μ is invariant, this coincides with the previous definition.

We can now give an overview of the proof. Our aim is to show that a mixing Markov chain is a unilateral factor of every ergodic process of equal or greater entropy. Equivalently:

THEOREM: *If P, m_P is an ergodic process and Q, m_Q is a mixing Markov process such that $h(Q) \leq h(P)$, then there exists a unilateral graph joining μ of m_P and m_Q.*

To prove this, we first prove a Copying Lemma (section 3). The proof uses the traditional framework of Rokhlin towers and block codes, but in lieu of the customary marriage lemma it resorts to a random coding argument. Note that the pre-unilaterality constraint is indispensable in the hypothesis of the Copying Lemma; indeed, Claims 2 and 9 imply that if μ is not pre-unilateral, then there exists a neighborhood of μ containing *no* unilateral graph joinings. Section 4 contains a proof of a Joining Lemma; it is an adaptation of Ornstein's proof [8] that independent processes are finitely determined. In Section 5, this result is combined with the Copying Lemma to yield an Improvement Lemma, which is shown to imply the Theorem.

3. The Copying Lemma

Fix ergodic processes P, m_P and Q, m_Q with $h(P) \geq h(Q)$. Given an ergodic pre-unilateral joining μ of P, m_P and Q, m_Q and a fidelity criterion, we wish to construct a $P_{-\infty}^{0}$-measurable partition $\tilde{Q}$ such that the joint distribution of P and $\tilde{Q}$ approximates that of P and Q, and the $\tilde{Q}$-process has approximately the same entropy as the Q-process. More precisely, we require:

(a) $$\frac{1}{2}\,|\mathrm{dist}_\mu((P \vee \tilde{Q})_1^r) - \mathrm{dist}_\mu((P \vee Q)_1^r)| < \delta \quad \text{and}$$

(b) $$|h(\tilde{Q}) - h(Q)| < \delta.$$

(We may suppose $r \geq 1$, $\delta < 1/10$, $\delta < 1/\log \#(P)$.)

Our construction, in outline, is as follows. We take a block-length n and a set E such that $TE, T^2E, \ldots, T^nE$ are disjoint and such that their union T^*E has

measure nearly 1. (We do not insist that the level $T^n E$ of the Rokhlin tower be disjoint from the "basement" E.) We devise a measurable map $f : P_{-\infty}^n \to Q_1^n$ taking semi-infinite P-names to finite Q-names and use it to give $\tilde{Q}$-names to all the points in the tower $T^* E$. Specifically, if a point x lies in the kth level of the tower ($1 \le k \le n$) and if its $P_{-\infty}^n$-name is A, then we assign x to the set $\tilde{Q}_{(j)}$, where j is the kth symbol of $f(A) = f((P_{-\infty}^n)(x))$. (Points outside the tower may as well get assigned to $\tilde{Q}_{(1)}$.)

To guarantee that a $\tilde{Q}$ constructed in this fashion is $P_{-\infty}^0$-measurable, it suffices that two conditions be satisfied: first, E is $P_{-\infty}^0$-measurable, and second, f is "unilateral" in the sense that if α, α' are two semi-infinite P-names in $P_{-\infty}^n$ that agree up to time $k \ge 1$, then $f(\alpha), f(\alpha')$ must also agree up to time k. The first condition allows us to determine, given $(P_{-\infty}^0)(x)$, whether or not x lies in $T^* E$, and if so, what level $T^{k(x)} E$ it lies in; the second condition says that for $x \in T^k E$ (say $x = T^k x_0$), knowledge of $P_{-\infty}^0(x) = P_{-\infty}^k(x_0)$ gives knowledge of $\tilde{Q}^k(x_0) = \tilde{Q}(x)$; so that, if both conditions hold, we see that knowing $(P_{-\infty}^0)(x)$ tells us $\tilde{Q}(x)$.

Most of this section is devoted to showing that a suitable f exists. Before doing this, let us first make sure that there is no difficulty in building Rokhlin towers whose bases are measurable with respect to the past.

ONE-SIDED ROKHLIN LEMMA: *Let $(X, T, \mathcal{A}, \mu)$ be an ergodic process with generator P. Given any $n \ge 1$ and $\epsilon > 0$, there exists a set E in $P_{-\infty}^0$ such that the sets $TE, T^2 E, \ldots, T^n E$ are disjoint with total measure $> 1 - \epsilon$. Moreover, if G is some pre-specified set of points with measure $> 1 - \epsilon'$, we can choose E so as to satisfy $\mu(G \mid E) > 1 - \epsilon'$ as well.*

(We omit the proof.)

Given $\epsilon > 0$ and $n \ge 1$, call an atom α of $P_{-\infty}^0$ ϵ-good if for all but ϵ^4 in conditional measure of the points $x \in \alpha$ one has

$$\mu_\alpha\big((P_1^k \vee Q_1^n)(x)\big) = e^{-(k\,h(P \vee Q) + (n-k)\,h(Q) \pm n\epsilon)}$$

and

$$\mu_\alpha\big((P_1^n \vee Q_1^k)(x)\big) = e^{-(k\,h(P \vee Q) + (n-k)\,h(P) \pm n\epsilon)}$$

for all $k \le n$, where μ_α is μ conditioned on the past-name α. We wish to know that most of the past-atoms α are ϵ-good. Theorem 3 in [12] says:

EQUIDISTRIBUTION THEOREM: *Let P, Q, and R be finite partitions of an ergodic dynamical system, with $0 < \epsilon < 1$. Then for all n sufficiently large,*

there exists a set of fibers α of $R_{-\infty}^{0}$ of total measure $> 1-\epsilon$, on each of which it is the case that for all but ϵ in conditional measure of the points $x \in \alpha$,

$$\mu_\alpha\big((P_1^k \vee Q_{k+1}^n)(x)\big) = e^{-n(\theta_k \pm \epsilon)}$$

for all $k \leq n$, where

$$\theta_k = \frac{1}{n}\left(k\,h(P) + (n-k)\,h(Q)\right) .$$

(Note that for $P = Q$, this is the Shannon-McMillan Theorem [7].) It follows from this theorem (by setting P, Q, R, ϵ equal to $P \vee Q, Q, P, \epsilon^4/2$ and then $P \vee Q, P, P, \epsilon^4/2$) that for all large n, the ϵ-good atoms α have total measure $> 1-\epsilon^4$. Here ϵ is some specific quantity whose dependence on r and δ will be specified later; for now, we may stipulate that $\epsilon < \delta^5$. Similarly, we will take a surrogate sub-block length $s \geq r$, whose precise relationship to r and δ will be discussed later (see the paragraph following the proof of Fact 2, below).

Note that if P, m_P is an i.i.d. process, it is not necessary to condition on individual atoms α of $P_{-\infty}^{0}$, either here or in subsequent stages of the copying argument.

Let G be the union of the ϵ-good atoms of $P_{-\infty}^{0}$. (Hereafter, we shall merely call them "good".) $\mu(G) > 1-\epsilon^4 > 1-\epsilon$, so by our One-Sided Rokhlin Lemma, we may suppose our base E has the property that G has conditional measure $> 1-\epsilon$ on E; that is,

$$\mu_E(G) > 1-\epsilon. \tag{3}$$

We will also suppose that our tower T^*E satisfies

$$\mu(T^*E) > 1 - 1/n .$$

Now we give a procedure for constructing the map $f : P_{-\infty}^{n} \to Q_1^n$. We will do this by randomly selecting a map $f_\alpha : P_1^n \to Q_1^n$ for each atom α of $P_{-\infty}^{0}$, and letting $f(\alpha \cap A) = f_\alpha(A)$ for all $A \in P_1^n$. (Actually, only the α's satisfying $\alpha \subset E$ are involved; the other α's are disjoint from E and play no part in the construction.) Since there are uncountably many α's, we cannot choose all the maps f_α independently of one another without sacrificing measurability of the unified map f; fortunately, we do not need to assume that f_α and $f_{\alpha'}$ are uncorrelated for $\alpha \neq \alpha'$. Indeed, here is a concrete way to imagine the simultaneous selection of all the f_α's. Let

$$N = (\#(Q)^n)^{(\#(P)^n)} ,$$

the number of possible block maps from P_1^n to Q_1^n, and index these block maps as $f_{(1)}, \ldots, f_{(N)}$. For each α, the construction we are about to describe gives a probability distribution on the block-maps $f_{(i)}$; thus, for each past-name α in $P_{-\infty}^{0}$ we can partition the interval $[0,1]$ into sub-intervals $I_{\alpha,i}$ so that the length of $I_{\alpha,i}$ is equal to the probability associated with the block map $f_{(i)}$ for the past-name α. To randomly select an $f : P_{-\infty}^{n} \to Q_{-\infty}^{0}$, it suffices to choose a single random number $t \in [0,1]$, as we can then take $f_\alpha = f_{(i)}$ for the unique i such that $t \in I_{\alpha,i}$, and then put $f(\alpha \cap A) = f_\alpha(A)$ as above.

We will use F_α to denote the random variable taking its value in the set of maps $f_\alpha : P_1^n \to Q_1^n$; similarly, we will use F to denote the random variable whose values are maps $f : P_{-\infty}^{n} \to Q_1^n$.

Fix α. f_α must have the property that the first k symbols of a P-name $A \in P_1^n$ determine the first k symbols of the Q-name $f_\alpha(A) \in Q_1^n$; accordingly, we may think of f_α as a function that takes P-names of length k to Q-names of length k, for all k between 1 and n. Our plan for constructing this extended version of the map f_α is to proceed iteratively, defining the behavior of the map on length-k names for $k = 1, 2, \ldots, n$ in succession.

First ($k = 1$), for each atom A of P^1 we choose an atom $F_\alpha(A)$ of Q^1 at random, according to

$$\text{Prob } [F_\alpha(A) = B] = \frac{\mu_\alpha(A \cap B)}{\mu_\alpha(A)} = \mu_\alpha(B \mid A) .$$

For distinct atoms $A \in P^1$, we choose the $F_\alpha(A)$'s independently of one another.

Now suppose we have defined F_α as a mapping from P_1^{k-1} to Q_1^{k-1}. We extend F_α to P_1^k as follows. Given an atom A of P_1^k, let $\overline{A}$ be $P_1^{k-1}(A)$ (the atom of P_1^{k-1} containing A), and let $\overline{B} = F_\alpha(\overline{A}) \in Q^{k-1}_1$. We choose $F_\alpha(A)$ from among the atoms B of Q_1^k contained in $\overline{B}$, according to the conditional probability

$$\text{Prob } [F_\alpha(A) = B \mid F_\alpha(\overline{A}) = \overline{B}] = \frac{\mu_\alpha(A \cap B)}{\mu_\alpha(A \cap \overline{B})} = \mu_\alpha(B \mid A \cap \overline{B}) .$$

For distinct atoms $A \in P_1^k$, we choose the $F_\alpha(A)$'s independently of one another. (More precisely, if A and A' are distinct atoms of P_1^k, then $F_\alpha(A)$ and $F_\alpha(A')$ are conditionally independent given $F_\alpha(\overline{A})$ and $F_\alpha(\overline{A'})$, where $\overline{A} = P_1^{k-1}(A)$ and $\overline{A'} = P_1^{k-1}(A')$.)

If we iterate this construction n times, we get a random map $F_\alpha : P_1^n \to Q_1^n$ with the property that for all $k \leq n$, the first k symbols of $A \in P_1^n$ determine the first k symbols of $F(A)$.

We now verify a crucial fact, namely,

$$\text{Prob } [F_\alpha(A) = B] = \mu_\alpha(B \mid A) \tag{4}$$

for all names $A \in P_1^n$, $B \in Q_1^n$. For all $0 \leq k \leq n$, let $A_k = P_1^k(A)$ and $B_k = Q_1^k(B)$. Thus in particular A_0 and B_0 are trivial names, while $A_n = A$ and $B_n = B$. Since μ is a pre-unilateral joining,

$$P_1^\infty \perp Q_{-\infty}^{\ 0} \mid P_{-\infty}^{\ 0} ,$$

so that stationarity yields

$$P_{k+1}^\infty \perp Q_{-\infty}^k \mid P_{-\infty}^k .$$

A fortiori we have

$$P_{k+1}^n \perp Q_1^k \mid P_{-\infty}^k ,$$

which implies that

$$P_{k+1}^n \perp Q_1^k \mid P_{-\infty}^k \vee Q^{k-1}{}_1$$

(since $Q^{k-1}{}_1 \subset Q_1^k$). Hence

$$\text{dist}_\mu(Q_1^k \mid P_{-\infty}^k \vee Q^{k-1}{}_1) = \text{dist}_\mu(Q_1^k \mid P_{-\infty}^n \vee Q^{k-1}{}_1) ;$$

separating out the conditioning on $P_{-\infty}^{\ 0}$, we get

$$\text{dist}_{\mu_\alpha}(Q_1^k \mid P_1^k \vee Q^{k-1}{}_1) = \text{dist}_{\mu_\alpha}(Q_1^k \mid P_1^n \vee Q^{k-1}{}_1) .$$

In particular, the nested atoms $A_0, \ldots, A_n$ and $B_0, \ldots, B_n$ satisfy

$$\mu_\alpha(B_k \mid A_k \cap B_{k-1}) = \mu_\alpha(B_k \mid A_n \cap B_{k-1}) \tag{5}$$

for all $1 \leq k \leq n$. Therefore

$$\begin{aligned}
\text{Prob } [F_\alpha(A) = B] &= \prod_{k=1}^n \text{Prob } [F_\alpha(A_k) = B_k \mid F_\alpha(A_{k-1}) = B_{k-1}] \\
&= \prod_{k=1}^n \mu_\alpha(B_k \mid A_k \cap B_{k-1}) \text{ (by construction)} \\
&= \prod_{k=1}^n \mu_\alpha(B_k \mid A_n \cap B_{k-1}) \text{ (by (5))} \\
&= \mu_\alpha(B_n \mid A_n) \text{ (by successive conditioning)} \\
&= \mu_\alpha(B \mid A)
\end{aligned}$$

as claimed.

With the aid of (4), we are nearly in a position to verify that with high probability our randomly selected map $f : P_{-\infty}^n \to Q_1^n$ will determine a $\tilde{Q}$ satisfying $\text{dist}_\mu(P \vee \tilde{Q})_1^r \approx \text{dist}_\mu(P \vee Q)_1^r$. Only one ingredient is missing, namely the following fact:

NORMALITY LEMMA: *Let χ be a bounded measurable function on an ergodic dynamical system (X, T, μ); put*

$$\chi_n^*(x) = \frac{1}{n} \sum_{k=0}^{n-1} \chi(T^k x)$$

and

$$\bar{\chi} = \int \chi(x)\, d\mu(x) \,.$$

Call the point x n, ϵ-normal if $|\chi_n^*(x) - \bar{\chi}| < \epsilon$. Then, given any $\epsilon > 0$, it will hold that for all Rokhlin towers of sufficiently large height n and total mass $> \epsilon$ (sic), all but ϵ in conditional measure of the points in the base are n, ϵ-normal.

Proof: Without loss of generality, suppose the bounded function χ satisfies

$$\sup \chi - \inf \chi \leq 1.$$

First note that if a point x is $n, \epsilon/2$-normal, then $T^k x$ is n, ϵ-normal for $|k| \leq (\epsilon/2)n$. For, we have

$$|\chi_n^*(T^{i+1}x) - \chi_n^*(T^i x)| = \left| \frac{1}{n} \left(\chi(T^{i+n}x) - \chi(T^i x)\right) \right| \leq \frac{1}{n} \quad \text{for all } i;$$

adding together $|k| \leq (\epsilon/2)n$ such terms, we find that

$$|\chi_n^*(T^k x) - \chi_n^*(x)| \leq |k| \cdot \frac{1}{n} \leq \frac{\epsilon}{2}$$

and

$$|\chi_n^*(T^k x) - \bar{\chi}| \leq |\chi_n^*(T^k x) - \chi_n^*(x)| + |\chi_n^*(x) - \bar{\chi}| < \epsilon$$

as claimed. Now, the ergodic theorem guarantees that if n is large enough, the set of points x that are not $n, \epsilon/2$-normal has measure $< \epsilon^3/2$. Fix such an n. If $T^* E$ is a Rokhlin tower of height n and total mass exceeding ϵ, the base E must have mass at least ϵ/n. Suppose that a proportion of more than ϵ of the points in the base failed to be n, ϵ-normal. Then the absolute measure of these abnormal points $x \in E$ would be at least ϵ^2/n. For all such x, $T^k x$ is $n, \epsilon/2$-abnormal for all k with $0 \leq k \leq (\epsilon/2)n$. The set of such points $T^k x$ has measure at least $((\epsilon/2)n)(\epsilon^2/n) = \epsilon^3/2$. This contradicts our assumption about n, and proves that in fact no more than an ϵ fraction of the base can be n, ϵ-abnormal. □

Clearly the same result holds if we desire that points in the base be n, ϵ-normal not with respect to a single bounded function χ but simultaneously with respect to any finite set of bounded functions. In particular, we may take the indicator functions $\chi_C(x)$, where C varies over the components of $(P \vee Q)_1^s$. Then we see that for nearly all points x in the base E of a Rokhlin tower of sufficiently large height n, the frequency statistics of length-s subnames in the length-n name $(P \vee Q)_1^n$ will be as close to $\mathrm{dist}_\mu((P \vee Q)_1^s)$ as we like.

We now return to the context of our random selection of $f : P_{-\infty}^n \to Q_1^n$. Recall that ϵ and s are yet-to-be-specified functions of δ and r. Put $\epsilon' = \epsilon^2/\#((P \vee Q)_1^s)$, and call a $(P \vee Q)_1^n$-name $A \times B$ ($A \in P_1^n$, $B \in Q_1^n$) **normal** if each length-s subname $C \in (P \vee Q)_1^s$ occurs in the name $A \times B$ with frequency $\mu(C) \pm \epsilon'$. Define the **abnormality set** of a map $F : P_{-\infty}^n \to Q_1^n$ and a Rokhlin base E as the union of the atoms $\alpha \cap A \subset E$ (α in $P_{-\infty}^0$, A in P_1^n) for which $A \times F_\alpha(A)$ is abnormal. Lastly, say that the map F itself is normal (with respect to the base E) if the abnormality set has conditional measure $< \epsilon^2$ on E.

FACT 1: *If n is sufficiently large, the likelihood that the random map F is normal is at least $1 - \epsilon$.*

Proof: By the Normality Lemma and the remark that follows its proof, if n is large enough all but an ϵ^8 fraction of the points in E will have normal $(P \vee Q)_1^n$-names. This implies that off of a set of $P_{-\infty}^0 \vee P_1^n$-names $\alpha \cap A$ of total conditional measure $< \epsilon^4$ on E, the $\alpha \cap A$-conditional measure of the normal $(P \vee Q)_1^n$-names exceeds $1 - \epsilon^4$. Focus on a non-exceptional $\alpha \cap A$; the set of names $B \in Q_1^n$ for which $A \times B$ is normal has conditional measure $> 1 - \epsilon^4$, so if one chooses such a name $F_\alpha(A)$ at random with Prob $[F_\alpha(A) = B] = \mu(B \mid \alpha \cap A)$ (which is precisely what our construction of F has us do), the probability that $A \times F_\alpha(A)$ will be normal is at least $1 - \epsilon^4$. If we now let $\alpha \cap A$ vary, we see that the expected conditional measure (relative to E) of the abnormality set is $< \epsilon^4 + \epsilon^4 = 2\epsilon^4$, so that with likelihood $> 1 - 2\epsilon^2 > 1 - \epsilon$, the abnormality set has measure $< \epsilon^2$ on E, in which case F is normal. □

FACT 2: *If F is normal, the resulting $\tilde{Q}$ has the property that*

$$\left|\mathrm{dist}_\mu((P \vee \tilde{Q})_1^s) - \mathrm{dist}_\mu((P \vee Q)_1^s)\right| < \epsilon\,.$$

Proof: Let E' be the complement of the abnormality set in E, and let $X' =$

$\{T^k : x \in E', 0 \le k \le n-s\}$. Then the normality of the points in E' implies

$$\begin{aligned}|\mathrm{dist}_\mu((P \vee \tilde{Q})_1^s | X') - \mathrm{dist}_\mu((P \vee Q)_1^s)| &= \sum_{C \in (P \vee Q)_1^s} |\mu(\tilde{C} \mid X') - \mu(C)| \\ &\le \#((P \vee Q)_1^s) \cdot \epsilon' \\ &= \epsilon^2\end{aligned}$$

(where $\tilde{C}$ denotes the atom of $(P \vee \tilde{Q})_1^s$ corresponding to $C \in (P \vee Q)_1^s$). On the other hand, the points $T^k x$ with $x \in E \setminus E'$ and $0 \le k \le n-s$ have total mass $< \epsilon^2$, the points $T^k x$ with $x \in E$ and $n-s < k < n$ have total mass $< s/n < \epsilon^2$ (for large n), and the points x that belong to none of the $T^k E$ $(0 \le k < n)$ have total mass $< 1/n < \epsilon^2$ (for large n), so $\mu(X \setminus X') < 3\epsilon^2$ and $|\mathrm{dist}_\mu((P \vee \tilde{Q})_1^s) - \mathrm{dist}_\mu((P \vee \tilde{Q})_1^s \mid X')| \le 2\mu(X \setminus X') < 6\epsilon^2$. Hence $|\mathrm{dist}_\mu((P \vee \tilde{Q})_1^s) - \mathrm{dist}_\mu((P \vee Q)_1^s)| < 6\epsilon^2 + \epsilon^2 < \epsilon$. □

We now complete the specification of the quantities s and ϵ in terms of r and δ. We have already stipulated that $s \ge r$ and $\epsilon \le \delta^5 < \delta$, so the inequality in Fact 2 implies

$$\frac{1}{2} |\mathrm{dist}_\mu((P \vee \tilde{Q})_1^r) - \mathrm{dist}_\mu((P \vee Q)_1^r)| < \delta ,$$

which is condition (a) from the beginning of this section. However, by requiring s to be even larger and ϵ to be even smaller, we can go part of the way toward ensuring that (b) is satisfied as well. For, suppose s is so large that $(1/s)H(Q_1^s) = h(Q) \pm \delta/2$. Then by requiring ϵ to be suitably small, we can ensure that the condition

$$|\mathrm{dist}_\mu(\tilde{Q}_1^s) - \mathrm{dist}_\mu(Q_1^s)| < \epsilon \tag{6}$$

implies $(1/s)H(\tilde{Q}_1^s) < (1/s)H(Q_1^s) + \delta/2$. It follows that if the partition $\tilde{Q}$ satisfies $|\mathrm{dist}_\mu((P \vee \tilde{Q})_1^s) - \mathrm{dist}_\mu((P \vee Q)_1^s)| < \epsilon$ (so that a fortiori (6) is satisfied), then $h(\tilde{Q}) \le \frac{1}{s} H(\tilde{Q}_1^s) < h(Q) + \delta$. This gives us half of condition (b); to prove the other, we need to show that $h(\tilde{Q}) > h(Q) - \delta$.

Recall (see (3)) that the union G of the good atoms of $P_{-\infty}^{\ 0}$ has conditional measure $> 1 - \epsilon$ on E. Suppose α is a good atom of $P_{-\infty}^{\ 0}$, and let G_α be the union of the $(P \vee Q)_1^n$-names C with the property that

$$\mu_\alpha((P_1^k \vee Q_1^n)(C)) = e^{-(kh(P \vee Q) + (n-k)h(Q) \pm n\epsilon)}$$

and

$$\mu_\alpha((P_1^n \vee Q_1^k)(C)) = e^{-(kh(P\vee Q)+(n-k)h(P)\pm n\epsilon)}$$

for all $k \leq n$; since α is good, $\mu_\alpha(G_\alpha) > 1 - \epsilon^4 > 1 - (\epsilon/2)^3$.

Let $\mathcal{A}_\alpha$ be the collection of atoms of $P_1^n \mid \alpha$ on which G_α has measure $> 1 - (\epsilon/2)^2$, so that $\mu(\bigcup \mathcal{A}) > 1 - \epsilon/2$, and let $\mathcal{B}_\alpha$ be the collection of atoms of $Q_1^n \mid \alpha$ on which G_α has measure $> 1 - (\epsilon/2)^2$, so that $\mu(\bigcup \mathcal{B}) > 1 - \epsilon/2$. Given $A \in \mathcal{A}_\alpha$ and $B \in \mathcal{B}_\alpha$, call A and B **compatible** if $A \cap B \subset G_\alpha$; we write $B \in \mathrm{Comp}_\alpha(A)$, $A \in \mathrm{Comp}_\alpha(B)$. Note that for each $A \in \mathcal{A}_\alpha$, the probability of choosing F_α such that $F_\alpha(A) \in \mathrm{Comp}_\alpha(A)$ is at least $1 - (\epsilon/2)^2$. Hence, off of a set of exceptional f_α's of total probability $\epsilon/2$, the atoms $A \in \mathcal{A}_\alpha$ satisfying $F_\alpha(A) \in \mathrm{Comp}_\alpha(A)$ occupy all but $\epsilon/2$ of $\bigcup \mathcal{A}_\alpha$, and thus all but ϵ of X. For each f_α, let

$$\mathcal{A}_\alpha^* = \{A \in \mathcal{A}_\alpha : f_\alpha(A) \in \mathrm{Comp}_\alpha(A)\}$$

(this depends not only on α but on f_α as well) and let f_α^* be the restriction of f_α to $\mathcal{A}_\alpha^*$. In this way each map f_α determines a restricted map f_α^*, and the map-valued random variable F_α determines another map-valued random variable F_α^*. With probability $> 1 - \epsilon/2 > 1 - \epsilon$, $\mu_\alpha(\bigcup \mathcal{A}_\alpha^*) > 1 - \epsilon$. Call the map f_α **good** if $\mu_\alpha(\bigcup \mathcal{A}_\alpha^*) > 1 - \epsilon$. Then we have shown:

FACT 3: *If α is good, then the likelihood that F_α is good exceeds $1 - \epsilon$.*

(Here, as in Facts 4 through 8, the words "For all sufficiently large n..." are implicit.) □

Say that the P_1^n-name A is **balanced** with respect to a particular map f_α if $\#(f_\alpha^{*-1}(f_\alpha^*(A))) \leq e^{n(h(P)-h(Q)+6\epsilon)}$, where we adopt the convention that $f_\alpha^{*-1}(f_\alpha^*(A))$ is empty when $A \notin \mathcal{A}_\alpha^* = \mathrm{dom}\, f_\alpha^*$. Say that the map $f_\alpha : P_1^n \to Q_1^n$ is balanced if the set of balanced names $A \in P_1^n$ has μ_α-measure $> 1 - \epsilon$. Lastly, say that the map $f : P_{-\infty}^n \to Q_1^n$ is **excellent** if the set of α's for which α is good and f_α is good and balanced has μ_E-measure $> 1 - \delta^3$. We will show that (for large n) the likelihood that the random map F is excellent exceeds $1 - \delta$, and that if F is excellent, the corresponding partition $\tilde{Q}$ satisfies $h(\tilde{Q}) > h(Q) - \delta$.

FACT 4: *For all $A \in \mathcal{A}_\alpha$, the expected cardinality of $F_\alpha^{*-1}(F_\alpha^*(A))$ is less than $e^{n(h(P)-h(Q)+5\epsilon)}$.*

Proof: Fix B, and condition our random selection of F_α on the event $\{f_\alpha : f_\alpha(A) = B\}$. If $B \notin \mathrm{Comp}_\alpha(A)$, we get $F_\alpha^{*-1}(F_\alpha^*(A)) = \phi$. On the other hand,

suppose $B \in \mathrm{Comp}_\alpha(A)$, so that $A \in \mathcal{A}_\alpha^*$; if we can show that in this case the conditional expectation of $\#(F_\alpha^{*-1}(F_\alpha^*(A)))$ must be less than $e^{n(h(P)-h(Q)+5\epsilon)}$, then we will have proved Fact 4.

We have $A \in P_1^n$, $B \in Q_1^n$ with $B \in \mathrm{Comp}_\alpha(A)$; the conditional expectation of $\#(F_\alpha^{*-1}(F_\alpha^*(A)))$ is equal to

$$\sum_{A' \in \mathrm{Comp}_\alpha(B)} \mathrm{Prob}\ [F_\alpha(A') = B \mid F_\alpha(A) = B]\ . \tag{7}$$

Fix $A' \in \mathrm{Comp}_\alpha(B)$; suppose the length-$n$ name A' agrees with the length-n name A up to time k but no further, with $0 \le k \le n$. If we let $\overline{A} = P_1^k(A) = P_1^k(A')$ and $\overline{B} = Q_1^k(B)$, then the conditional independence of $F_\alpha(A)$ and $F_\alpha(A')$ given $F_\alpha(\overline{A})$ implies

$$\mathrm{Prob}\ [F_\alpha(A') = B \mid F_\alpha(A) = B] = \frac{\mu_\alpha(A' \cap B)}{\mu_\alpha(A' \cap \overline{B})}\ .$$

Since $A' \in \mathrm{Comp}_\alpha(B)$, we have $A' \cap B \subset G_\alpha$, so that the contribution of A' to (7) is

$$\begin{aligned} \frac{\mu_\alpha(A' \cap B)}{\mu_\alpha(A' \cap \overline{B})} &= \frac{e^{-(n\,h(P \vee Q) \pm n\epsilon)}}{e^{-(k\,h(P \vee Q) + (n-k)\,h(P) \pm n\epsilon)}} \\ &< e^{-(n-k)(h(P \vee Q) - h(P)) + 2n\epsilon}\ . \end{aligned}$$

Now hold $k < n$ fixed and let A' vary over all B-compatible names that match A for k steps but no further. Each such A' must satisfy

$$\begin{aligned} \frac{\mu_\alpha(\overline{A} \cap B)}{\mu_\alpha(A' \cap B)} &= \frac{e^{-(k\,h(P \vee Q) + (n-k)\,h(Q) \pm n\epsilon)}}{e^{-(n\,h(P \vee Q) \pm n\epsilon)}} \\ &< e^{(n-k)(h(P \vee Q) - h(Q)) + 2n\epsilon}\ , \end{aligned}$$

so the number of such atoms A' is at most $e^{(n-k)(h(P \vee Q) - h(Q)) + 2n\epsilon}$; and since each contributes at most $e^{-(n-k)(h(P \vee Q) - h(P)) + 2n\epsilon}$ to (7), their joint contribution is at most $e^{(n-k)(h(P) - h(Q)) + 4n\epsilon} \le e^{n(h(P) - h(Q) + 4\epsilon)}$. Hence, summing over all $0 \le k \le n$, we can bound (7) by $(n+1)e^{n(h(P) - h(Q) + 4\epsilon)}$, which is less than $e^{n(h(P) - h(Q) + 5\epsilon)}$ for large n. □

FACT 5: *If α is good, then the likelihood that F_α is balanced exceeds $1 - \epsilon$.*

Proof: Suppose not. That is, suppose that the likelihood is ϵ or greater of choosing F_α such that $\#(F_\alpha^{*-1}(F_\alpha^*(A)))$ exceeds $e^{n(h(P) - h(Q) + 6\epsilon)}$ for a set of A's

of total measure $\geq \epsilon$. Then the expected value (as $F_\alpha : P_1^n \to Q_1^n$ and $A \in \mathcal{A}_\alpha$ vary) of $\#(F_\alpha^{*-1}(F_\alpha^*(A)))$ must be at least $\epsilon \cdot \epsilon \cdot e^{n(h(P)-h(Q)+6\epsilon)}$. But for large n this exceeds $e^{n(h(P)-h(Q)+5\epsilon)}$, whereas the estimate from Fact 4, averaged over all A, implies that this expected value must be less than $e^{n(h(P)-h(Q)+5\epsilon)}$. This contradiction establishes Fact 5. □

FACT 6: *The likelihood that F is excellent exceeds $1-\delta$.*

Proof: We already know (see (3)) that all but ϵ of the atoms of $P_{-\infty}^0 \mid E$ are good. Fact 3 tells us that if α is good, the likelihood that F_α is good is $> 1-\epsilon$, while Fact 5 tells us that if α is good, the likelihood that F_α is balanced is $> 1-\epsilon$. Hence if we let α vary over all the atoms of $P_{-\infty}^0 \mid E$, the expected μ_E-mass of the atoms α for which α is good and F_α is good and balanced is $> (1-\epsilon)(1-2\epsilon) > 1-3\epsilon$. Since ϵ was chosen $< \delta^5 < \delta^4/3$, we have $1-3\epsilon > 1-\delta^4$, so that with likelihood $> 1-\delta$, the good atoms α for which F_α is good and balanced have total μ_E-measure $> 1-\delta^3$, in which case F is excellent. □

FACT 7: *If α is good and f_α is good and balanced, then $H_{\mu_\alpha}(P_1^n \mid \tilde{Q}_1^n) < n(h(P)-h(Q)+\delta^2)$.*

(Hereafter, we will write H_α instead of H_{μ_α}.)

Proof: Fix α good and f_α good and balanced. Since f_α is good, $\mu_\alpha(\bigcup \mathcal{A}_\alpha^*) > 1-\epsilon$; and since f_α is balanced, the names A that are balanced with respect to f_α have μ_α-measure $> 1-\epsilon$. Therefore, for a set of names $A \in \mathcal{A}_\alpha^*$ of total μ_α-measure $> 1-2\epsilon$, the cardinality of $f_\alpha^{*-1}(f_\alpha^*(A))$ is less than $M = e^{n(h(P)-h(Q)+6\epsilon)}$. Put lexicographic ordering on the collection $\mathcal{A}_\alpha^*$, and for each balanced name A in $\mathcal{A}_\alpha^*$, let $r(A)$ be the rank of the P-name A in the ordered sub-collection $f_\alpha^{*-1}(f_\alpha^*(A))$ of P-names, so that $1 \leq r(A) \leq M$. Then there is a function from $Q_1^n \times \{1,\ldots,M\}$ to P_1^n which takes $(f_\alpha(A), r(A))$ to A, for all balanced names $A \in \mathcal{A}_\alpha^*$.

For $1 \leq i \leq M$, let $R_{(i)} = \bigcup\{A \in \mathcal{A}_\alpha^* : A \text{ is balanced and } r(A) = i\}$, with $R_{(0)}$ consisting of everything else. On each component $R_{(i)}$ with $1 \leq i \leq M$, $f_\alpha(A)$ and i jointly determine A. Recall that $\tilde{Q}_1^n(x) = f_\alpha(P_1^n(x))$ for all $x \in \alpha$. Thus

$$\begin{aligned} H_\alpha(P_1^n \mid \tilde{Q}_1^n \vee R) &= \mu_\alpha(R_{(0)}) H_{R_{(0)}}(P_1^n \mid \tilde{Q}_1^n) \\ &< (2\epsilon)(\log \#(P_1^n)) \\ &= (2\epsilon)(n \log \#(P)) \\ &< n(2\epsilon/\delta) \\ &< n(2\delta^4)\,. \end{aligned}$$

On the other hand,

$$\begin{aligned} H_\alpha(R) &\le \log(1+M) \\ &< n(h(P)-h(Q)+7\epsilon) \\ &< n(h(P)-h(Q)+7\delta^5)\ . \end{aligned}$$

Therefore

$$\begin{aligned} H_\alpha(P_1^n \mid \tilde{Q}_1^n) &= H_\alpha(P_1^n \mid \tilde{Q}_1^n \vee R) + I_\alpha(P_1^n; R \mid \tilde{Q}_1^n) \\ &\le H_\alpha(P_1^n \mid \tilde{Q}_1^n \vee R) + H_\alpha(R) \\ &< n(h(P)-h(Q)+7\delta^5+2\delta^4) \\ &< n(h(P)-h(Q)+\delta^2)\ . \end{aligned}$$

□

FACT 8: *If F is excellent, $h(\tilde{Q}) > h(Q) - \delta$.*

Proof: By Fact 7, the excellence of F implies that

$$H_\alpha(P_1^n \mid \tilde{Q}_1^n) < n(h(P)-h(Q)+\delta^2)$$

for all but a δ^3-fraction of the α's in E. On the remaining α's, we have

$$H_\alpha(P_1^n \mid \tilde{Q}_1^n) \le \log \#(P)^n = n \log \#(P) < n/\delta\ ,$$

so the average value of $H_\alpha(P_1^n \mid \tilde{Q}_1^n)$ is at most

$$\begin{aligned} &(1-\delta^3)\, n\,(h(P)-h(Q)+\delta^2) + (\delta^3)\, n/\delta \\ &= n\left[(1-\delta^3)\,(h(P)-h(Q)+\delta^2)+\delta^2\right] \\ &\le n[h(P)-h(Q)+2\delta^2]\ . \end{aligned}$$

So, using the fact that the partition of E into past-atoms α is $P_{-\infty}^{\ 0} \vee \tilde{Q}_1^n$-measurable, we get

$$H_E(P_1^n \mid P_{-\infty}^{\ 0} \vee \tilde{Q}_1^n) \le n(h(P)-h(Q)+2\delta^2)$$

and a fortiori

$$H_E(P_1^n \mid P_{-\infty}^{\ 0} \vee \tilde{Q}_{-\infty}^{\infty}) \le n(h(P)-h(Q)+2\delta^2)\ .$$

Now, since the partition of X into TE, T^2E, ..., T^nE, $X \setminus T^*E$ is $P_{-\infty}^{-1}$-measurable, we have

$$\begin{aligned} H_\mu(P \mid P_{-\infty}^{-1} \vee \tilde{Q}_{-\infty}^{\infty}) = & \sum_{i=1}^{n} \mu(T^iE) H_{T^iE}(P \mid P_{-\infty}^{-1} \vee \tilde{Q}_{-\infty}^{\infty}) \\ & + \mu(X \setminus T^*E) H_{X \setminus T^*E}(P \mid P_{-\infty}^{-1} \vee \tilde{Q}_{-\infty}^{\infty}) \\ = & \ \mu(E) \left[\sum_{i=1}^{n} H_{T^iE}(P \mid P_{-\infty}^{-1} \vee \tilde{Q}_{-\infty}^{\infty}) \right] \\ & + \mu(X \setminus T^*E) \left[H_{X \setminus T^*E}(P \mid P_{-\infty}^{-1} \vee \tilde{Q}_{-\infty}^{\infty}) \right] . \end{aligned}$$

But the first bracketed expression is equal to $H_E(P_1^n \mid P_{-\infty}^{0} \vee \tilde{Q}_{-\infty}^{\infty})$, which is less than or equal to $n(h(P)-h(Q)+2\delta^2)$, while the second bracketed expression is at most $\log \#(P)$. Also, $\mu(E), \mu(X \setminus T^*E) < 1/n$. Hence

$$\begin{aligned} H_\mu(P \mid P_{-\infty}^{-1} \vee \tilde{Q}_{-\infty}^{\infty}) &\le \frac{1}{n} \left[n(h(P) - h(Q) + 2\delta^2) \right] + \frac{1}{n} \left[\log \#(P) \right] \\ &= h(P) - h(Q) + 2\delta^2 + \frac{\log \#(P)}{n} , \end{aligned}$$

which is less than $h(P) - h(Q) + 3\delta^2 < h(P) - h(Q) + \delta$ for large n.

On the other hand, $H(P \mid P_{-\infty}^{-1} \vee \tilde{Q}_{-\infty}^{\infty}) = h(P \vee \tilde{Q}) - h(\tilde{Q})$ (see [10], p. 66, Theorem 8). Note $h(P \vee \tilde{Q}) = h(P)$ since $\tilde{Q}_{-\infty}^{\infty} \subset P_{-\infty}^{\infty}$. Hence $h(P \vee \tilde{Q}) - h(\tilde{Q}) = h(P) - h(\tilde{Q})$. Thus

$$\begin{aligned} h(P) - h(\tilde{Q}) &= H(P \mid P_{-\infty}^{-1} \vee \tilde{Q}_{-\infty}^{\infty}) \\ &< h(P) - h(Q) + \delta, \end{aligned}$$

which implies $h(\tilde{Q}) > h(Q) - \delta$. □

Combining all of the above ingredients, we get the

COPYING LEMMA: *We are given an ergodic pre-unilateral joining of two dynamical systems with respective finite generators P and Q, such that $h(Q) \le h(P)$. For all $\delta > 0$ and all $r \ge 1$, there exists a partition $\tilde{Q} \subset P_{-\infty}^{0}$ such that*

$$\frac{1}{2} \left| \operatorname{dist} (P \vee \tilde{Q})_1^r - \operatorname{dist} (P \vee Q)_1^r \right| < \delta \qquad \text{and} \qquad |h(\tilde{Q}) - h(Q)| < \delta .$$

Proof: Choose a map F at random, with n sufficiently large; this determines a partition $\tilde{Q}$. By Facts 1 and 2 (together with the remarks following the proof of Fact 2), the probability is close to 1 that the conditions

$$\frac{1}{2} \left| \operatorname{dist} (P \vee \tilde{Q})_1^r - \operatorname{dist} (P \vee Q)_1^r \right| < \delta \qquad \text{and} \qquad h(\tilde{Q}) < h(Q) + \delta$$

are satisfied. On the other hand, by Facts 6 and 8 above, the probability is close to 1 that the condition

$$h(\tilde{Q}) > h(Q) - \delta$$

is satisfied.

Therefore, there exists a $\tilde{Q}$ satisfying all three conditions simultaneously. □

4. The Joining Lemma

In this chapter we will establish the fundamental property of mixing Markov processes ("special finite determinedness") that enables us to prove a unilateral factor-map result. We will first show that under suitable hypotheses, two processes $\tilde{Q}, \tilde{m}_Q$ and Q, m_Q have a "non-stationary one-sided joining" that is tight and satisfies an analogue of pre-unilaterality. We then stationarify the joining and make it two-sided without affecting the pre-unilaterality or the degree of tightness. Finally, we show that the stationary tight pre-unilateral joining may be assumed ergodic without loss of generality.

Outside of this chapter, the approximate copy of Q, m_Q is called $\tilde{Q}, \tilde{m}_Q$; here, however, to avoid a plague of tildes, we will use the name P, m_P instead. This will cause no confusion, since the P, m_P of the main theorem is nowhere in sight throughout the proof of the Joining Lemma.

A (stationary) one-sided joining of two processes P, m_P and Q, m_Q is a measure π defined on $P_0^\infty \vee Q_0^\infty$ that projects to m_P on P_0^∞, projects to m_Q on Q_0^∞, and is invariant under the product action. If the last condition is not satisfied, we call μ a non-stationary one-sided joining. In both cases, we say π is ϵ-tight if $\#(P) = \#(Q)$ and $\pi(P^n \triangle Q^n) \leq \epsilon$ for all n.

A one-sided joining (stationary or not) will be called pre-unilateral if

$$P^{n+1} \perp Q_0^n \mid P_0^n$$

for all $n \geq 0$.

We begin with two information-theoretic lemmas.

INFORMATION LEMMA: *If $I(P; Q \mid R) < \epsilon^4$, then all but ϵ of the atoms $Q_j \cap R_k$ satisfy*

$$\frac{1}{2}\Big|\text{dist}\,(P \mid Q_j \cap R_k) - \text{dist}\,(P \mid R_k)\Big| < \epsilon\,.$$

Proof: In 1967, Csiszár, Kemperman, and Kullback proved (independently of one another; see Theorem 4.1 in [2], Theorem 6.11 in [4], and [5]) that for all

probability vectors $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$,

$$\sum_{k=1}^{n} x_k \log \frac{x_k}{y_k} \geq \frac{1}{2}\left(\sum_{k=1}^{n} |x_k - y_k|\right)^2 ;$$

we will use a weaker version of this estimate, with the constant 1/2 replaced by 1/4. Combining the weakened estimate with Jensen's inequality (and the fact that $(t/2)^2$ is convex in t), we have

$$\begin{aligned}
\epsilon^4 &> I(P; Q \mid R) \\
&= \sum_{j,k} m(Q_j \cap R_k) \sum_i m(P_i \mid Q_j \cap R_k) \log \frac{m(P_i | Q_j \cap R_k)}{m(P_i | R_k)} \\
&\geq \sum_{j,k} m(Q_j \cap R_k) \left(\frac{1}{2} \sum_i \left| m(P_i \mid Q_j \cap R_k) - m(P_i \mid R_k) \right| \right)^2 \\
&\geq \left(\frac{1}{2} \sum_{j,k} m(Q_j \cap R_k) \sum_i \left| m(P_i \mid Q_j \cap R_k) - m(P_i \mid R_k) \right| \right)^2 \\
&= \left(\sum_{j,k} m(Q_j \cap R_k) \frac{1}{2} \left| \operatorname{dist}(P \mid Q_j \cap R_k) - \operatorname{dist}(P \mid R_k) \right| \right)^2 .
\end{aligned}$$

□

CONCAVITY LEMMA: *$H_\mu(P \mid Q)$ is a concave function of μ. That is, if μ_1 and μ_2 are measures on $P \vee Q$ and $\nu = t\mu_1 + (1-t)\mu_2$ with $0 \leq t \leq 1$, then*

$$H_\nu(P \mid Q) \geq t H_{\mu_1}(P \mid Q) + (1-t) H_{\mu_2}(P \mid Q) .$$

Proof: Without loss of generality, we may assume μ_1 and μ_2 are disjointly supported. Let R be a partition of X into two sets that support the respective measures μ_1, μ_2. Then

$$\begin{aligned}
H_\nu(P \mid Q) &\geq H_\nu(P \mid Q \vee R) \\
&= t H_{\mu_1}(P \mid Q) + (1-t) H_{\mu_2}(P \mid Q) .
\end{aligned}$$

□

We now proceed to prove the Joining Lemma. Suppose Q, m_Q is a mixing Markov chain and $\epsilon > 0$; we wish to find $\delta > 0$ such that for any process P, m_P satisfying

$$\begin{aligned}
&\text{(a)} \frac{1}{2} \mid \operatorname{dist} P_0^1 - \operatorname{dist} Q_0^1 \mid < \delta \quad \text{and} \\
&\text{(b)} |h(P) - h(Q)| < \delta ,
\end{aligned}$$

there exists an ϵ-tight pre-unilateral ergodic joining of the two processes. We will construct such joinings in three stages: first, as one-sided, not necessarily stationary joinings; then, as two-sided and stationary but not necessarily ergodic joinings; and lastly, as two-sided, stationary, and ergodic joinings. The properties of ϵ-tightness and pre-unilaterality will hold at all three stages.

For the first stage, our strategy will be to successively construct joining-measures on $P \vee Q$, $P_0^1 \vee Q_0^1$, $P_0^2 \vee Q_0^2$, etc., such that each measure extends the previous one. An atom of $P_1^n \vee Q_1^n$ will be called **matched** if the last P-symbol and the last Q-symbol match, and **mis-matched** if not; we must make sure that for each n, the mis-matched atoms of $P_1^n \vee Q_1^n$ have total measure $\leq \epsilon$.

Our extension procedure will be a modified greedy algorithm; under most circumstances, we will simply join $\text{dist}(P^{n+1} \mid P_0^n)$ and $\text{dist}(Q^{n+1} \mid Q_0^n)$ in as tight a way as possible. This ensures that mis-matches will seldom arise spontaneously. The danger is that when a mis-match does occur, it will force a mis-match at the next stage, and the stage after, and so on — so that, even though spontaneous mis-matches hardly ever happen, they might tend to last a long time. Indeed, this is what happens if you try to join a non-mixing Markov chain with a mixing Markov chain whose distribution and entropy are close; once the two processes "get out of sync", it may take an extremely long time before they match again, no matter how you devise the joining. Fortunately, this cannot happen in the mixing Markov case. When the two processes get out of sync, we let them run independently of one another for a while. The expected time it takes for the two processes to randomly come into matching is finite, and indeed does not depend in any significant way on δ, but only on the statistics of the original process Q, m_Q. Once the two processes match, we continue as before, matching the conditional distributions of P^n and Q^n in as tight a way as possible.

To force our joining $P_0^\infty \vee Q_0^\infty$ to satisfy the pre-unilaterality condition

$$P^{n+1} \perp Q_0^n \mid P_0^n$$

for every n, we must make sure that

$$\text{dist}(P^{n+1} \mid P_0^n \vee Q_0^n) = \text{dist}(P^{n+1} \mid P_0^n) \, .$$

In fact, our constructed joining will also satisfy

$$\text{dist}(Q^{n+1} \mid P_0^n \vee Q_0^n) = \text{dist}(Q^{n+1} \mid Q_0^n) \, ,$$

so that when we pass to the limit, we get a non-stationary one-sided joining in which pre-unilaterality holds "in both directions".

We now begin the proof proper. Let M denote the transition matrix of Q, m_Q. Since the Q-process is mixing, there exists a positive integer r and a real number $\eta > 0$ such that all entries of M^r exceed η.

Let

$$\epsilon_1 = \frac{\eta}{r}\frac{\epsilon}{3}, \quad \epsilon_2 = \frac{1}{2}\epsilon_1^4 .$$

There exists $0 < \delta \leq \epsilon_2$ such that every process P, m_P (with $\#(P) = \#(Q)$) that satisfies (a) also satisfies

$$\frac{1}{2}\left| \operatorname{dist}(P^1 \mid P_{(i)}) - \operatorname{dist}(Q^1 \mid Q_{(i)}) \right| < \epsilon_1 \tag{8}$$

(for all $1 \leq i \leq \#(Q)$) and

$$|H(P^1 \mid P) - H(Q^1 \mid Q)| < \epsilon_2 . \tag{9}$$

(For, $H(P^1 \mid P)$ varies continuously with $\operatorname{dist} P_0^1$, and so does $\operatorname{dist}(P^1 \mid P_{(i)})$ provided we stay away from distributions that give $P_{(i)}$ measure 0; since

$$m_Q(Q_{(i)}) > 0$$

for all i, we are safe if we take δ sufficiently small.) $H(Q^1 \mid Q) = h(Q)$ since Q, m_Q is a Markov process.

Now suppose P, m_P satisfies both (a) and (b) with δ as defined above. Then (9) implies that for all n

$$\begin{aligned} I(P^{n+1}; P_0^{n-1} \mid P^n) &= H(P^{n+1} \mid P^n) - H(P^{n+1} \mid P_0^n) \\ &= H(P^1 \mid P^0) - H(P^1 \mid P_{-n}^0) \\ &\leq H(P^1 \mid P^0) - h(P) \\ &\leq |H(P^1 \mid P^0) - H(Q^1 \mid Q^0)| \\ &+ \ |H(Q^1 \mid Q^0) - h(Q)| \\ &+ \ |h(Q) - h(P)| \\ &< \epsilon_2 + 0 + \delta \\ &\leq 2\epsilon_2 \\ &= \epsilon_1^4 . \end{aligned}$$

This implies, by the Information Lemma, that

$$\frac{1}{2}\left| \operatorname{dist}(P^{n+1} \mid A \cap B) - \operatorname{dist}(P^{n+1} \mid A) \right| < \epsilon_1 \tag{10}$$

with the exception of a set of atoms $A \cap B$ of $P^n \vee P_0^{n-1}$ whose union — call it $E(n)$ — satisfies

$$\mu(E(n)) < \epsilon_1 \ . \tag{11}$$

If $A \cap B$ lies outside the exceptional set $E(n)$, and C is the atom of Q^n corresponding to A, then (by (8) and stationarity)

$$\frac{1}{2} \left| \operatorname{dist}(P^{n+1} \mid A) - \operatorname{dist}(Q^{n+1} \mid C) \right| < \epsilon_1 \ , \tag{12}$$

and (10) and (12) together imply that

$$\frac{1}{2} \left| \operatorname{dist}(P^{n+1} \mid A \cap B) - \operatorname{dist}(Q^{n+1} \mid C) \right| < 2\epsilon_1 \ . \tag{13}$$

Now we create a one-sided joining μ by iteratively constructing joinings of P_0^n and Q_0^n. Informally, we describe the inductive step as follows: If the P-name and Q-name seen so far agree at the most recent symbol, then the distributions at the next step are to be joined in as tight a way as possible. If however they disagree, then the two distributions are to be joined independently at the next step.

More precisely: We first determine μ on $P^0 \vee Q^0$ by joining the two time-0 distributions as tightly as possible; i.e.,

$$\mu(P^0 \triangle Q^0) = \frac{1}{2} \mid \operatorname{dist}(P^0) - \operatorname{dist}(Q^0) \mid \leq \frac{1}{2} \mid \operatorname{dist}(P_0^1) - \operatorname{dist}(Q_0^1) \mid < \delta < \epsilon_1 \ .$$

Now suppose μ has been defined on $P_0^n \vee Q_0^n$, and we wish to extend the definition to $P_0^{n+1} \vee Q_0^{n+1}$. Let A, B, C, D be atoms of $P^n, P_0^{n-1}, Q^n, Q_0^{n-1}$ respectively. Whether or not A and C match (i.e., whether or not $A = P_{(i)}^n$ and $C = Q_{(i)}^n$ for some i), we will put

$$\begin{aligned} \operatorname{dist}_\mu(P^{n+1} \mid A \cap B \cap C \cap D) &= \operatorname{dist}(P^{n+1} \mid A \cap B) \ , \\ \operatorname{dist}_\mu(Q^{n+1} \mid A \cap B \cap C \cap D) &= \operatorname{dist}(Q^{n+1} \mid C \cap D) \\ &= \operatorname{dist}(Q^{n+1} \mid C) \ . \end{aligned}$$

This guarantees that μ will be a (one-sided) joining, yet still leaves us much freedom in determining the joint distribution

$$\operatorname{dist}_\mu(P^{n+1} \vee Q^{n+1} \mid A \cap B \cap C \cap D).$$

If A and C match, then we join P^{n+1} and Q^{n+1} as tightly as possible on $A \cap B \cap C \cap D$, i.e.,

$$\mu(P^{n+1} \triangle Q^{n+1} \mid A \cap B \cap C \cap D) = \frac{1}{2} \left| \operatorname{dist}(P^{n+1} \mid A \cap B) - \operatorname{dist}(Q^{n+1} \mid C) \right| ;$$

if moreover $A \cap B$ lies outside of $E(n)$, this equality implies

$$\mu(P^{n+1} \triangle Q^{n+1} \mid A \cap B \cap C \cap D) < 2\epsilon_1 , \tag{14}$$

by (13). If A and C don't match, then we make P^{n+1} and Q^{n+1} conditionally independent on $A \cap B \cap C \cap D$.

Iterating this construction, we get a non-stationary one-sided joining μ on $P_0^\infty \vee Q_0^\infty$. For all n it satisfies $P^{n+1} \perp Q_0^n \mid P_0^n$, so it is a non-stationary one-sided pre-unilateral joining.

For future convenience, we re-index (14) and write it as

$$\mu(P^n \triangle Q^n \mid B \cap D) < 2\epsilon_1 ; \tag{15}$$

the inequality is achieved whenever the names $B \in P_0^{n-1}$ and $D \in Q_0^{n-1}$ match and B lies outside of $E(n-1)$.

It now remains to show that $\mu(P^n \triangle Q^n) \leq \epsilon$ for all n. To prove this, let $F_k(n)$ $(1 \leq k \leq n+1)$ be the union of the atoms of $(P \vee Q)_0^n$ such that the P-name and Q-name disagree in the last k positions but agree in the position preceding that; thus $P^n \triangle Q^n$ is the union of $F_1(n), F_2(n), \ldots, F_{n+1}(n)$.

First we will show that $F_1(n)$ (the set of mis-matches arising spontaneously at stage n) is small. (Note that with $n = 0$, $\mu(F_1(0)) = \mu(P^0 \triangle Q^0) < \epsilon_1$.) Suppose $A \cap B \cap C \cap D$ (an atom of $P^n \vee P_0^{n-1} \vee Q^n \vee Q_0^{n-1}$) is in $F_1(n)$, so that B and D match. Then either B is in $E(n-1)$ or else $\mu(P^n \triangle Q^n | B \cap D) < 2\epsilon_1$ (by (15)). The atoms $A \cap B \cap C \cap D \subset F_1(n)$ of the former sort have total measure at most $\mu(E(n-1)) < \epsilon_1$ (by (11)), while those of the latter sort have total measure at most $2\epsilon_1$. Thus $\mu(F_1(n)) < 3\epsilon_1$ for all n. Since $F_2(n) \subset F_1(n-1)$, we also get $\mu(F_2(n)) < 3\epsilon_1$ for all n, and similarly $\mu(F_k(n)) < 3\epsilon_1$ for all k, n.

We will now show that $F_{k+r}(n+r)$ is uniformly strictly smaller than $F_k(n)$ ("mis-matches decay with uniformly positive probability in r steps"). Fix $A \cap B \cap C \cap D$ in $F_k(n)$. Then $\mu(F_{k+r}(n+r) \mid A \cap B \cap C \cap D)$ is equal to the probability that if one starts the P-process with past $A \cap B$ and the Q-process with past $C \cap D$ and continues them in an independent way for r steps, they will disagree for *all* r steps. Thus $\mu(F_{k+r}(n+r) \mid A \cap B \cap C \cap D)$ is bounded above by the probability that if one runs the independent continuation of the two processes

starting from respective histories $A \cap B$ and $C \cap D$, they will disagree at the rth step (that is, at time $n+r$). But we chose r so that $\text{dist}(Q^{n+r} \mid C \cap D)$ has all its entries $> \eta$, so regardless of what $\text{dist}(P^{n+r} \mid A \cap B)$ is, the probability exceeds η that the two processes in the independent continuation will agree at time $n+r$. Hence

$$\mu(F_{k+r}(n+r) \mid A \cap B \cap C \cap D) < 1 - \eta$$

for all $A \cap B \cap C \cap D$ in $F_k(n)$, and it follows that $\mu(F_{k+r}(n+r)|F_k(n)) < 1-\eta$ for all k, n.

Applying the inequalities

$$\mu(F_k(n)) < 3\epsilon_1 \quad \text{and} \quad \mu(F_{k+r}(n+r)) < (1-\eta)\mu(F_k(n)),$$

we get

$$\begin{array}{lll} \mu(F_k(n)) & < 3\epsilon_1 & \text{for } k > 0, \\ \mu(F_k(n)) & < 3\epsilon_1(1-\eta) & \text{for } k > r, \\ \mu(F_k(n)) & < 3\epsilon_1(1-\eta)^2 & \text{for } k > 2r, \end{array}$$

etc. Therefore

$$\begin{aligned} \mu(P^n \triangle Q^n) &= \sum_{k=1}^{n+1} \mu(F_k(n)) \\ &< 3\epsilon_1 \cdot (r \cdot 1 + r \cdot (1-\eta) + r \cdot (1-\eta)^2 + \dots) \\ &= 3\epsilon_1 \cdot \frac{r}{\eta} \\ &= \epsilon, \end{aligned}$$

as claimed.

Now, for the second stage, we will see how to turn the not necessarily stationary one-sided ϵ-tight pre-unilateral joining μ into a *stationary two-sided* ϵ-tight pre-unilateral joining $\hat{\mu}$. Because of pre-unilaterality, μ satisfies

$$\begin{aligned} H_\mu(P_{b+1} \mid P_a^b \vee Q_a^b) &\geq H_\mu(P_{b+1} \mid P_0^b \vee Q_0^b) \\ &= H_\mu(P_{b+1} \mid P_0^b) \\ &= H_{m_P}(P_{b+1} \mid P_0^b) \\ &\geq H_{m_P}(P_{b+1} \mid P_{-\infty}^b) \\ &= h(P) \end{aligned}$$

for all $0 \leq a \leq b$. Let μ_n be the n-step translate of μ; i.e.,

$$\text{dist}_{\mu_n}(P \vee Q)_0^k = \text{dist}_\mu(P \vee Q)_n^{n+k}.$$

The μ_n's will be ϵ-tight joinings that satisfy

$$H_{\mu_n}(P^{b+1} \mid P_a^b \vee Q_a^b) \geq h(P) \tag{16}$$

for all $0 \leq a \leq b$. Let

$$\nu_n = \frac{1}{n} \sum_{k=0}^{n-1} \mu_k \ ;$$

ν_n is an ϵ-tight joining, and by (16) and the Concavity Lemma,

$$H_{\nu_n}(P^{b+1} \mid P_a^b \vee Q_a^b) \geq h(P) \tag{17}$$

for all $0 \leq a \leq b$. By the compactness of the space of non-stationary joinings (an easy generalization of Claim 6), the sequence of ν_n's has an accumulation point ν; ν is an ϵ-tight joining, and by (17) and the continuity of dist$(P^{b+1} \mid P_a^b \vee Q_a^b)$,

$$H_{\nu}(P^{b+1} \mid P_a^b \vee Q_a^b) \geq h(P) \tag{18}$$

for all $0 \leq a \leq b$. What is more, ν, being an accumulation point of measures ν_n that approximate stationarity arbitrarily closely for large n, is stationary. Let $\hat{\mu}$ be the natural extension of the stationary one-sided process ν; $\hat{\mu}$ is an ϵ-tight two-sided joining, and by (28) and stationarity,

$$H_{\hat{\mu}}(P^{b+1} \mid P_a^b \vee Q_a^b) \geq h(P) \tag{19}$$

for all $a \leq b$ (not just non-negative values). Replacing b by n and sending $a \to -\infty$, we get

$$\begin{aligned} H_{\hat{\mu}}(P^{n+1} \mid P_{-\infty}^n \vee Q_{-\infty}^n) &\geq h(P) \\ &= H_{\hat{\mu}}(P^{n+1} \mid P_{-\infty}^n) \\ &\geq H_{\hat{\mu}}(P^{n+1} \mid P_{-\infty}^n \vee Q_{-\infty}^n) \ . \end{aligned}$$

Hence the equality $H_{\hat{\mu}}(P^{n+1} \mid P_{-\infty}^n \vee Q_{-\infty}^n) = H_{\hat{\mu}}(P^{n+1} \mid P_{-\infty}^n)$ holds, so that $\hat{\mu}$ is pre-unilateral in the two-sided sense.

Lastly, for the third stage of the proof, note that by Claim 10, almost every ergodic component of $\hat{\mu}$ is pre-unilateral. Also note that since $\hat{\mu}$ is ϵ-tight, a positive fraction of its ergodic components are ϵ-tight. For, if we write $\hat{\mu} = \int \lambda_\alpha \, d\alpha$, as in the proof of Claim 10, then we have $\hat{\mu}(P \triangle Q) = \int \lambda_\alpha(P \triangle Q) \, d\alpha$. If $\hat{\mu}(P \triangle Q) \leq \epsilon$, then certainly $\lambda_\alpha(P \triangle Q) \leq \epsilon$ for a set of α's of positive measure. Thus, $\hat{\mu}$ has an ϵ-tight pre-unilateral ergodic component, proving:

JOINING LEMMA: *If Q, m_Q is a mixing (finite-state) Markov chain and $\epsilon > 0$, then there exists $\delta > 0$ such that for any process P, m_P satisfying*

$$(a) \quad \frac{1}{2} \mid dist\, P_0^1 - dist\, Q_0^1 \mid < \delta \quad \textit{and}$$

$$(b) \quad |h(P) - h(Q)| < \delta \,,$$

there exists an ergodic (two-sided, stationary) ϵ-tight pre-unilateral joining of P, m_P and Q, m_Q. That is, every mixing Markov chain is UFD. □

5. Conclusion of the Proof

Let P, m_P be an ergodic process and Q, m_Q be a UFD process with $h(Q) \leq h(P)$. The Improvement Lemma says that every neighborhood of an ergodic pre-unilateral joining of P, m_P and Q, m_Q contains an ergodic pre-unilateral joining μ' with $H_{\mu'}(Q \mid P_{-\infty}^0) < \eta$. More precisely:

IMPROVEMENT LEMMA: *Given an ergodic pre-unilateral joining μ of P, m_P and Q, m_Q, and given any $r \geq 1$ and any $\epsilon, \eta > 0$, there exists an ergodic pre-unilateral joining μ' of m_P and m_Q such that*

$$\frac{1}{2} \mid dist_{\mu'} (P \vee Q)_0^{r-1} - dist_{\mu} (P \vee Q)_0^{r-1}) \mid < \epsilon$$

and

$$H_{\mu'}(Q \mid P_{-\infty}^0) < \eta \,.$$

Proof: We may assume $r \geq 2$. Given μ, r, ϵ, and η, take ϵ' so small that $r\epsilon' < \epsilon/2$ and

$$\nu(Q \triangle \tilde{Q}) < \epsilon' \quad \text{implies} \quad H_\nu(Q \mid \tilde{Q}) < \eta$$

for all partitions $\tilde{Q}$ with $\#(\tilde{Q}) = \#(Q)$ and all measures ν on $Q \times \tilde{Q}$ (the existence of such an ϵ' follows from Claim 7). Take $\delta \leq \delta(\epsilon')$ (with $\delta(\cdot)$ defined as in the Joining Lemma) satisfying $0 < \delta < \epsilon/2$, and apply the Copying Lemma: there exists $\tilde{Q} \subset P_{-\infty}^0$ such that

$$\frac{1}{2} \mid \text{dist}_\mu (P \vee \tilde{Q})_0^{r-1} - \text{dist}_\mu (P \vee Q)_0^{r-1} \mid < \delta \tag{20}$$

and $|h(\tilde{Q}) - h(Q)| < \delta$. Since $(P \vee \tilde{Q})_0^{r-1}$ and $(P \vee Q)_0^{r-1}$ refine $\tilde{Q}_0^1$ and Q_0^1 (respectively), inequality (20) implies

$$\frac{1}{2} \mid \text{dist}_\mu \tilde{Q}_0^1 - \text{dist}_\mu Q_0^1 \mid \; < \delta \,.$$

Therefore, if we let $\tilde{m}_Q$ denote the restriction of μ to $\tilde{Q}_{-\infty}^{\infty}$, the hypotheses of the Joining Lemma are satisfied, and there exists an ergodic ϵ'-tight pre-unilateral joining ν of $\tilde{Q}, \tilde{m}_Q$ and Q, m_Q.

Regarding μ as a measure on $P_{-\infty}^{\infty} \vee \tilde{Q}_{-\infty}^{\infty} \vee Q_{-\infty}^{\infty}$, let $\tilde{\mu}$ denote the restriction of μ to $P_{-\infty}^{\infty} \vee \tilde{Q}_{-\infty}^{\infty}$. Since μ is ergodic, so are $\tilde{\mu}$ and $\tilde{m}_Q$. Since $\tilde{Q} \subset P_{-\infty}^{0}$ modulo μ, we get $\tilde{Q}_{-\infty}^{0} \subset P_{-\infty}^{0}$, whence $P_1^{\infty} \perp \tilde{Q}_{-\infty}^{0} \mid P_{-\infty}^{0}$ trivially, so that $\tilde{\mu}$ is a pre-unilateral joining of m_P and $\tilde{m}_Q$. Let π be the conditionally independent joining of $\tilde{\mu}$ and ν over $\tilde{m}_Q$ (see Fig. 1(a)). We get

$$P_1^{\infty} \perp \tilde{Q}_{-\infty}^{0} \mid P_{-\infty}^{0}$$

(since $\tilde{\mu}$ is pre-unilateral),

$$\tilde{Q}_1^{\infty} \perp Q_{-\infty}^{0} \mid \tilde{Q}_{-\infty}^{0}$$

(since ν is pre-unilateral), and

$$P_{-\infty}^{\infty} \perp Q_{-\infty}^{\infty} \mid \tilde{Q}_{-\infty}^{\infty}$$

(since π is the conditionally independent joining of $\tilde{\mu}$ and ν). These three facts are easily shown to imply

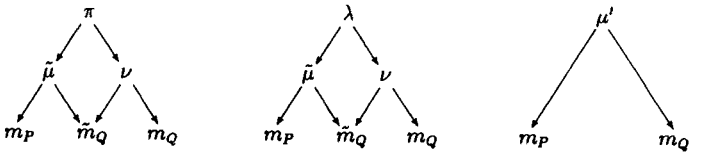
$$P_1^{\infty} \perp (\tilde{Q} \vee Q)_{-\infty}^{0} \mid P_{-\infty}^{0} ,$$

so that π is a pre-unilateral joining of P, m_P and $(\tilde{Q} \vee Q), \nu$. Let λ be an ergodic component of π; since $\tilde{\mu}$ and ν are both ergodic, the diagram in Fig. 1(b) applies almost surely. Furthermore, since π is a pre-unilateral joining of m_P and ν, Claim 10 implies that λ is pre-unilateral also (provided we chose λ outside a set of bad ergodic components of measure 0).

Modulo λ, we have

$$+P_1^{\infty} \perp (\tilde{Q} \vee Q)_{-\infty}^{0} \mid P_{-\infty}^{0} \quad \text{and} \quad P_1^{\infty} \perp Q_{-\infty}^{0} \mid P_{-\infty}^{0} ,$$

so that the restriction μ' of λ to $(P \vee Q)_{-\infty}^{\infty}$ is an ergodic pre-unilateral joining



(a) (b) (c)

Fig. 1.

of P, m_P and Q, m_Q (see Fig. 1(c)). Also,

$$\begin{aligned}
&\frac{1}{2} \mid \mathrm{dist}_\lambda (P \vee Q)_0^{r-1} - \mathrm{dist}_\lambda (P \vee \tilde{Q})_0^{r-1} \mid \\
&= \frac{1}{2} \sum_{A \in (P \vee Q)_0^{r-1}} |\lambda(A) - \lambda(\tilde{A})| \\
&\quad \text{(here } \tilde{A} \text{ denotes the atom of} \\
&\quad (P \vee \tilde{Q})_0^{r-1} \text{ corresponding to } A) \\
&\leq \frac{1}{2} \sum_{A \in (P \vee Q)_0^{r-1}} \lambda(A \triangle \tilde{A}) \\
&\leq r \cdot \frac{1}{2} \sum_{B \in P \vee Q} \lambda(B \triangle \tilde{B}) \\
&\quad \text{(by stationarity)} \\
&= r \cdot \frac{1}{2} \sum_{C \in Q} \lambda(C \triangle \tilde{C}) \quad \text{(sic)} \\
&= r \cdot \nu(Q \triangle \tilde{Q}) \\
&\leq r \cdot \epsilon' \quad \text{(since } \nu \text{ is } \epsilon'\text{-tight)} \\
&< \epsilon/2 \,,
\end{aligned}$$

where the equality marked "sic" holds because B and $\tilde{B}$ are in the same component of P. Therefore

$$\begin{aligned}
&\frac{1}{2} \mid \mathrm{dist}_{\mu'}(P \vee Q)_0^{r-1} - \mathrm{dist}_\mu(P \vee Q)_0^{r-1} \mid \\
&\leq \frac{1}{2} \mid \mathrm{dist}_{\mu'}(P \vee Q)_0^{r-1} - \mathrm{dist}_\mu(P \vee \tilde{Q})_0^{r-1} \mid \\
&\quad + \frac{1}{2} \mid \mathrm{dist}_\mu(P \vee \tilde{Q})_0^{r-1} - \mathrm{dist}_\mu(P \vee Q)_0^{r-1} \mid \\
&= \frac{1}{2} \mid \mathrm{dist}_\lambda(P \vee Q)_0^{r-1} - \mathrm{dist}_\lambda(P \vee \tilde{Q})_0^{r-1} \mid \\
&\quad + \frac{1}{2} \mid \mathrm{dist}_\mu(P \vee \tilde{Q})_0^{r-1} - \mathrm{dist}_\mu(P \vee Q)_0^{r-1} \mid \\
&< \epsilon/2 + \delta \quad \text{(by the previous inequality and (20))} \\
&< \epsilon.
\end{aligned}$$

Lastly, we have

$$\begin{aligned} H_{\mu'}(Q \mid P^0_{-\infty}) &= H_\lambda(Q \mid P^0_{-\infty}) \\ &\leq H_\lambda(Q \vee \tilde{Q} \mid P^0_{-\infty}) \\ &= H_\lambda(Q \mid \tilde{Q} \vee P^0_{-\infty}) \quad (\text{since } H_\lambda(\tilde{Q} \mid P^0_{-\infty}) = 0) \\ &\leq H_\lambda(Q \mid \tilde{Q}) \\ &= H_\nu(Q \mid \tilde{Q}) \\ &< \eta \quad (\text{since } \nu \text{ is } \epsilon' - \text{tight}). \end{aligned}$$

□

COROLLARY 1: *Given an ergodic pre-unilateral joining μ of P, m_P and Q, m_Q and given $\eta > 0$, every distribution neighborhood of μ contains an ergodic pre-unilateral joining μ' such that $H_{\mu'}(Q \mid P^0_{-\infty}) < \eta$.*

Proof: Given an open set U containing the measure μ, there exist $r \geq 1$ and $\epsilon > 0$ such that the "r, ϵ-ball" about μ lies in U. Apply the previous result. □

COROLLARY 2: *The set of joinings μ satisfying $H_\mu(Q \mid P^0_{-\infty}) = 0$ is dense in the set of ergodic pre-unilateral joinings.*

Proof: Let μ_0 be an ergodic pre-unilateral joining, and let U_0 be some neighborhood of μ_0. We must find $\mu_\infty \in U_0$ satisfying $H_{\mu_\infty}(Q \mid P^0_{-\infty}) = 0$ (such a μ_∞ is automatically an ergodic pre-unilateral joining). To do so, we iteratively define $\mu_1, \mu_2, \ldots$ as follows. Suppose measures μ_i and open sets U_i have been defined for all $i < k$. By Corollary 1, there exists an ergodic pre-unilateral joining $\mu_k \in U_{k-1}$ with $H_{\mu_k}(Q \mid P^0_{-\infty}) < 1/k$. Since the distribution topology is metrizable, there exists a neighborhood of μ_k whose closure lies in U_{k-1}; and by semi-continuity (Claim 8), there exists a neighborhood of μ_k in which $H(Q \mid P^0_{-\infty}) < 1/k$. Intersecting these two sets, we get a neighborhood U_k of μ_k such that the closure $\overline{U_k}$ lies in U_{k-1} and

$$H_\mu(Q \mid P^0_{-\infty}) < 1/k \quad \text{for all } \mu \in U_k \, .$$

In this way we get a nested sequence $U_0 \supset U_1 \supset U_2 \supset \ldots$ of open sets, each containing the closure of the next, such that

$$H_\mu(Q \mid P^0_{-\infty}) < 1/n \quad \text{for all } \mu \in U_n \, .$$

By compactness (Claims 6 and 9), there exists

$$\mu_\infty \in \bigcap_{n=1}^{\infty} \overline{U_n} \subset \bigcap_{n=0}^{\infty} U_n \, .$$

Since $H_{\mu_\infty}(Q \mid P^0_{-\infty}) < 1/n$ for all n, we have $H_{\mu_\infty}(Q \mid P^0_{-\infty}) = 0$, as desired. □

Remark: The preceding demonstration is nothing more than a proof of the following variant of the Baire category theorem for complete separable metric spaces: Given a set E and a set O not necessarily contained in E, say that O is "dense in E" if every open set that meets E also meets $O \cap E$. If $O_1, O_2, \ldots$ are open sets dense in E, then their intersection is non-empty, and if moreover $\bigcap_{n=1}^{\infty} O_n$ lies in E, it is dense in E.

COROLLARY 3: *The set of joinings μ satisfying $H_\mu(Q \mid P_{-\infty}^{0}) = 0$ is non-empty.*

Proof: Corollary 2 says that the set of joinings μ satisfying $H_\mu(Q \mid P_{-\infty}^{0}) = 0$ is dense in the set of ergodic pre-unilateral joinings, so it suffices to show that the set of ergodic pre-unilateral joinings is non-empty. But this is clear; e.g., the product joining $m_P \times m_Q$ is an ergodic pre-unilateral joining. □

Hence we have shown:

THEOREM: *If P, m_P is an ergodic process and Q, m_Q is a mixing Markov chain (or other UFD process) such that $h(Q) \leq h(P)$, then Q, m_Q is a unilateral factor of P, m_P.* □

6. Remarks

Let us say that a process Q, m_Q is **inherently unilaterally codable** if for all processes P, m_P, Q, m_Q is a factor of P, m_P if and only if it is a unilateral factor of P, m_P. Sinai's work [15] showed that every independent process is inherently unilaterally codable, and Ornstein and Weiss [9] extended this result to the class of Markov chains with all transition probabilities positive. The current article further extends this result to all mixing Markov chains. Also, by Claim 4 of section 2, all zero-entropy processes are inherently unilaterally codable.

It would be nice to have some negative results to counter-balance these, i.e. examples of processes that are not inherently unilaterally codable (no such examples are currently known). At the same time, one could hope for further progress in the positive direction, showing that ever-more processes are inherently unilaterally codable.

One easy extension of the theorem proved here is to the class of multi-stage mixing Markov chains. Let Q, m_Q be an m-stage Markov chain; then $\hat{Q} = Q_{-m+1}^{0}$ is a Markov partition of the process Q, m_Q, so by our main theorem, if P, m_P is any process with $h(P) \geq h(Q)$, there exists a joining μ such that $\hat{Q}_{-\infty}^{0} \subset P_{-\infty}^{0}$ modulo μ. But since $Q_{-\infty}^{0} = \hat{Q}_{-\infty}^{0}$, this μ gives a unilateral factor map from $P_{-\infty}^{0}$ to $Q_{-\infty}^{0}$.

A less trivial extension is to the class of non-mixing Markov chains. Suppose Q, m_Q is an (irreducible) Markov chain with minimal period d; we would like to know that if P, m_P is any process with entropy $\geq h(Q)$ and with the d-point cycle as a factor (i.e., with the dth roots of unity in its discrete spectrum) then Q, m_Q is a factor of P, m_P. To this end, we must make some modifications in the proof of the main theorem — more specifically, in the copying and joining constructions (sections 4 and 5).

Fix P, m_P and Q, m_Q as in the claim, living on the spaces X and Y, respectively. Since P, m_P and Q, m_Q have the d-point cycle as a factor, there exist partitions $\overline{P} = \{\overline{P}_{(i)} : 1 \leq i \leq d\}$ of X and $\overline{Q} = \{\overline{Q}_{(i)} : 1 \leq i \leq d\}$ of Y such that $T\overline{P}_{(i)} = \overline{P}_{(i+1)}$ and $T\overline{Q}_{(i)} = \overline{Q}_{(i+1)}$ for all i (with $i+1$ interpreted modulo d); moreover, since Q, m_Q is a Markov chain, we may suppose $\overline{Q} \subset Q$, so that the states of the Markov chain are divided into d classes in the usual way. Since $h(\overline{P}) = 0$ and P is a generating partition, we have $\overline{P} \subset P_{-\infty}^{0}$; hence $(P \vee \overline{P})_{-\infty}^{0} = P_{-\infty}^{0}$, and for purposes of constructing a unilateral coding we may as well suppose that P refines $\overline{P}$ (if not, replace P by $P \vee \overline{P}$).

To get a suitable copying construction, one should choose a Rokhlin base E that is confined entirely to a single $\overline{P}_{(i)}$. Then the return-time from E to itself is always a multiple of d, and it is possible to choose a $P_{-\infty}^{0}$-measurable approximation to Q (call it $\tilde{Q}$) such that $\tilde{Q}, \tilde{m}_Q$ not only approximates Q, m_Q in distribution and entropy but also has the property that class i symbols of $\tilde{Q}$ can only be followed by class $i+1$ symbols (modulo $\tilde{m}_Q$ and modulo d).

Such a $\tilde{Q}, \tilde{m}_Q$ is grist for the mill of a modified joining lemma. The key point is that, relative to the periodic factor (of period d) that it has in common with $\tilde{Q}, \tilde{m}_Q$, the process Q, m_Q is uniformly mixing, so that the iterative construction of a non-stationary one-sided joining goes through almost exactly as before. It is only necessary to verify that the d-point factors of Q, m_Q and $\tilde{Q}, \tilde{m}_Q$ are forced to remain in sync; the other details are unaffected.

As a final remark, it is worth noting that the results of this paper translate naturally into the category of endomorphisms of a measure space. In this setting, the full σ-algebra $P_{-\infty}^{\infty}$ is replaced by the one-sided σ-algebra $P_0^\infty = \bigvee_{i=0}^{\infty} T^{-i}P$, and the natural notion of a "factor map" from P, m_P to Q, m_Q is a measure-preserving shift-commuting map from P_0^∞ to Q_0^∞ (not necessarily invertible). The task of finding a factor map is reducible to that of finding a partition $\tilde{Q} \subset P_0^\infty$ such that the $\tilde{Q}$-process obeys the same statistical law as the Q-process. This is just like the unilateral coding problem, except that the roles of past and future have been exchanged. In the case of Markov chains, this time-reversal doesn't matter (a retrograde Markov chain is just another Markov chain); so we

conclude that a mixing Markov endomorphism is a factor (in the endomorphism sense) of every ergodic endomorphism of greater or equal entropy with a finite generating partition. In particular, two one-sided mixing Markov processes of the same entropy are "weakly isomorphic" (mutual factors).

Acknowledgment

I wish to thank Jack Feldman for his many useful suggestions during the course of this research.

References

1. I. P. Cornfeld, S. V. Fomin, and Ya. G. Sinai, *Ergodic Theory*, Springer Verlag, New York, 1980.
2. I. Csiszár, *Information-type measures of difference of probability distributions and indirect observations*, Studia Sci. Math. Hungar. **2** (1967), 299–318.
3. H. Furstenberg, *Recurrence in Ergodic Theory and Combinatorial Number Theory*, Princeton University Press, Princeton, 1981.
4. J. H. B. Kemperman, *On the optimum rate of transmitting information in Probability and Information Theory*, Springer Lecture Notes in Mathematics **89** (1969), 126–169.
5. S. Kullback, *A lower bound for discrimination in terms of variation*, IEEE Trans. Inf. Theory **13** (1967), 126–127.
6. R. McEliece, *The Theory of Information and Coding: A Mathematical Framework for Communication*, Addison-Wesley, Reading, Massachusetts, 1977.
7. B. McMillan, *The basic theorems of information theory*, Ann. Math. Stat. **24** (1953), 196–219.
8. D. Ornstein, *Ergodic Theory, Randomness, and Dynamical Systems*, Yale University Press, New Haven, 1974.
9. D. Ornstein and B. Weiss, *Unilateral codings of Bernoulli systems*, Isr. J. Math. **21** (1975), 159–166.
10. W. Parry, *Topics in Ergodic Theory*, Cambridge University Press, Cambridge, 1981.
11. J. Propp, *Coding from the Past*, Ph.D. thesis, University of California at Berkeley, 1987.
12. J. Propp, *A Shannon-McMillan theorem for motley names*, Isr. J. Math. **69** (1990), 225–234.
13. A. Rothstein and R. Burton, *Isomorphism Theorems in Ergodic Theory* (in preparation).
14. H. L. Royden, *Real Analysis*, Macmillan, New York, 1963.

15. Ya. G. Sinai, *Weak isomorphism of transformations with invariant measure*, Mat. Sb. **63** (1964), 23-42; A.M.S. Translations **57** (1966), 123–143.