# SUBSHIFTS OF MULTI-DIMENSIONAL SHIFTS OF FINITE TYPE

Anthony N. Quas and Paul B. Trow

Abstract. We show that every shift of finite type $X$ with positive entropy has proper subshifts of finite type with entropy strictly smaller than the entropy of $X$, but with entropy arbitrarily close to the entropy of $X$. Consequently, $X$ contains an infinite chain of subshifts of finite type which is strictly decreasing in entropy.

*Introduction*

For a dynamical system $X$, one can ask what are the subsystems of $X$, and what are the possible values of entropies of subsystems of $X$? In the case of a one-dimensional irreducible shift of finite type, the subshifts of finite type in $X$ are characterized by the Krieger embedding theorem ([5], Theorem 10.1.1). But for higher dimensional shifts of finite type, the answers are not known. So we ask some weaker questions: are there infinitely many subshifts of $X$, and if so, do the entropies of those subshifts take on infinitely many values? If the topological entropy of $X$ is positive, the answer to both these questions is yes. In [4], a related question of embedding multi-dimensional subshifts is considered.

In §2, we prove the first of these statements using elementary methods (Corollary 2.7). We then prove a stronger result (which implies both statements), using a theorem due to Ornstein and Weiss on recurrence for stationary random fields. Given a shift of finite type $X$, with positive entropy, we show that there exists a proper subshift of finite type contained in $X$, with entropy arbitrarily close to and less than that of $X$ (see Theorem 2.9).

In §3, we discuss continuous, shift-commuting maps (or *codes*) between higher dimensional shifts of finite type, and generalize some known results for shifts in one dimension (Theorem 3.2).

We would like to thank Kathleen Madden for her helpful comments on an earlier version of this paper.

---

1991 *Mathematics Subject Classification.* 58F03, 28D20.
*Key words and phrases.* entropy, recurrence.

*1. Background*

Let $\mathcal{A}$ be a finite set, $d \in \mathbb{N}$, and let $\Sigma = \mathcal{A}^{\mathbb{Z}^d}$ be the space of all maps $\mathbb{Z}^d \to \mathcal{A}$. Elements of $\mathcal{A}$ are called *symbols*, and points in $\Sigma$ can be thought of as infinite $d$-dimensional arrays of symbols. We give $\Sigma$ the product topology, where $\mathcal{A}$ has the discrete topology. If $x \in \Sigma$ and $S \subseteq \mathbb{Z}^d$, we let $x_S$ denote the restriction of $x$ to $S$. (If $S = \{\mathbf{a}\}$ is a singleton, we simply write $x_{\mathbf{a}}$.) A *configuration* on $S$ is a map $E : S \to \mathcal{A}$. For any subset $X$ of $\Sigma$, a configuration $E : S \to \mathcal{A}$ is said to be *allowed for* $X$ (or simply *allowed*) if there exists $x \in X$ such that $x_S = E$. We say that $E$ *occurs* in $x$. If $B\colon S \to \mathcal{A}$ is any configuration, then we write $[B]$ for $\{x \in X\colon x_S = B\}$. These sets are also known as *cylinder sets*. If $E$ and $F$ are configurations and $S$ and $T$ are disjoint subsets of the domains of $E$ and $F$ respectively, then the configuration $E_S F_T$ is the map $S \cup T \to \mathcal{A}$ which agrees with $E$ on $S$ and with $F$ on $T$.

For each $\mathbf{a} \in \mathbb{Z}^d$, the *shift map* $\sigma_{\mathbf{a}} : \Sigma \to \Sigma$ is defined by $(\sigma_{\mathbf{a}}(x))_{\mathbf{c}} = x_{\mathbf{c}+\mathbf{a}}$. Clearly, $\sigma_{\mathbf{a}}$ is a homeomorphism, and $\sigma_{\mathbf{a}}\sigma_{\mathbf{b}} = \sigma_{\mathbf{b}}\sigma_{\mathbf{a}}$ for all $\mathbf{a}, \mathbf{b} \in \mathbb{Z}^d$. Thus $\mathbb{Z}^d$ acts on $\Sigma$ by homeomorphisms. A closed, non-empty subset $X$ of $\Sigma$ which is invariant under $\sigma_{\mathbf{a}}$, for all $\mathbf{a} \in \mathbb{Z}^d$, is called a *d-dimensional shift space* (or simply a *shift space*). If $Y \subseteq X$ is closed, non-empty and invariant under $\sigma_{\mathbf{a}}$, for all $\mathbf{a} \in \mathbb{Z}^d$, we say that $Y$ is a *subshift* of $X$.

The *k-cube with lowest corner at the origin* is the set $\Lambda(k)$ consisting of all $\mathbf{b} = (b_1, \ldots b_d) \in \mathbb{Z}^d$ such that $0 \leqslant b_i < k$, for $1 \leqslant i \leqslant d$. For $\mathbf{a} \in \mathbb{Z}^d$, the set $\mathbf{a} + \Lambda(k)$ is the *k-cube with lowest corner at* $\mathbf{a}$. The $2k-1$*-cube centered at the origin* is the set $\bar{\Lambda}(2k-1)$ consisting of all $\mathbf{b} = (b_1, \ldots b_d) \in \mathbb{Z}^d$ such that $|b_i| < k$. For any $S \subset \mathbb{Z}^d$, the *border* of $S$, denoted $\partial S$, is the set of $\mathbf{a} \in S$ such that there exists $\mathbf{b}$ in the complement of $S$ with $\|\mathbf{a} - \mathbf{b}\| = 1$, where $\|\cdot\|$ is the usual norm on $\mathbb{R}^d$. The *border* of an allowed configuration $B = x_S$ is defined to be $\partial(B) = x_{\partial S}$.

A *k-block* is an allowed configuration on $\Lambda(k)$. We write $\mathcal{B}_k(X)$ for the set of $k$-blocks and $\mathcal{B}(X)$ for $\bigcup_{k=1}^{\infty} \mathcal{B}_k(X)$. Elements of $\mathcal{B}(X)$ are called *allowed blocks*. If $B$ is a $k$-block and $x \in X$, we say that $B$ *occurs in* $x$ *with lowest corner at* $\mathbf{a} \in \mathbb{Z}^d$ if $B(\mathbf{j}) = x(\mathbf{a}+\mathbf{j})$ for $\mathbf{j} \in \Lambda(k)$. When this holds, we will abuse notation and write $B = x_{\mathbf{a}+\Lambda(k)}$. If $N \geqslant k$, we say that $B$ *occurs in* $x_{\mathbf{b}+\Lambda(N)}$ if there exists $\mathbf{a} \in \mathbb{Z}^d$ such that $\mathbf{a} + \Lambda(k) \subseteq \mathbf{b} + \Lambda(N)$ and $B = x_{\mathbf{a}+\Lambda(k)}$. If $E = x_{\Lambda(N)}$ is an $N$-block, $x \in X$, we say that $B$ *occurs in* $E$ if $B$ occurs in $x_{\Lambda(N)}$.

A shift space $X$ is a *shift of finite type* if there is a finite set $S \subseteq Z^d$ and a non-empty subset $P \subseteq \mathcal{A}^S$ such that

$$X = \{x \in \Sigma : x_{S+\mathbf{a}} \in P \text{ for every } \mathbf{a} \in \mathbb{Z}^d\}.$$

We may think of $P$ as a finite set of allowed finite configurations. $X$ is a *matrix shift* if there is a collection of $d$ transition matrices $A_1, \ldots A_d$, each indexed by $\mathcal{A}(X)$, such that

$$X = X(A_1, \ldots A_d) = \{x \in \Sigma : A_i(x_{\mathbf{a}}, x_{\mathbf{a}+\mathbf{e}_i}) = 1 \text{ for all } \mathbf{a} \in \mathbb{Z}^d,\, 1 \leqslant i \leqslant d\}.$$

where $\mathbf{e}_i$ is the $i$'th standard basis vector. Any shift of finite type is topologically conjugate to a matrix shift. A proof of this for the case $d = 1$ is given in [5], Prop. 2.3.9 (3), and the proof for higher dimensions is similar. For the remainder of this paper, we will assume that all shifts of finite type are presented as a matrix shifts.

If $X$ is a shift space, and $E$ is an allowed block for $X$, then we let $X \setminus E$ denote the set of points in $X$ which do not contain an occurrence of $E$. Formally, $X \setminus E = X \setminus \bigcup_{\mathbf{a}\in\mathbb{Z}^d} \sigma_{\mathbf{a}}([E])$. If $X \setminus E \neq \emptyset$, then it is a subshift of $X$, and if $X$ is a shift of finite type, so is $X \setminus E$.

Let $X$ be a $d$-dimensional shift space and $N$ a positive integer. Define a map $\Gamma_N : X \to \mathcal{B}_N(X)^{\mathbb{Z}^d}$ by $(\Gamma_N(x))_{\mathbf{a}} = x_{N\mathbf{a}+\Lambda(N)}$. The image of $\Gamma_N$ is a subshift of $\mathcal{B}_N(X)^{\mathbb{Z}^d}$, called the *$N$'th power of $X$*, and denoted $X^N$. Any point $x \in X$ corresponds naturally under $\Gamma_N$ to a point $\bar{x} = \Gamma_N(x) \in X^N$, and for any $\mathbf{a} \in \mathbb{Z}^d$, we have $\Gamma_N \sigma^N_{\mathbf{a}} = \sigma_{\mathbf{a}}\Gamma_N$. An allowed $k$-block $E$ for $X^N$ corresponds to an allowed $Nk$-block for $X$ (which we also denote by $E$, abusing notation).

The *topological entropy* of a $d$-dimensional shift space $X$ is defined to be

$$h(X) = \lim_{k\to\infty} \frac{1}{k^d} \log |\mathcal{B}_k(X)|.$$

It is easy to verify that $h(X^N) = Nh(X)$.

The *measure-theoretic entropy* of a shift space $X$ with respect to a $\mathbb{Z}^d$-invariant measure $\mu$ is

$$h_\mu(X) = \lim_{k\to\infty} \frac{1}{k^d} \sum_{B\in\mathcal{B}_k} -\mu[B]\log\mu[B].$$

A shift space $X$ is *irreducible* if for any allowed blocks $U$ and $V$, there is a point $x \in X$ and disjoint sets of coordinates $S$ and $T$ such that $x_S = U$ and $x_T = V$.

We make use at many points of the Variational Principle which is stated below. For a proof of the one-dimensional version, see [8].

**Theorem 1.1.** *Let $G = \mathbb{Z}^d$ act on a compact topological space $X$ by homeomorphisms. Then $h(X) = \sup_{\mu\in M(X)} h_\mu(X)$ where $M(X)$ is the set of $G$-invariant measures on $X$. Further if the action of $G$ is expansive, then the supremum is attained on a non-empty compact set of measures.*

A *measure of maximal entropy* for $X$ is a shift-invariant measure $\mu$ such that $h_\mu(X) = h(X)$. Such measures always exist in the case when $X$ is a shift space as then the action of $\mathbb{Z}^d$ on $X$ is necessarily expansive.

*2. Positive Entropy and Subshifts*

In this section, we show that if $X$ is a shift of finite type, and $h(X) > 0$, then $X$ contains infinitely many subshifts of finite type, all of positive entropy. The key observation for all that follows is contained in the following simple lemma.

**Lemma 2.1.** *Let $X$ be a shift space and let $\mu$ be an invariant probability measure for $X$. If $Y$ is a subshift of $X$ such that $h(Y) < h_\mu(X)$ (or if $Y$ is empty and $0 < h_\mu(X)$), then for any positive integer $M$, there exist $M$ distinct blocks, all having the same border, which are not allowed in $Y$, but are allowed in the support of $\mu$ (and hence have positive $\mu$-measure).*

*Proof.* First, suppose that $Y \neq \emptyset$. Let $\bar{X}$ denote the support of $\mu$. Then

$$h(Y) < h_\mu(X) = h_\mu(\bar{X}) \leqslant h(\bar{X}).$$

Arguing by contradiction, suppose that there is a positive integer $K$ such that for any configuration $H$, which is the border of an allowed block for $\bar{X}$, there are at most $K$ blocks in $\mathcal{B}(\bar{X}) \setminus \mathcal{B}(Y)$ with border $H$. Since $|\partial(\Lambda(n))| \leqslant 2^d n^{d-1}$, it would then follow that $|\mathcal{B}_n(\bar{X}) \setminus \mathcal{B}_n(Y)| \leqslant K|\mathcal{A}|^{2^d n^{d-1}}$, where $\mathcal{A}$ is the symbol set for $X$. Therefore

$$|\mathcal{B}_n(\bar{X})| = |\mathcal{B}_n(Y)| + |\mathcal{B}_n(\bar{X}) \setminus \mathcal{B}_n(Y)| \leqslant |\mathcal{B}_n(Y)|\, K\, |\mathcal{A}|^{2^d n^{d-1}}.$$

Consequently,

$$h(\bar{X}) \leqslant \lim_{n\to\infty} \left[\frac{1}{n^d} \log |\mathcal{B}_n(Y)| + \frac{1}{n^d} \log K|\mathcal{A}|^{2^d n^{d-1}}\right] = h(Y),$$

since the limit of the second term in the sum is 0. Since $h(Y) < h(\bar{X})$, this is a contradiction.

The proof for the case $Y = \emptyset$ is similar. □

**Lemma 2.2.** *Let $X$ be a matrix shift of finite type. Let $B, C$ be distinct blocks such that $\partial(B) = \partial(C)$. For any $N$-block $E$ in which $C$ occurs, there is an $N$-block $F$ in which $B$ occurs, but $C$ does not occur, such that $\partial(E) = \partial(F)$.*

*Proof.* Put an arbitrary order on $\mathcal{A}$, the symbol set of $X$. This extends to a lexicographic order on $\mathcal{B}_m(X)$ for any $m$. Specifically, for $d = 2$, this is defined as follows: $G < H$ if there exists a pair $r, s$ such that $G_{rs} < H_{rs}$, $G_{rj} = H_{rj}$ for $0 \leqslant j < s$ and $G_{ij} = H_{ij}$ for $0 \leqslant i < r$ and $0 \leqslant j \leqslant m-1$.

Let $E$ be a block in which $C$ occurs. Assume $C < B$. Then changing any occurrence of $C$ in $E$ to a $B$ (which can be done since $\partial(B) = \partial(C)$ and $X$ is a matrix shift of finite type), produces a new block $E_1$, with $E < E_1$. Since $\partial(B) = \partial(C)$, we have that $\partial(E_1) = \partial(E)$. If $C$ occurs in $E_1$, we can repeat this procedure, to obtain a block $E_2$, with $E_1 < E_2$. So we obtain a sequence $E < E_1 < E_2 < \dots$, in which no $E_i$ can be repeated and $\partial(E_i) = \partial(E)$. Since there are only finitely many $N$-blocks, we must eventually reach a block $E_k = F$ in which $C$ does not occur. Since $F$ was obtained from $E_{k-1}$ by changing a $C$ to a $B$, it follows that $B$ occurs in $F$. If $B < C$, then a similar argument, with inequalities reversed, proves the result. □

**Corollary 2.3.** *Let $X$ be a shift of finite type. If $h(X) > 0$, then $X$ contains a proper subshift of finite type. In particular, $X$ is not minimal.*

*Proof.* By recoding, we may assume that $X$ is a matrix shift of finite type. By Lemma 2.1, there exist two distinct blocks $B, C$ such that $\partial(B) = \partial(C)$. It follows from Lemma 2.2 and compactness of $X$ that $X \setminus C$ is nonempty, so $X \setminus C$ is a proper subshift of finite type contained in $X$. □

**Lemma 2.4.** *If $X$ is a shift of finite type, and $Y$ is a subshift of $X$ which is not of finite type, then $X$ contains infinitely many subshifts of finite type.*

*Proof.* Since $Y$ is a proper subshift of $X$, there is a block $B_1$ which is allowed for $X$ but not for $Y$. Let $Y_1 = X \setminus B_1$. Then $Y_1$ is a shift of finite type, since $X$ is one. Also, $X \supsetneq Y_1 \supsetneq Y$ since $Y$ is not a shift of finite type. Repeating this argument, we obtain an infinite strictly decreasing sequence $X \supsetneq Y_1 \supsetneq Y_2 \supsetneq \dots$ of subshifts of finite type contained in $X$. □

In Corollary 2.13, we will show that every shift of finite type of positive entropy contains a subshift which is not of finite type.

*Definition 1.*

Given a $k$-block $B$ and a point $x \in X$, let $f_n(B, x)$ denote the number of occurrences of $B$ in $x_{\Lambda(n)}$; that is, the number of points $\mathbf{a} \in \Lambda(n-k)$ such that $x_{\mathbf{a}+\Lambda(k)} = B$. A block $B$ is *positively recurrent* in $x$ if

$$\liminf_{n\to\infty} \frac{f_n(B,x)}{n^d} > 0.$$

If such an $x$ exists, we say that $B$ is *positively recurrent*.

It can be shown that a block $B$ is positively recurrent if and only there exists an ergodic invariant measure $\mu$ on $X$ such that $\mu[B] > 0$.

**Lemma 2.5.** *Let $X$ be a shift space, and let $x \in X$. If $C$ is a $k$-block which is positively recurrent in $x$, then there exists $N \in \mathbb{Z}^+$ such that for any $\mathbf{a} \in \mathbb{Z}^d$, there exists $\mathbf{a}' \equiv \mathbf{a} \pmod{N\mathbb{Z}^d}$ such that $C$ occurs in $x_{\mathbf{a}'+\Lambda(N)}$.*

*Proof.* For simplicity, we will give the proof for the case $d = 2$, the proof of the general case being similar. Arguing by contradiction, suppose that for all $N \in \mathbb{Z}^+$, there exists $\mathbf{a}^N = (a_1^N, a_2^N) \in \mathbb{Z}^d$ such that for any $\mathbf{a}' \equiv \mathbf{a}^N \pmod{N\mathbb{Z}^d}$, we have that $C$ does not occur in $x_{\mathbf{a}'+\Lambda(N)}$. The condition that $C$ does not occur in $x_{\mathbf{a}'+\Lambda(N)}$ implies that the lowest corner of $C$ must lie in a horizontal strip of width $k-1$ below one of the $H_r$ or a vertical strip of width $k-1$ to the left of one of the $V_s$.

We consider $x_{\Lambda(N)}$, the $N$-block of $x$ with lowest corner at the origin. At most one of the vertical lines of the form $V_s$, and at most one of the horizontal lines of the form $H_r$ can intersect $\Lambda(N)$. The vertical and horizontal strips of width $k-1$ about these lines cover at most $2N(k-1)$ of the coordinates of $\Lambda(N)$. It follows that the number of occurrences of $C$ in $x_{\Lambda(N)}$ can be at most $2N(k-1)$. Since $\frac{2N(k-1)}{N^2} \to 0$ as $N \to \infty$, this contradicts the fact that $C$ is positively recurrent in $x$. □

If $E$ is a block and $\mathbf{a} \in \mathbb{Z}^d$, we say that $E$ occurs in $y \in X$ with lowest corner in $\mathbf{a} + N\mathbb{Z}^d$ if there exists $\mathbf{a}' \equiv \mathbf{a} \pmod{N\mathbb{Z}^d}$ such that $E$ occurs in $y$ with lowest corner at $\mathbf{a}'$.

**Theorem 2.6.** *Let $X$ be a shift of finite type. If $Y$ is a subshift of $X$ and $h(Y) < h(X)$ (or if $Y$ is empty and $0 < h(X)$), then there exists a shift of finite type $Z$ such that $Y \subsetneq Z \subsetneq X$ and $h(Z) > 0$.*

*Proof.* We may assume that $X$ is a matrix shift of finite type. Assume that $Y$ is nonempty, the proof in the case $Y = \emptyset$ being similar. Let $\mu$ be an ergodic measure of maximal entropy for $X$, and let $\bar{X}$ denote the support of $\mu$. Then $h(Y) < h(X) = h_\mu(X) = h_\mu(\bar{X})$. By Lemma 2.1, there exist three distinct blocks $B$, $C$ and $D$ in $\mathcal{B}(\bar{X}) \setminus \mathcal{B}(Y)$, such that $\partial(B) = \partial(C) = \partial(D)$. We may assume that $C < B < D$, were $<$ is the lexicographic ordering on blocks, as in the proof of Lemma 2.2. Since $C \in \mathcal{B}(\bar{X})$ we have $\mu[C] > 0$, and it follows from the ergodic theorem that there exists a point $x \in \bar{X}$ in which $C$ is positively recurrent. Let $\mathbf{a}_1, \mathbf{a}_2, \dots \mathbf{a}_{N^d}$ be a complete set of representatives of $\mathbb{Z}^d/N\mathbb{Z}^d$. By Lemma 2.5, there exists $N \in \mathbb{Z}^+$ such that for each $\mathbf{a}_i$, there exists $\mathbf{a}'_i \equiv \mathbf{a}_i \pmod{N\mathbb{Z}^d}$ such that $C$ occurs in $x_{\mathbf{a}'_i+\Lambda(N)}$. Let $x_{\mathbf{a}'_i+\Lambda(N)} = E_i \in \mathcal{B}(\bar{X})$. By Lemma 2.2, for each $i$ there exists a block $F_i$, in which $B$ occurs but $C$ does not, such that $\partial(E_i) = \partial(F_i)$. If $E_i = E_j$, then we can choose $F_i = F_j$.

Let $W$ denote the set of points in $X$ in which $E_i$ does not occur with lowest corner in $N\mathbb{Z}^d$, for $1 \leqslant i \leqslant N^d$. Now $W$ may not be shift invariant, but it is invariant under $\sigma_{\mathbf{a}}$, for $\mathbf{a} \in N\mathbb{Z}^d$, and is therefore a subshift of $X^N$. Let

$$Z' = \bigcup_{\mathbf{a}\in\mathbb{Z}^d} \sigma_{\mathbf{a}}(W). \tag{1}$$

Clearly $Z'$ is shift-invariant. Since $\sigma_{\mathbf{a}}(W) = \sigma_{\mathbf{b}}(W)$ if $\mathbf{a} \equiv \mathbf{b} \pmod{N\mathbb{Z}^d}$, the union in (2.1) can be taken over a complete set of representatives of $\mathbb{Z}^d/N\mathbb{Z}^d$, so $Z'$ is a finite union of closed sets, and therefore closed. Consequently, $Z'$ is a subshift of $X$. Since $C$ occurs in $E_i$ for each $i$, it follows that $E_i$ is not an allowed block for $Y$, and so $Y \subset W \subset Z'$. If $Z'$ is not a shift of finite type (see Example 1), then by the proof of Lemma 2.4, there is a shift of finite type $Z$ with $Z' \subsetneq Z \subsetneq X$ and we show in the final paragraph below that $h(Z) > 0$. So suppose that $Z'$ is a shift of finite type, and let $Z = Z'$.

We show that $Y \neq Z$, by showing that $W$ contains a point in which $B$ occurs. Let $y \in X$ be a point in which some $E_k$ occurs with lowest corner in $N\mathbb{Z}^d$. Let $z$ be the point obtained by changing all occurrences of $E_i$ in $y$, with lowest corner in $N\mathbb{Z}^d$, to $F_i$, for $1 \leqslant i \leqslant N^d$. (This can be done, since $X$ is a matrix shift of finite type, $\partial(E_i) = \partial(F_i)$, and any two such occurrences do not overlap.) Since $C$ does not occur in any $F_i$, but does occur in every $E_i$, we have $F_i \neq E_j$, for all $i, j$. It follows that $E_i$ does not occur in $z$ with lowest corner in $N\mathbb{Z}^d$, for all $i$, so that $z \in W \subset Z$. On the other hand, $F_k$ occurs in $z$, since $E_k$ occurs in $y$. Since $F_k$ contains a $B$, it follows that $z$ is not in $Y$.

Next, we show that $Z \neq X$, by showing that $x$ is not in $Z$. Note that $\sigma_{-\mathbf{a}}(W)$ consists of the set of points in which $E_i$ does not occur with lowest corner in $\mathbf{a}+N\mathbb{Z}^d$, for $1 \leqslant i \leqslant N^d$. By the choice of the blocks $E_i$, for every $\mathbf{a} \in \mathbb{Z}^d$, there is an $i$ such that $E_i$ occurs in $x$ with lowest corner in $\mathbf{a} + N\mathbb{Z}^d$. It follows that $x$ is not in $\sigma_{-\mathbf{a}}(W)$ for any $\mathbf{a} \in \mathbb{Z}^d$, so $x$ is not in $Z$.

Finally, we show that $h(Z) > 0$, by showing that $h(W) > 0$, where $W$ is considered as a subshift of $X^N$. Note that in $X^N$, the blocks $E_i$ and $F_i$ correspond to one-blocks, and $W$ corresponds to the set of points which contain no occurrences of

$E_i$, for $1 \leqslant i \leqslant N^d$. Let $E_m$ be the largest of the $E_i$, in the lexicographic ordering. By the proof of Lemma 2.2, $E_m < F_m$, and $B$ occurs in $F_m$. Now, replace any single occurrence of $B$ in $F_m$ with $D$, to produce a block $G$. Then $\partial(F_m) = \partial(G)$ and we have $E_i \leqslant E_m < F_m < G$, for $1 \leqslant i \leqslant N^d$. Since $E_m \in \mathcal{B}(\bar{X})$, we have $\mu[E_m] > 0$. It follows from the ergodic theorem that there is a point $y \in X^N$ in which $E_m$ is positively recurrent. Now, replace all occurrences of $E_i$ in $y$ with $F_i$, for $1 \leqslant i \leqslant N^d$, to obtain a point $z \in W$. Clearly, $F_m$ is positively recurrent in $z$. Therefore there is an $\epsilon > 0$ such that for all sufficiently large $k$, the number of occurrences of $F_m$ in $z_{\Lambda(k)}$ is at least $\epsilon\, k^d$. Since $G \neq E_i$ for all $i$, and $\partial(G) = \partial(F_m)$, any occurrence of $F_m$ in $z_{\Lambda(k)}$ can be replaced with $G$, and the resulting block is allowed in $W$. It follows that $\mathcal{B}_k(W) \geqslant 2^{\epsilon k^d}$ for sufficiently large $k$. Therefore, $h(W) \geqslant \epsilon \log 2 > 0$. Now, since $W \subset Z$, we have $h(Z) > 0$. □

The reader may wonder why we could not prove Theorem 2.6 by simply letting $Z = X \setminus C$, which is clearly a proper subshift of $X$. The reason is that $X \setminus C$ might equal $Y$, which would be the case if for every point $x \in X$, $C$ occurs in $x$ if and only if $B$ occurs in $x$. Later, we show that given a pair $B, C$, as in Theorem 2.6, we can extend $B$ and $C$ to a pair of larger blocks, $\tilde{B}$ and $\tilde{C}$, with $\partial(\tilde{B}) = \partial(\tilde{C})$, such that $Y \neq X \setminus \tilde{C}$ (see the remark preceding Corollary 2.13).

**Corollary 2.7.** *Let $X$ be a shift of finite type. If $Y$ is a subshift of $X$ and $h(Y) < h(X)$ (or if $Y$ is empty and $0 < h(X)$), then there exists an infinite chain of subshifts of finite type (ordered by inclusion) $Z_i$, with $Y \subsetneq Z_i \subsetneq X$ and $h(Z_i) > 0$ for each $i$.*

*Proof.* Assume that $Y$ is non-empty, the proof in the case $Y = \emptyset$ being similar. By Theorem 2.6, there exists a subshift of finite type $Z$, with $Y \subset Z \subset X$ and $h(Z) > 0$. Proceeding inductively, suppose that we have found a sequence of subshifts of finite type $Y = Z_1 \subsetneq Z_2 \subsetneq \cdots \subsetneq Z_n = X$, and $h(Z_i) > 0$ for $2 \leqslant i \leqslant n$. Since $h(Y) < h(X)$, there exists an $i$ with $h(Z_i) < h(Z_{i+1})$. Then we can apply Theorem 2.6 to obtain a subshift of finite type $Z'$, with $Z_i \subsetneq Z' \subsetneq Z_{i+1}$ and $h(Z') > 0$. The corollary now follows by induction. □

If $X$ is a shift of finite type, and $h(X) > 0$, then Corollary 2.7 implies that $X$ contains infinitely many subshifts of finite type, all having positive entropy.

*Example 1.* We give an example to show that the subshift $Z'$ constructed in Theorem 2.6 may not be of finite type. Let $X$ be the full 2-dimensional shift on two symbols, 0 and 1. Let $E$ and $F$ be two distinct $N$-blocks whose borders consist of all 1s.

As in the proof of Theorem 2.6, let $W$ denote the set of points which contain no occurrence of $E$ with lowest corner in $\mathbb{Z}^2/N\mathbb{Z}^2$. Let

$$Z' = \bigcup_{\mathbf{a} \in \mathbb{Z}^2} \sigma_{\mathbf{a}}(W).$$

Then $Z'$ is not a shift of finite type. To see this, let $a_1, a_2, \ldots, a_{N^2}$ be a complete set of representatives for $\mathbb{Z}^2/N\mathbb{Z}^2$. Then for any positive integer $M$, it is easy to construct a point $y_M$ with the following properties:

(1) For any $\mathbf{a} \in \mathbb{Z}^2$, $y_M$ contains an occurrence of $E$ with lowest corner in $\mathbf{a} + N\mathbb{Z}^2$.
(2) Any two $E$s in $y_M$ occur at a distance at least $2M$ apart.
(3) All symbols in $y_M$ outside an occurrence of $E$ are 0.

Then $y_M$ is not in $Z'$ by property (1), but it is easy to see that every $M$-block in $y_M$ is allowed in $Z'$. The details are left to the reader. This shows that for each $M$, there exists a point, all of whose $M$-blocks are allowed in $Z'$, but such that the point itself does not belong to $Z'$. Consequently, $Z'$ is not a shift of finite type.

In what follows, we work for simplicity in the two-dimensional case ($d = 2$). All of the material generalizes in a straightforward manner to higher dimensions. We will show that by using recurrence arguments due to Ornstein and Weiss that we may control the entropy of $X \setminus C$ and hence create a sequence of such subshifts of $X$ of finite type with entropy close to $X$.

**Proposition 2.8.** *Every shift space $X$ contains an entropy minimal subshift $Y$ with the property that $h(Y) = h(X)$.*

*Proof.* Let $B_1,\ B_2,\dots$ be a list of all possible finite blocks listed in order of increasing size. Set $X_0 = X$ and inductively set

$$X_n = \begin{cases} X_{n-1} \setminus B_n & \text{if } h(X_{n-1} \setminus B_n) = h(X_{n-1}) \\ X_{n-1} & \text{otherwise.} \end{cases}$$

Then $X_n$ is a decreasing sequence of compact shift-invariant subsets of $X$. It follows that the limit $Y = \bigcap_{n=0}^{\infty} X_n$ is also a compact shift-invariant subset of $X$. Since topological entropy is upper semi-continuous for shift spaces (see [5]) and $h(X_n) = h(X)$, it follows that $h(Y) = h(X)$. To show that $Y$ is entropy minimal, suppose for a contradiction that $Y$ has a proper subshift of equal entropy. Then there certainly exists a block $B$ such that $[B] \cap Y \neq \emptyset$ and $h(Y \setminus B) = h(Y)$. The block $B$ must occur in the list as $B_n$ for some $n$ and since $\emptyset \neq [B] \cap Y \subseteq [B] \cap X_n$, we conclude that $[B] \cap X_n \neq \emptyset$. It follows from the definition of $X_n$ that $h(X_{n-1} \setminus B_n) < h(X_{n-1})$ from which we deduce a contradiction as follows: $h(Y \setminus B_n) \leqslant h(X_{n-1} \setminus B_n) < h(X_{n-1}) = h(X)$. The entropy minimality of $Y$ follows. □

Note: It is not known whether the space $Y$ constructed above is itself necessarily a shift of finite type.

**Theorem 2.9.** *Let $X$ be a subshift of finite type with positive topological entropy. Then for all $\epsilon > 0$, there exists a proper subshift $Y$ of $X$ which is also a subshift of finite type with the property that $h(X) - \epsilon < h(Y) < h(X)$.*

We make use in the proof of the theorem of three key results. The first is the Variational Principle, Theorem 1.1.

The second key result is due to Ornstein and Weiss ([6]) and gives a characterization of measure-theoretic entropy in terms of return time.

The return time of the (centered) central $2k-1$-block of $x$, $x_{\bar\Lambda(2k-1)}$ is defined to be

$$R_k(x) = \inf\{n > 0\colon x_{\bar\Lambda(2k-1)} \text{ occurs in } x_{\bar\Lambda(2n-1)}, \text{ other than at } 0\}.$$

**Theorem 2.10.** (Ornstein and Weiss) *If $\mu$ is an ergodic stationary random field, then*

$$\lim_{k\to\infty} \frac{d \log R_k(x)}{(2k-1)^d} = h_\mu(X)$$

*for $\mu$-almost every $x$.*

The final result is a lemma due to Burton and Steif giving an important property of measures of maximal entropy for subshifts of finite type. The proof is in [1], where it is stated for a special class of shifts of finite type. The proof however applies verbatim in the general setting.

**Lemma 2.11.** (Burton and Steif) *If $\mu$ is a measure of maximal entropy for a matrix shift, then for any finite set $G \subset \mathbb{Z}^d$, the conditional distribution of $\mu$ on $G$ given the configuration on $\partial G$ is $\mu$-a.s. uniform over all configurations on $G$ which extend the configuration on $\partial G$.*

The following lemma is then central in the proof of Theorem 2.9.

**Lemma 2.12.** *If $\mu$ is an ergodic invariant measure on $X$ and $B$ is a configuration on $\bar\Lambda(2n-1)$ such that $\mu[B] > 0$, then there exists a configuration $\tilde B$ on $\bar\Lambda(2N-1)$ for some $N > n$ extending $B$ (in the sense that $\tilde B|_{\bar\Lambda(2n-1)} = B$) such that*

*(i)* $\mu[\tilde B] > 0$;

*(ii)* *If $\tilde D = \tilde B_{\bar\Lambda(2N-1)\setminus\bar\Lambda(2n-1)} D_{\Lambda(n)}$ and $\tilde E = \tilde B_{\bar\Lambda(2N-1)\setminus\bar\Lambda(2n-1)} E_{\Lambda(2n-1)}$ are any two configurations (not necessarily distinct) on $\bar\Lambda(2N-1)$ which agree with $\tilde B$ on $\bar\Lambda(2N-1) \setminus \bar\Lambda(2n-1)$, then if $\tilde E$ and $\tilde D$ occur in $x \in \Sigma$, then they either occur at the same point $\mathbf{a} \in \mathbb{Z}^d$ or they occur at two points $\mathbf{a}, \mathbf{b} \in \mathbb{Z}^d$ with the property that $\|\mathbf{a} - \mathbf{b}\|_\infty \geqslant N + n - 1$.*

The following figure illustrates the closest that two blocks, $\tilde D$ and $\tilde E$, as described in the lemma above may occur in a single point $x$ of $X$. The small squares are the central $2n-1$-blocks of $\tilde D$ and $\tilde E$.

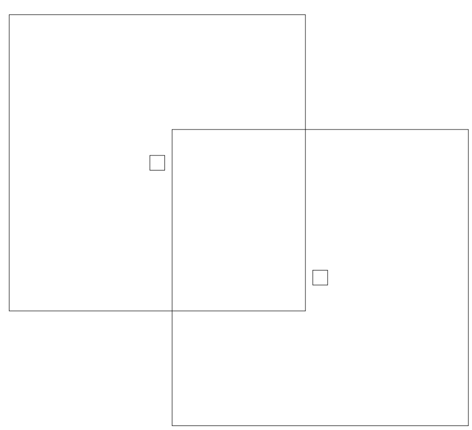

Figure 1. Two padded blocks with minimum separation

*Proof.* Let $\epsilon$ be less than $\mu[B]/8$. Since $\mu$ is an ergodic invariant measure on $X$, Theorem 2.10 applies so that

$$\lim_{k\to\infty} \frac{d \log R_k(x)}{(2k-1)^d} = h_\mu(X)$$

for $\mu$-almost every $x$. It follows that for almost every $x$, there exists an $n(x)$ such that for $k > n(x)$, $d \log R_k(x)/(2k-1)^d \geqslant h_\mu(X)/2$. This may be rewritten

$$\log R_k(x) \geqslant \exp\left(\frac{h(2k-1)^d}{2d}\right).$$

In particular, there exists an $m(x) > n(x)$ such that $\log R_k(x) \geqslant 10k$ for each $k \geqslant m(x)$. Letting $m(x)$ be the minimum such number, the function $m(x)$ becomes measurable and integer-valued. It follows that there exists an $M$ such that $m(x) \leqslant M$ for $x$ belonging to a set of measure at least $1-\epsilon$. Write $S$ for $\{x\colon m(x) \leqslant M\}$. The complement of this set has measure at most $\epsilon$.

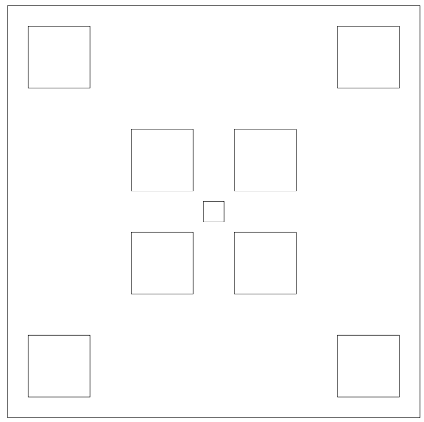

Figure 2. Arrangement of $\tilde{B}$

Consider the arrangement of blocks shown in Figure 2. The (smaller) central block is a centered square of side $2n-1$ placed at the origin and the surrounding (*marker*) squares are of side $2M-1$ with centers at $\mathbf{a}_1, \mathbf{a}_2, \dots \mathbf{a}_8$. The surrounding larger squares are separated by distances greater than $2n-1$. The frame containing all of the squares is the subset $F = \Lambda(2N-1)$ of $\mathbb{Z}^2$. The frame is constructed so as to ensure that $2N < 10M$. We will call the central part the *core*

Now forming

$$A = [B] \cap \bigcap_{i=1}^{8} \sigma_{\mathbf{a}_i}(S) = [B] \setminus \left( \bigcup_{i=1}^{8} \sigma_{\mathbf{a}_i}(S^c) \right),$$

we see that $\mu(A) > 0$. We let $x$ be any point in $A$ and set $\tilde{B} = x|_F$. The block $\tilde{B}$ now has the property that none of its marker squares are repeated within $\tilde{B}$. We then show that this set has the required properties. By elementary measure theory, for almost all $x \in A$, the set $\tilde{B}$ as defined above has positive measure so we can ensure that (i) is satisfied.

To show that (ii) is satisfied, suppose that $\tilde{D}$ and $\tilde{E}$ are two configurations as in the statement of the lemma. Suppose also that they occur in a point $x$ at positions $\mathbf{a}$ and $\mathbf{b}$ with $0 < \|\mathbf{a}-\mathbf{b}\|_\infty < N+n+1$. It follows that in $x$, the core of $\tilde{D}$ overlaps the frame of $\tilde{E}$. We will then show that this contradicts the recurrence properties of points in $A$. To establish the contradiction, we distinguish three modes of overlapping of $\tilde{D}$ and $\tilde{E}$ as follows:

Case (1): The corner of the frame of $\tilde{E}$ lies within the core of $\tilde{D}$;

Case (2): The corners of the frame of $\tilde{E}$ lie outside the core of $\tilde{D}$, but the core of $\tilde{D}$ is not entirely contained within the frame of $\tilde{E}$;

Case (3): The core of $\tilde{D}$ lies entirely within the frame of $\tilde{E}$.

These possibilities are illustrated in Figures 3(1), 3(2) and 3(3).

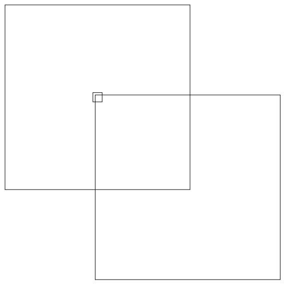

Figure 3(1).

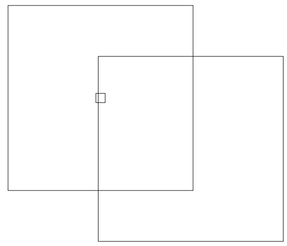

Figure 3(2).

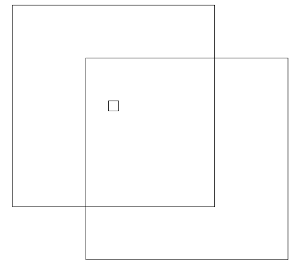

Figure 3(3)

We will show that in each of these cases, the intersection is such that there is a marker square of $\tilde{E}$ entirely contained within the frame of $\tilde{D}$, but not intersecting the core of $\tilde{D}$. This will provide the required contradiction.

In cases (1) and (2), we assume (without loss of generality) that the top of $\tilde{E}$ lies above the bottom of the core of $\tilde{D}$ and the left side of $\tilde{E}$ lies to the left of the right of the core of $\tilde{D}$. This is as shown in Figure 4. The top left middle marker square of $\tilde{E}$ is then completely contained within the frame of $\tilde{D}$, but does not intersect the core of $\tilde{D}$.

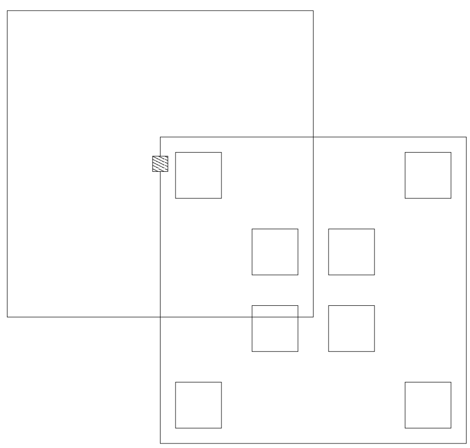

Figure 4.

The case (3) is similar. A typical configuration is shown in Figure 5. Here (assuming again that the top left corner of $\tilde{E}$ lies in the top left quadrant of $\tilde{D}$), we see that both of the marker squares lying in the top left quadrant of $\tilde{E}$ are entirely within the frame of $\tilde{D}$. Since the gap between the marker squares is greater than $2n-1$, it follows that at least one of them does not intersect the core of $\tilde{D}$.

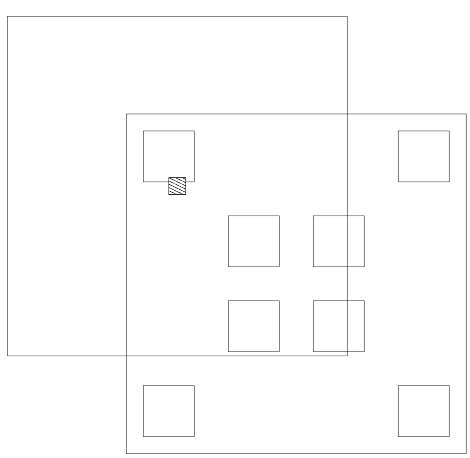

Figure 5.

The marker square in question now occurs twice in $\tilde{D}$ with neither occurrence overlapping the core. It follows that the marker square in question occurs twice in $\tilde{B}$ and this gives the required contradiction. □

*Proof of Theorem 2.9.* Let $X$ be a subshift of finite type with positive topological entropy. Then by Proposition 2.8, $X$ has a subshift $X_0$ with the property that $h(X_0) = h(X)$ and that $X_0$ is entropy minimal. That is if $Y$ is any proper subshift of $X_0$, then $h(Y) < h(X_0) = h(X)$.

Since any subshift is expansive, it follows that the set of measures of maximal entropy is a non-empty compact convex set. Let $\mu$ be any ergodic measure of maximal entropy on $X_0$. Then $\mu$ may also be regarded as an invariant measure on $X$. It is also a measure of maximal entropy on $X$. The support of $\mu$ is $X_0$ (as otherwise the support of $\mu$ is a subshift of $X_0$ of entropy at least $h_\mu(X_0) = h(X_0)$ by the variational principle).

Now by Lemma 2.1, since $h_\mu(X_0) > 0$, there exist $n$-blocks $B$ and $C$ of positive measure with the property that $\partial B = \partial C$ and $\mu[B] > 0$. We now apply Lemma 2.12 to get an extension $\tilde{B}$ of $B$ with the required properties. We then define $\tilde{C} = \tilde{B}_{\bar{\Lambda}(2N-1)\setminus\bar{\Lambda}(2n-1)}C_{\Lambda(2n-1)}$. The block $\tilde{C}$ is an allowed $N$-block because all the adjacent pairs outside the core are allowed as they occur in $\tilde{B}$; the adjacent pairs in $\tilde{C}$ consisting of one symbol from the core and one from outside the core are allowed as they occur in $\tilde{B}$ (because $\partial B = \partial C$); and the adjacent pairs in the core of $\tilde{C}$ are allowed as they occur in $C$. Since $\partial\tilde{B} = \partial\tilde{C}$ and $\tilde{C}$ is an allowed block, it follows from Lemma 2.11 that $\mu[\tilde{C}] = \mu[\tilde{B}] > 0$.

Next, we show that $X \setminus \tilde{C}$ is non-empty. We note that now if we take a configuration $x$ in $X$ and replace $\tilde{C}$s with $\tilde{B}$s, then we no longer have the difficulty which was present in the context of Lemma 2.2. Changing a $\tilde{C}$ to a $\tilde{B}$ cannot produce any new $\tilde{C}$s because if it did, the new $\tilde{C}$ would have to overlap the core of the replacement $\tilde{B}$ which contradicts Lemma 2.12. It follows that we can take a point $x \in X$ and simultaneously replace all $\tilde{C}$ blocks by $\tilde{B}$ blocks to get a new point $x' \in X$ with no $\tilde{C}$s, but $\tilde{B}$s in each of the places where $\tilde{C}$ previously occurred. It follows that $X \setminus \tilde{C}$ is non-empty.

This argument can be modified to show that $X_0 \setminus \tilde{C}$ is non-empty as follows: Since the support of $\mu$ is $X_0$, for every $x \in X_0$ we have $\mu[x|_{\bar{\Lambda}(2k-1)}] > 0$ for each $k$. By Lemma 2.11, replacing the $\tilde{C}$s by $\tilde{B}$s in $x|_{\bar{\Lambda}(2k-1)}$ gives a point $x'$ such that $\mu[x'|_{\bar{\Lambda}(2k-1)}] > 0$. Since $\mu$ is concentrated on $X_0$, it follows that $x' \in X_0$.

Since $\mu[\tilde{C}] > 0$, $X_0 \setminus \tilde{C}$ is a proper subshift of $X_0$. It follows from the entropy minimality of $X_0$ that $h(X_0 \setminus \tilde{C}) < h(X)$.

We now seek a lower bound for the entropy of $X_0 \setminus \tilde{C}$ and will then produce a subshift of finite type which is a proper subshift of $X$ which has the required entropy properties. We will use in this part certain properties of measure-theoretic entropy. The relevant material is contained (in the one-dimensional case) in [8]. The proofs are the same in higher dimensions.

To this end, let $S = \bigcup_{\mathbf{a}\in\bar{\Lambda}(2n-1)} \sigma_{-\mathbf{a}}([\tilde{B}] \cup [\tilde{C}])$. We then define two partitions of $X_0$ as follows: $\mathcal{P}_1 = \{[i] \cap S^c \colon i \in \mathcal{A}\} \cup \{S\}$ and $\mathcal{P}_2 = \{[\tilde{B}], [\tilde{C}], (X \setminus ([\tilde{B}] \cup [\tilde{C}]))\}$. Then we see that $\mathcal{P} = \mathcal{P}_1 \vee \mathcal{P}_2$ is a generating partition as follows: If we know which element of $\mathcal{P}$ each of $\sigma_{\mathbf{a}}(x)$ lies in for $\mathbf{a}$ running over $\bar{\Lambda}(2N-1)$, then we know $x_0$. To see this, note that if $x_0$ lies inside the core of a $\tilde{C}$ or $\tilde{B}$ occurrence, then knowing the element of the $\mathcal{P}_2$ partition which the centralized copy of $x$ lies in tells us whether it is in a $\tilde{B}$ or a $\tilde{C}$ and this allows us to determine the value of $x_0$. If $x_0$ does not lie inside the core of a $\tilde{C}$ or $\tilde{B}$ occurrence, then knowing the element of the $\mathcal{P}_1$ partition in which $x$ lies tells us the value of $x_0$. It follows that the partition $\mathcal{P}$ generates the Borel $\sigma$-algebra of $X_0$. Putting this together, we have

$$\begin{aligned} h(X) = h(X_0) = h_\mu(X_0) &= h_\mu(X_0, \mathcal{P}_1 \vee \mathcal{P}_2) \\ &\leqslant h_\mu(X_0, \mathcal{P}_1) + H_\mu(\mathcal{P}_2|\mathcal{P}_1) \\ &\leqslant h_\mu(X_0, \mathcal{P}_1) + H_\mu(\mathcal{P}_2). \end{aligned}$$

This yields $h_\mu(X_0, \mathcal{P}_1) \geqslant h(X) - H_\mu(\mathcal{P}_2)$. Write $\theta$ for the map sending $X_0$ to $X_0 \setminus \tilde{C}$ by replacing each occurrence of $\tilde{C}$ with a $\tilde{B}$ (simultaneously). Then we see that $\theta$ is a conjugacy between the systems $(X_0, \bigvee_{\mathbf{a}\in\mathbb{Z}^2} \sigma_{\mathbf{a}}\mathcal{P}_1, \mu)$ and $(X_0 \setminus \tilde{C}, \mathcal{B}, \mu \circ \theta^{-1})$ where $\mathcal{B}$ is the Borel $\sigma$-algebra of $X_0 \setminus \tilde{C}$. It follows that

$$\begin{aligned} h(X_0 \setminus \tilde{C}) \geqslant h_{\mu\circ\theta^{-1}}(X_0 \setminus \tilde{C}) &= h_\mu(X_0, \mathcal{P}_1) \\ &\geqslant h(X) - H_\mu(\mathcal{P}_2). \end{aligned}$$

Now let $\epsilon > 0$ be given. Since $H_\mu(\mathcal{P}_2) = -2\mu[\tilde{C}]\log(\mu[\tilde{C}]) - (1-2\mu[\tilde{C}])\log(1-2\mu[\tilde{C}])$, we may ensure that $H_\mu(\mathcal{P}_2) < \epsilon$ if we can ensure that $\tilde{C}$ has arbitrarily small measure. But $\tilde{C}$ had the property that any two occurrences must be at least $N+n+1$ apart, so it follows that $\mu[\tilde{C}] \leqslant (N+n+1)^{-2}$. In particular by ensuring that $N$ is sufficiently large, $H_\mu(\mathcal{P}_2)$ may be made less than $\epsilon$. We now have $h(X) - \epsilon < h(X \setminus \tilde{C}) < h(X)$. There exists a decreasing sequence of subshifts $Y_n$ of $X$ which are shifts of finite type and satisfy $X_0 \setminus \tilde{C} = \bigcap_{n=1}^\infty Y_n$. It is then known (see [5], prop. 4·4·6) that $h(Y_n) \to h(X_0 \setminus \tilde{C})$. In particular, there exists an $n$ such that $h(X_0 \setminus \tilde{C}) \leqslant h(Y_n) < h(X)$. This completes the proof. □

*Remark* With $\tilde{C}$ and $\tilde{B}$ defined as above, changing a $\tilde{C}$ to a $\tilde{B}$ in a configuration $x$ neither creates nor destroys any existing $\tilde{C}$s. Formally if $\tilde{C}$ occurs in $x$ with left corner at $\mathbf{a}$ and if $y$ is the corresponding point with $\tilde{C}$ replaced by $\tilde{B}$, then for each $\mathbf{b} \in \mathbb{Z}^2 \setminus \{\mathbf{a}\}$, $x$ has a $\tilde{C}$ at $\mathbf{b}$ if and only if $y$ has a $\tilde{C}$ at $\mathbf{b}$.

**Corollary 2.13.**
*If $X$ is a shift of finite type and $h(X) > 0$, then $X$ has a subshift $Y$ which is not of finite type.*

*Proof.* Let $\tilde{C}$ and $\tilde{B}$ be as in Theorem 2.9. Then let $Y$ be the subset of $X$ consisting of those points of $X$ in which $\tilde{C}$ occurs at most once. Then $Y$ is shift-invariant, non-empty (it contains $X \setminus \tilde{C}$) and is closed (since the limit of points containing no more than one $\tilde{C}$ contains no more than one $\tilde{C}$). To see that $Y$ is not of finite type, observe that for each $N$, there is a point $x$ of $X$ containing exactly two $\tilde{C}$s, separated by at least $N$ (this follows from the above remark). Then all $N$-blocks in $x$ are allowed $N$-blocks in $Y$, but $x$ does not belong to $Y$. □

*3. Factor Maps and Diamonds*

In this section we discuss continuous, shift-commuting maps between shift spaces, and generalize some known results for one-dimensional shifts (see [4], Chapters 8-10).

Suppose that $X$ and $Y$ are $d$-dimensional shift spaces. Let $k$ be a positive integer, and $g : \mathcal{B}_{2k-1}(X) \to \mathcal{B}_1(Y)$ a map on finite blocks. A *sliding block code* (or simply a code) is a map $f : X \to Y$ defined by $f(x)_{\mathbf{a}} = g(x_{\mathbf{a}+\bar{\Lambda}(2k-1)})$ for $x \in X$. If $k = 0$, we say that $f$ is a *one-block code*.

By an easy generalization of the Curtis-Hedlund-Lyndon Theorem, any continuous map $f : X \to Y$ such that $f\sigma_{\mathbf{a}} = \sigma_{\mathbf{a}} f$, for all $\mathbf{a} \in \mathbb{Z}^d$, is a sliding block code for some $k$ (see [5], Theorem 6.2.9 for a proof in dimension one). Any sliding block code can be recoded to a one-block map; that is, there is a shift space $\bar{X}$, a one-block code $\bar{f} : \bar{X} \to Y$ and a conjugacy $\alpha : X \to \bar{X}$ such that $\bar{f}\alpha = f$. See [5], Prop. 1.5.12 for a proof in dimension one. A surjective code is also known in the literature as a factor map.

If $f : X \to Y$ is a code between shift spaces, a *diamond* for $f$ is a pair of points $x, y \in X$ such that there is a finite subset $S \subseteq \mathbb{Z}^d$, with $f(x) = f(y)$ and $x_S \neq y_S$, $x_{\mathbf{a}} = y_{\mathbf{a}}$ for $\mathbf{a} \in \mathbb{Z}^d \setminus S$. If no such pair exists, we say that $f$ has no diamonds. If $X$ is a shift of finite type and $f$ is a one-block code, this definition is equivalent to saying that there is a pair of allowed blocks $B \neq C$, with $\partial(B) = \partial(C)$, such that $f(B) = f(C)$.

**Proposition 3.1.** *(see [5], Theorem 8.1.16). Let $f : X \to Y$ be a code, where $X$ is an irreducible shift of finite type. If $f$ is countable-to-one, then it has no diamonds.*

*Proof.* We may assume that $f$ is a one-block code. Suppose $f$ has a diamond, so that there is a pair of allowed $k$-blocks $B, C$, with $\partial(B) = \partial(C)$, such that $f(B) = f(C)$. Since $X$ is irreducible, there exists a point in which $B$ occurs infinitely often, in non-overlapping coordinates. Since the block $B$ can be replaced by $C$ wherever it occurs, and $f(B) = f(C)$, we obtain an uncountable collection in $f^{-1}(y)$. □

For one-dimensional shift spaces (not necessarily shifts of finite type), the converse is true, and in fact $f$ must be uniformly finite-to-one ([5], Theorem 8.1.16). But for higher dimensional shifts, the converse is false. For example, if $X$ is the three-dot system (see [3] for a definition), and $f$ collapses all points in $X$ to a single fixed point, then $f$ is uncountable-to-one, since $X$ is uncountable. However, $f$ has no diamonds, since no two distinct $N$-blocks for $X$ can have the same border.

**Theorem 3.2.** *(see [5], Theorem 8.1.16). Let $X$ be a shift of finite type, $Y$ be a shift space and $f : X \to Y$ be a surjective code. If $f$ has no diamonds, then $h(X) = h(Y)$. If $X$ is entropy minimal, the converse holds.*

*Proof.* We may assume that $f$ is a one-block code. Suppose that $f$ has no diamonds. Let $\mathcal{A}_X$ denote the alphabet of $X$. Then for any $B \in \mathcal{B}_k(Y)$, we have $|f^{-1}(B)| \leqslant |\mathcal{A}_X|^{2^d k^{d-1}}$, since $|\partial(\Lambda(k))| \leqslant 2^d k^{d-1}$ and no two preimages of $B$ can share the same border. It follows that

$$|\mathcal{B}_k(X)| \leqslant |\mathcal{A}_X|^{2^d k^{d-1}} |\mathcal{B}_k(Y)|.$$

Therefore

$$\frac{1}{k^d} \log |\mathcal{B}_k(X)| \leqslant \frac{1}{k^d} \log |\mathcal{A}_X|^{2^d k^{d-1}} |\mathcal{B}_k(Y)| = \frac{2^d}{k} \log |\mathcal{A}_X| + \frac{1}{k^d} \log |\mathcal{B}_k(Y)|.$$

Since $2^d \log |\mathcal{A}_X|$ is constant, the first term in the sum on the right tends to 0 as $k \to \infty$, and so $h(X) = h(Y)$.

Now assume that $X$ is an entropy minimal shift of finite type. Suppose $f$ has a diamond $B, C$, where $B$ and $C$ are $N$-blocks. Then $f$ restricts to a map $X \setminus C \to Y$, and we claim that the restriction is still onto. To see this, observe that if $f(x) = y$, and we change any occurrence of $C$ in $x$ to $B$, the resulting point still maps to $y$. Let $y \in Y$. Since $f$ is surjective, there exists $x \in f^{-1}(y)$. Now, following the proof of Lemma 2.2, for any positive integer $N$ we may replace all occurrences of $C$ in $x_{\bar{\Lambda}(2N-1)}$ with $B$, so that the resulting point $x^N \in f^{-1}(y)$, and $x^N_{\bar{\Lambda}(2N-1)}$ contains no occurrences of $C$. Now, by compactness, choose a limit point $x'$ of the sequence $x^N$. Then $x' \in X \setminus C$ and $f(x') = y$. Therefore $f|_{X \setminus C}$ is surjective.

It follows that $h(X \setminus C) \geqslant h(Y)$. Since $X$ is entropy minimal, we have $h(X) > h(X \setminus C) \geqslant h(Y)$. Therefore, $h(Y^N) \leqslant h(X^N \setminus C) < h(X^N)$, and so $h(Y) < h(X)$. □

If $f : X \to Y$ is a surjective, entropy preserving code (i.e. $h(X) = h(Y)$), and $\nu$ is a measure of maximal entropy for $Y$, then there exists an invariant measure $\mu$ for $X$ such that $\hat{f}\mu = \nu$, where $\hat{f}$ is the induced map on measures (see [2], Theorem 1.1). It is easy to see that $\mu$ is a measure of maximal entropy for $X$. Consequently, an entropy preserving code is measure preserving for some measures of maximal entropy for $X$ and $Y$. If $X$ and $Y$ are one-dimensional irreducible shifts of finite type, they have unique measures of maximal entropy, which are preserved by $f$ ([7]and [2]). But in higher dimensions there may be more than one measure of maximal entropy ([1]).

**Lemma 3.3.** *Let $X$ be a shift of finite type, $Y$ be a shift space and $f : X \to Y$ be a surjective code. Suppose that $Y$ is entropy minimal, $f$ has no diamonds and $h(X) = h(Y)$. Then $f$ is onto.*

*Proof.* Assume that $f$ is a one-block code. We have $h(X) = h(f(X))$, by Theorem 3.2. Therefore $h(f(X)) = h(Y)$. Since $Y$ is entropy minimal, we must have $f(X) = Y$. □

The following is a generalization of [5], Corollary 8.1.20], in which $f$ having no diamonds replaces $f$ being finite-to-one.

**Theorem 3.4.** *Let $f : X \to Y$ be a code between entropy minimal shifts of finite type. Then any two of the following statements implies the third.*

*(1) $f$ has no diamonds.*
*(2) $f$ is surjective.*
*(3) $h(X) = h(Y)$.*

*Proof.* The fact that (1) and (2) implies (3), and that (2) and (3) imply (1), follows from Theorem 3.2. That (1) and (3) imply (2) follows from Lemma 3.3. □

Sliding block codes are important in the classification of shifts of finite type up to finite equivalence (see [5], Section 8.3). Two shift spaces are *finitely equivalent* if there is a shift of finite type which is a common finite-to-one extension of both of them ([5], Def. 8.3.1). It is known that one-dimensional shifts of finite type are finitely equivalent if and only if they have the same entropy ([5], Theorem 8.3.7). In higher dimensions this is false: for example, no infinite zero-entropy shift of finite type, such as the three-dot system, can be finitely equivalent to a shift consisting of a finite periodic orbit. So we ask, under what conditions are two shift spaces finitely equivalent? We can define a weaker equivalence relation by saying that two shift spaces $X_1$ and $X_2$ are *equal-entropy equivalent* if there is a shift of finite type $W$ and surjective codes $\phi_1 : W \to X_1$ and $\phi_2 : W \to X_2$ which preserve entropy. We conclude by asking the following question: if $X_1$ and $X_2$ are shifts of finite type, with $h(X_1) = h(X_2)$, are they equal-entropy equivalent?

## References


1. R. Burton and J. Steif, *Non-uniqueness of measures of maximal entropy for subshifts of finite type*, Ergodic Theory Dynamical Systems **14** (1994), 213–235.
2. E. Coven and M. Paul, *Endomorphisms of irreducible shifts of finite type*, Math. Sys. Theory **8** (1974), 167–175.
3. F. Ledrappier, *Un champ Markovien peut être d'entropie nulle et mélangeant*, C. R. Acad. Sc. Paris, Ser. A Nol. **287**, 561–562.
4. S. Lightwood, *PhD dissertation: University of Maryland 1998.*
5. D. Lind and B. Marcus, *An Introduction to Symbolic Dynamics and Coding*, Cambridge University Press, New York, 1995.
6. D. S. Ornstein and B. Weiss, *Preprint: Entropy and recurrence rates for stationary random fields* (1995).
7. W. Parry, *Intrinsic Markov chains*, Trans. Amer. Math. Soc. **112** (1964), 55-66.
8. P. Walters, *An introduction to ergodic theory*, Springer, New York, 1981.



Department of Mathematical Sciences, University of Memphis, Memphis, TN 38152 USA

*E-mail address*: quasa@msci.memphis.edu, trowp@msci.memphis.edu