Universal Scanning of Mixing Random Fields and the Performance of the Peano-Hilbert Scan

Asaf Cohen
Department of Electrical Engineering
Technion - I.I.T.
Haifa 32000, ISRAEL
Email: soofsoof@tx.technion.ac.il

Neri Merhav
Department of Electrical Engineering
Technion - I.I.T.
Haifa 32000, ISRAEL
Email: merhav@ee.technion.ac.il

Tsachy Weissman
Department of Electrical Engineering
Stanford University
Stanford 94305-9510, USA
Email: tsachy@stanford.edu

***Abstract*— We investigate the problem of scanning and prediction ("scandiction", for short) of multidimensional data arrays. This problem arises in several aspects of image and video processing, such as predictive coding, where an image is compressed by coding the prediction error sequence resulting from scandicting it.**

Specifically, given a strongly mixing random field, we show that there exists a scandiction scheme which is independent of the field's distribution, yet almost surely asymptotically achieves the same performance as if this distribution was known. We then discuss the scenario where the Peano-Hilbert scanning order is used, accompanied by an optimal predictor, and derive a bound on the excess loss compared to optimal finite state scandiction, which is valid for any individual image and any bounded loss function.

I. INTRODUCTION

Consider the problem of sequentially scanning and predicting a multidimensional data array, while minimizing a given loss function. Particularly, at each time instant t, $0 \leq t \leq |B|-1$, where $|B|$ is the number of sites ("pixels") in the data array, the scandictor chooses a site to be visited, denoted by Ψ_t, and gives a prediction, F_t, for the value at that site. Both Ψ_t and F_t may depend of the previously observed values - the values at sites Ψ_0 to Ψ_{t-1}. It then observes the true value, x_{Ψ_t}, suffers a loss $l(x_{\Psi_t}, F_t)$, and so on. The goal is to minimize the cumulative loss after scandicting the entire data array.

Important applications of this problem can be found in image and video processing, such as predictive coding or denoising. In these applications, one wishes to code the prediction error (predictive coding), or use the predicted value as an estimate to the noiseless pixel value (denoising). It is clear that different scanning patterns of the image may result in different prediction errors, thus, it is natural to ask what is the optimal scanning strategy, whether there exists a universal scan, which does not depend on the image model yet performs as if this model was known in advance, and what is the loss when non-optimal strategies are use.

In [1], Lempel and Ziv showed that the Peano-Hilbert scan is optimal for compression of individual images. In [2], Merhav and Weissman formally defined the "scandictability" of a random field, namely, the best achievable performance, and discussed particular cases where this value can be computed and the optimal scanning order can be identified. A more comprehensive survey can be found in [3].

Although the problem is strongly related to its one-dimensional analogue, the numerous scanning possibilities result in a substantially richer and more challenging problem, and many of the important results in prediction and filtering of one-dimensional arrays have only partial solutions in the multi-dimensional case. In this paper, we prove the following two results. In Section III, we show that, under certain conditions on the random field and loss function, there exists a universal algorithm which does not depend on the probability measure of the data array, Q, yet asymptotically Q-a.s. achieves the scandictability of the source. In Section IV, we consider the scenario where the data array to be scandicted is an *individual image*, and bound the excess loss when the Peano-Hilbert scan is used, compared to any other finite state scan, establishing it as an advantageous scan in the individual image scenario.

II. PROBLEM FORMULATION

Let A denote a finite alphabet set. Let $\Omega = A^{\mathbb{Z}^d}$ denote the space of all possible data arrays in $\mathbb{Z}^d$. Although the results in this paper are applicable to any $d \geq 1$, for simplicity, we assume from now on that $d = 2$. The extension to $d \geq 2$ is straightforward. A probability measure Q on Ω is stationary if it is invariant under translations τ_i, where for each $x \in \Omega$ and $i, j \in \mathbb{Z}^2$, $\tau_i(x)_j = x_{j+i}$ (namely, stationarity means shift invariance). Denote by $\mathcal{M}(\Omega)$ and $\mathcal{M}_S(\Omega)$ the spaces of all probability measures and stationary probability measures on Ω, respectively. Elements of $\mathcal{M}(\Omega)$, *random fields*, will be denoted by upper case letters while elements of Ω, *individual data arrays*, will be denoted by the corresponding lower case.

Let $\mathcal{V}$ denote the set of all finite subsets of $\mathbb{Z}^2$. For $V \in \mathcal{V}$, denote by X_V the restrictions of the data array X to V. For $i \in \mathbb{Z}^2$, X_i is the random variable corresponding to X at site i. Denote by V_n the square $\{0, \ldots, n-1\} \times \{0, \ldots, n-1\}$. Throughout, $\log(\cdot)$ will denote the natural logarithm, and entropies will be measured in nats.

Definition 1 ([2]): A *scandictor* for a finite set of sites $B \in \mathcal{V}$ is the following pair (Ψ, F):

- $\{\Psi_t\}_{t=1}^{|B|}$ is a sequence of measurable mappings, $\Psi_t : A^{t-1} \mapsto B$ determining the site to be visited at time t, with the property that

$$\Big\{\Psi_1, \Psi_2(x_{\Psi_1}), \Psi_3(x_{\Psi_1}, x_{\Psi_2})\ldots, \\ \Psi_{|B|}\big(x_{\Psi_1}, \ldots, x_{\Psi_{|B|-1}}\big)\Big\} = B, \quad \forall x \in A^B. \quad (1)$$

- $\{F_t\}_{t=1}^{|B|}$ is a sequence of measurable predictors, where for each t, $F_t : A^{t-1} \mapsto D$ determines the prediction for the site visited at time t based on the observations at previously visited sites, and D is a finite prediction alphabet.

We allow *randomized scandictors*, namely, scandictors such that $\{\Psi_t\}_{t=1}^{|B|}$ or $\{F_t\}_{t=1}^{|B|}$ can be chosen randomly from some set of possible functions. Scandictors for *infinite* data arrays are not considered in this paper. We will consider, though, the limit as the size of the array tends to infinity.

It is also beneficial to consider *finite-state scandictors*, similarly to [1]. In this case, Ψ_t and F_t are determined according the scandictor's *state*, namely, $\Psi_t = \Psi_{t-1} + d(s_{t-1})$ and $F_t = F(s_{t-1})$, where d is the displacement function. The next state is chosen according to the next state function g, $s_t = g(s_{t-1}, x_{\Psi_t})$.

Denote by $L_{(\Psi,F)}(x_{V_n})$ the cumulative loss of (Ψ, F) over x_{V_n}, that is

$$L_{(\Psi,F)}(x_{V_n}) = \sum_{t=1}^{|V_n|} l\left(x_{\Psi_t}, F_t(x_{\Psi_1}, \ldots, x_{\Psi_{t-1}})\right), \quad (2)$$

where $l : A \times D \to [0, \infty)$ is a given loss function. We assume that $l(\cdot, \cdot)$ is non-negative and bounded by $l_{max} < \infty$. The scandictability of a source $Q \in \mathcal{M}(\Omega)$ on $B \in \mathcal{V}$ is defined by

$$U(l, Q_B) = \inf_{(\Psi,F)\in\mathcal{S}(B)} E_{Q_B} \frac{1}{|B|} L_{(\Psi,F)}(X_B), \quad (3)$$

where Q_B is the marginal probability measure of X restricted to B and $\mathcal{S}(B)$ is the set of *all* possible scandictors for B. The scandictability of $Q \in \mathcal{M}(\Omega)$ is defined by

$$U(l, Q) = \lim_{n\to\infty} U(l, Q_{V_n}). \quad (4)$$

By [2, Theorem 1], the limit in (4) exists for any $Q \in \mathcal{M}_S(\Omega)$.

For $A, B \in \mathbb{Z}^2$, define

$$\alpha^Q(A, B) = \sup\{|Q(U \cap V) - Q(U)Q(V)|, \\ U \in \sigma(X_A), V \in \sigma(X_B)\}, \quad (5)$$

where $\sigma(X_B)$ is the smallest sigma algebra generated by X_B. Let $\alpha^Q_{a,b}(k)$ denote the strong mixing coefficient of the random field Q,

$$\alpha^Q_{a,b}(k) = \sup\{\alpha^Q(A, B), |A| \le a, |B| \le b, d(A, B) \ge k\}, \quad (6)$$

where d is a metric on $\mathbb{Z}^2$ and $d(A, B)$ is the distance between the closest points, i.e., $d(A, B) = \min_{i\in A, j\in B} d(i, j)$. A measure Q is strongly mixing if for all $a, b \in \mathbb{N} \cup \{\infty\}$, $\alpha^Q_{a,b}(k) \to 0$ as $k \to \infty$.

Denote by $\mathcal{I}_m$ the σ-algebra of m-invariant sets,

$$A \in \mathcal{I}_m \text{ iff } \tau_j(A) = A \quad \text{for all } j = i \cdot m, i \in \mathbb{Z}^2. \quad (7)$$

If $\mathcal{I}_m$ is the trivial σ-algebra, the measure Q is called m-block ergodic. It is not hard to show that if the measure Q is strongly mixing, then it is block-ergodic for any finite block size (i.e., totally ergodic).

III. Universal Scandiction

The following theorem is the main result of this paper.

Theorem 2: Let X be a stationary strongly mixing random field with a probability measure Q. Then, there exists a sequence of scandictors $\{(\Psi, F)_n\}$, independent of Q, where $(\Psi, F)_n$ is a scandictor for V_n and operates in blocks of size $m \times m$, $m < n$, for which

$$\liminf_{n\to\infty} \frac{1}{|V_n|} L_{(\Psi,F)_n}(X_{V_n}) \le U(l, Q) + \delta(m) \quad Q - a.s. \ (8)$$

for any such Q and some $\delta(m)$ such that $\delta(m) \to 0$ as $m \to \infty$. Thus, when $m \to \infty$, the performance of $\{(\Psi, F)_n\}$ equals the scandictability of the source, $Q - a.s.$

Before we prove Theorem 2, we present the scandiction algorithm and two helpful results from [3]. For an individual data array x_{V_n} and $m < n$, define $K \triangleq \lceil \frac{n}{m} \rceil - 1$ and divide x_{V_n} into K^2 blocks of size $m \times m$ and $2K + 1$ blocks of possibly smaller size. Denote by x^i, $0 \le i \le (K + 1)^2 - 1$ the i'th block under some fixed scanning order of the blocks. Since we will later see that this scanning order is irrelevant in this case, assume from now on that it is a raster scan from the upper left corner. Let $\mathcal{F}_m$ be a set of scandictors for V_m and denote by $L_{j,i}$ the cumulative loss of $(\Psi, F)_j \in \mathcal{F}_m$ after scanning i blocks, where $(\Psi, F)_j$ is *restarted* after each block, namely, it scans each block separately and independently of the other blocks. Note that $L_{j,i} = \sum_{l=0}^{i-1} L_j(x^l)$ and that for $i = 0$, $L_{j,i} = 0$ for all j. For the $2K + 1$ possibly smaller (and not square) blocks the loss may be l_{max} throughout. For $\eta > 0$, and any i and j, define

$$P_i\left(j|\{L_{j,i}\}_{j=1}^{\lambda}\right) = \frac{e^{-\eta L_{j,i}}}{\sum_{j=1}^{\lambda} e^{-\eta L_{j,i}}}, \quad (9)$$

where $\lambda = |\mathcal{F}_m|$. We offer the following algorithm for a block-wise scan of the data array x. For each $0 \le i \le (K + 1)^2 - 1$, after scanning i blocks of data, the algorithm computes $P_i\left(j|\{L_{j,i}\}_{j=1}^{\lambda}\right)$ for each j. It then randomly selects a scandictor according to this distribution, independently of its previous selections, and uses this scandictor as its output for the $(i+1)$-st block. In [3], the following two results are proved.

Proposition 3: Let $L_{alg}(x_{V_n})$ be the cumulative loss of the proposed algorithm on x_{V_n}, and denote by $\bar{L}_{alg}(x_{V_n})$ its expected value, where the expectation is with respect to the randomized scandictor selection of the algorithm. Let L_{min} denote the cumulative loss of the best scandictor in $\mathcal{F}_m$, operating block-wise on x_{V_n}. Assume $|\mathcal{F}_m| = \lambda$, then

$$\bar{L}_{alg}(x_{V_n}) - L_{min}(x_{V_n}) \le m(n + m)\sqrt{\log \lambda}\frac{l_{max}}{\sqrt{2}}. \quad (10)$$

Proposition 4: Assume $m = o\left(n^{1/3}\right)$. Then, as $n \to \infty$, $L_{alg}(x_{V_n})$ converges to $L_{min}(x_{V_n})$ for any x_{V_n}, with probability 1 with respect to the randomization in the algorithm.

Proof: [Theorem 2] Fix n and a block size $m < n$. Take $\mathcal{F}_m = \mathcal{S}(m)$, i.e., the set of all possible scandictors for V_m. By Proposition 3, for each x_{V_n}, we have,

$$
\begin{aligned}
&\frac{1}{|V_n|}\bar{L}_{alg}(x_{V_n}) \\
&\quad \le \frac{1}{|V_n|}L_{min}(x_{V_n}) + \frac{m(n+m)}{|V_n|}\sqrt{\log\lambda}\frac{l_{max}}{\sqrt{2}} \\
&\quad = \frac{1}{|V_n|}\min_{(\Psi,F)_j\in\mathcal{S}(V_m)}\sum_{i=1}^{(K+1)^2} L_j(x^i) \\
&\qquad + \frac{m(n+m)}{|V_n|}\sqrt{\log\lambda}\frac{l_{max}}{\sqrt{2}} \\
&\quad \le \frac{1}{|V_m|}\min_{(\Psi,F)_j\in\mathcal{S}(V_m)}\frac{1}{K^2}\sum_{i=1}^{K^2} L_j(x^i) \\
&\qquad + 2\frac{m}{n}l_{max} + \frac{m(n+m)}{|V_n|}\sqrt{\log\lambda}\frac{l_{max}}{\sqrt{2}}. \qquad (11)
\end{aligned}
$$

The size of $\mathcal{F}_m$ can be bounded using a simple counting argument [3],

$$\lambda \le (m^2!)^{|A|^{m^2}}|D|^{m^2|A|^{m^2-1}}, \qquad (12)$$

thus,

$$
\begin{aligned}
&\liminf_{n\to\infty}\frac{1}{|V_n|}L_{alg}(x_{V_n}) = \liminf_{n\to\infty}\frac{1}{|V_n|}\bar{L}_{alg}(x_{V_n}) \\
&\quad \le \frac{1}{|V_m|}\liminf_{n\to\infty}\min_{(\Psi,F)_j\in\mathcal{S}(V_m)}\frac{1}{K^2}\sum_{i=1}^{K^2} L_j(x^i) \\
&\quad \le \frac{1}{|V_m|}\min_{(\Psi,F)_j\in\mathcal{S}(V_m)}\liminf_{n\to\infty}\frac{1}{K^2}\sum_{i=1}^{K^2} L_j(x^i). \qquad (13)
\end{aligned}
$$

where the first equality is due to Proposition 4. Since $K \to \infty$ as $n \to \infty$, by the block ergodicity of Q (see [4, Theorem 6.1′]) and the fact that for finite m and each $(\Psi,F)_j \in \mathcal{S}(V_m)$, $L_j(X)$ is a bounded function, it follows that

$$\liminf_{n\to\infty}\frac{1}{K^2}\sum_{i=1}^{K^2} L_j(x^i) = E_{Q_{V_m}}L_j(X^0) \quad Q-a.s. \qquad (14)$$

Finally, since $U(l,Q)$ exists, there exists $\delta(m)$ such that $\delta(m) \to 0$ as $m \to \infty$ and

$$\min_{(\Psi,F)_j\in\mathcal{S}(V_m)} E_{Q_{V_m}}L_j(X^0) \le U(l,Q) + \delta(m). \qquad (15)$$

■

IV. Individual Images and the Peano-Hilbert Scan

Consider now the scenario of predicting the next outcome of a binary *individual* source, with $D = [0,1]$ as the prediction space. Hence, $l : \{0,1\} \times [0,1] \to \mathbb{R}$ is the loss function. Let ϕ_l denote the Bayes envelope associated with l, i.e.,

$$\phi_l(p) = \min_{q\in[0,1]}[(1-p)l(0,q) + pl(1,q)]. \qquad (16)$$

We further define

$$\epsilon_l = \min_{\alpha,\beta}\max_{0\le p\le 1}|\alpha h_b(p) + \beta - \phi_l(p)|, \qquad (17)$$

where $h_b(p)$ is the binary entropy function. Thus, ϵ_l is the error in approximating $\phi_l(p)$ by the best affine function of $h_b(p)$. For example, when l is the Hamming loss function, denoted by l_H, we have $\epsilon_{l_H} = 0.08$ and when l is the squared error, denoted by l_s, $\epsilon_{l_s} = 0.0137$. For the log loss, however, $\epsilon_{l_{log}} = 0$.

Let Ψ_B be a scanner for the data array x_B. Let $x_1^{|B|}$ be the sequence resulting from scanning x_B with Ψ_B. Fix $k < |B|$ and for any $s \in \{0,1\}^{k+1}$ define the empirical distribution of order $k+1$ as

$$\hat{P}^{k+1}_{\Psi_B}(s) = \frac{1}{|B|-k}\left|\left\{k < i \le |B| : x^i_{i-k} = s\right\}\right|. \qquad (18)$$

The distributions of lower orders, and the conditional distribution are derived from $\hat{P}^{k+1}_{\Psi_B}(s)$, i.e., for $s' \in \{0,1\}^k$ and $x \in \{0,1\}$ we define

$$\hat{P}^{k+1}_{\Psi_B}(s') = \hat{P}^{k+1}_{\Psi_B}([s',0]) + \hat{P}^{k+1}_{\Psi_B}([s',1]) \qquad (19)$$

and

$$\hat{P}^{k+1}_{\Psi_B}(x|s') = \frac{\hat{P}^{k+1}_{\Psi_B}([s',x])}{\hat{P}^{k+1}_{\Psi_B}(s')}, \qquad (20)$$

where $0/0$ is defined as $1/2$ and $[\cdot,\cdot]$ denotes string concatenation. Let $\hat{H}^{k+1}_{\Psi_B}(X|X^k)$ be the empirical conditional distribution of order k, i.e.,

$$
\begin{aligned}
&\hat{H}^{k+1}_{\Psi_B}(X|X^k) \\
&\quad = -\sum_{s\in\{0,1\}^k}\hat{P}^{k+1}_{\Psi_B}(s)\sum_{x\in\{0,1\}}\hat{P}^{k+1}_{\Psi_B}(x|s)\log\hat{P}^{k+1}_{\Psi_B}(x|s). \qquad (21)
\end{aligned}
$$

Denote by $F^{k,opt}$ the optimal k-th order Markov predictor, in the sense that it minimizes the expected loss with respect to $\hat{P}^{k+1}_{\Psi_B}(\cdot|\cdot)$ and $x_1^{|B|}$. For $\Psi = \{\Psi_n\}$, where Ψ_n is a scan for V_n, and an infinite individual image x, define

$$L^k_\Psi(x) = \limsup_{n\to\infty}\frac{1}{|V_n|}L_{(\Psi_n,F^{k,opt})}(x_{V_n}) \qquad (22)$$

and

$$L_\Psi(x) = \lim_{k\to\infty}L^k_\Psi(x). \qquad (23)$$

Theorem 5 relates the asymptotic cumulative loss of any sequence of finite state scans Ψ to that resulting from the Peano-Hilbert sequence of scans, establishing the Peano-Hilbert sequence as an advantageous scanning order for any loss function.

Theorem 5: Let x be any individual image. Let PH denote the Peano-Hilbert sequence of scans. Then, for any sequence of finite state scans Ψ and any loss function $l : \{0,1\}\times[0,1] \to \mathbb{R}$,

$$L_{PH}(x) \le L_\Psi(x) + 2\epsilon_l. \qquad (24)$$

The following proposition relates the empirical conditional distribution to the cumulative loss under any scan Ψ_B, and will be useful in the proof of Theorem 5.

Proposition 6: Let x_B be any data array. Then,

$$\left|\alpha_l \hat{H}^{k+1}_{\Psi_B}(X|X^k) + \beta_l - \frac{1}{|B|} L_{(\Psi_B, F^{k,opt})}(x_B)\right| \leq \epsilon_l + \frac{k l_{max}}{|B|}, \tag{25}$$

where α_l and β_l are the achievers of the minimum in (17).

Proof:

$$\begin{aligned}
&\left|\alpha_l \hat{H}^{k+1}_{\Psi_B}(X|X^k) + \beta_l - \frac{1}{|B|} L_{(\Psi_B, F^{k,opt})}(x_B)\right| \\
&\leq \left|\alpha_l \hat{H}^{k+1}_{\Psi_B}(X|X^k) + \beta_l \right.\\
&\qquad \left. - \frac{1}{|B|-k} \sum_{t=k+1}^{|B|} l(x_t, F^{k,opt}(x_{t-k}^{t-1}))\right| + \frac{k l_{max}}{|B|} \\
&\leq \left|\alpha_l \hat{H}^{k+1}_{\Psi_B}(X|X^k) + \beta_l - \right.\\
&\qquad \left. \sum_{s \in \{0,1\}^{k+1}} \hat{P}^{k+1}_{\Psi_B}(s) l(s_{k+1}, F^{k,opt}(s_1^k))\right| + \frac{k l_{max}}{|B|} \\
&\leq \left| \sum_{s' \in \{0,1\}^k} \hat{P}^{k+1}_{\Psi_B}(s') \right.\\
&\qquad \left. \left(\alpha_l h_b(\hat{P}^{k+1}_{\Psi_B}(\cdot|s')) + \beta_l - \phi_l(\hat{P}^{k+1}_{\Psi_B}(\cdot|s'))\right)\right| + \frac{k l_{max}}{|B|} \\
&\leq \sum_{s' \in \{0,1\}^k} \hat{P}^{k+1}_{\Psi_B}(s') \max_p |\alpha_l h_b(p) + \beta_l - \phi_l(p)| + \frac{k l_{max}}{|B|} \\
&= \epsilon_l + \frac{k l_{max}}{|B|}.
\end{aligned} \tag{26}$$

■

Define the asymptotic k-th order empirical conditional entropy under $\{\Psi_n\}$ as

$$\hat{H}^{k+1}_{\Psi}(x) = \limsup_{n\to\infty} \hat{H}^{k+1}_{\Psi_n}(X|X^k) \tag{27}$$

and further define

$$\hat{H}_{\Psi}(x) = \lim_{k\to\infty} \hat{H}^{k+1}_{\Psi}(x). \tag{28}$$

The existence of $\hat{H}_\Psi(x)$ is established later in the proof of Theorem 5, where it is also shown that this limit equals $\lim_{k\to\infty} \limsup_{n\to\infty} \frac{1}{k} \hat{H}^k_{\Psi_n}(X^k)$. By [5, Theorem 3], the later limit is no other than the asymptotic finite state compressibility of x under the sequence of scans Ψ, namely,

$$\begin{aligned}
\lim_{k\to\infty} \limsup_{n\to\infty} \frac{1}{k} \hat{H}^k_{\Psi_n}(X^k) &= \rho(\Psi(x)) \\
&= \lim_{s\to\infty} \limsup_{n\to\infty} \rho_{E(s)}(\Psi_n(x_{V_n})),
\end{aligned} \tag{29}$$

where $\rho_{E(s)}(x_1^n)$ is the minimum compression ratio for x_1^n over the class of all finite state encoders with at most s states [5, eq. (1)-(4)]. We may now prove Theorem 5.

Proof: [Theorem 5] By Proposition 6,

$$\begin{aligned}
&\left|\alpha_l \hat{H}^{k+1}_{\Psi_n}(X|X^k) + \beta_l - \frac{1}{|V_n|} L_{(\Psi_n, F^{k,opt})}(x_{V_n})\right| \\
&\qquad \leq \epsilon_l + \frac{k l_{max}}{|V_n|}.
\end{aligned} \tag{30}$$

Taking the limsup as $n \to \infty$ yields

$$\left|\alpha_l \limsup_{n\to\infty} \hat{H}^{k+1}_{\Psi_n}(X|X^k) + \beta_l - L^k_\Psi(x)\right| \leq \epsilon_l. \tag{31}$$

For a stationary source, it is well known (e.g., [6, Theorem 4.2.1]) that $\lim_{k\to\infty} H(X_k|X_1^{k-1})$ exists and in fact

$$\lim_{k\to\infty} H(X_k|X_1^{k-1}) = \lim_{k\to\infty} \frac{1}{k} H(X_1^k). \tag{32}$$

To this end, we show that the same holds for empirical entropies. We start by showing that $\limsup_{n\to\infty} \hat{H}^{k+1}_{\Psi_n}(X|X^k)$ is a decreasing sequence in k. Since conditioning reduces the entropy, it is clear that $\hat{H}^{k+1}_{\Psi_n}(X|X^k) \leq \hat{H}^{k+1}_{\Psi_n}(X|X^{k-1})$, where both are calculated using $\hat{P}^{k+1}_{\Psi_n}(\cdot)$. However, the above may not be true when $\hat{H}^{k+1}_{\Psi_n}(X|X^{k-1})$ is replaced by $\hat{H}^k_{\Psi_n}(X|X^{k-1})$, as the later is calculated using $\hat{P}^k_{\Psi_n}(\cdot)$. Nevertheless, using a simple counting argument, it is not too hard to show that for every k, $0 < j \leq k$ and $s \in \{0,1\}^i$, where $0 < i \leq j$, we have

$$\hat{P}^{k+1}_{\Psi_n}(s) - \frac{k+1-j}{|V_n|-k} \leq \hat{P}^j_{\Psi_n}(s) \leq \hat{P}^{k+1}_{\Psi_n}(s) + \frac{k+1-j}{|V_n|-k}. \tag{33}$$

Thus, by the continuity of the entropy function, we have

$$\begin{aligned}
\limsup_{n\to\infty} \hat{H}^{k+1}_{\Psi_n}(X|X^k) &\leq \limsup_{n\to\infty} \hat{H}^{k+1}_{\Psi_n}(X|X^{k-1}) \\
&= \limsup_{n\to\infty} \hat{H}^k_{\Psi_n}(X|X^{k-1}),
\end{aligned} \tag{34}$$

hence $\limsup_{n\to\infty} \hat{H}^k_{\Psi_n}(X|X^{k-1})$ is decreasing in k. Since it is a non negative sequence, $\hat{H}_\Psi(x)$ as defined in (28) exists and we have

$$\left|\alpha_l \hat{H}_\Psi(x) + \beta_l - L_\Psi(x)\right| \leq \epsilon_l. \tag{35}$$

We now show that indeed $\hat{H}_\Psi(x)$ equals $\rho(\Psi(x))$ for every sequence of finite state scans Ψ, hence when Ψ is a sequence of finite state scans the results of [1] can be applied. The method is similar to that in [6, Theorem 4.2.1]), with an adequate handling of empirical entropies. By (33),

$$\begin{aligned}
\limsup_{n\to\infty} \frac{1}{k} \hat{H}^k_{\Psi_n}(X^k) &= \limsup_{n\to\infty} \frac{1}{k} \sum_{i=1}^k \hat{H}^k_{\Psi_n}(X_i|X_1^{i-1}) \\
&= \limsup_{n\to\infty} \frac{1}{k} \sum_{i=1}^k \hat{H}^i_{\Psi_n}(X_i|X_1^{i-1}).
\end{aligned} \tag{36}$$

But the sequence $\limsup_{n\to\infty} \hat{H}^i_{\Psi_n}(X_i|X_1^{i-1})$ converges to $\hat{H}_\Psi(x)$ as $i \to \infty$, thus its Cesaro mean converges to the same limit and we have

$$\begin{aligned}
\hat{H}_\Psi(x) &= \lim_{k\to\infty} \limsup_{n\to\infty} \frac{1}{k} \hat{H}^k_{\Psi_n}(X^k) \\
&= \rho(\Psi(x)).
\end{aligned} \tag{37}$$

Consider now the Peano-Hilbert sequence of finite state scans, denoted by PH. Let $\rho(x)$ denote the (finite state) compressibility of x as defined in [1, eq. (4)]. For any other sequence of finite state scans $\tilde{\Psi}$ we have

$$\begin{aligned}\hat{H}_{PH}(x) &\leq \rho(x)\\ &\leq \hat{H}_{\tilde{\Psi}}(x),\end{aligned} \tag{38}$$

where the first inequality is by [1, eq. (9) and (16)] and the second is straightforward from the definition of $\rho(x)$. Finally,

$$\begin{aligned}L_{PH}(x) &\overset{(a)}{\leq} \epsilon_l + \beta_l + \alpha_l \hat{H}_{PH}(x)\\ &\leq \epsilon_l + \beta_l + \alpha_l \hat{H}_{\tilde{\Psi}}(x)\\ &\overset{(b)}{\leq} 2\epsilon_l + L_{\tilde{\Psi}}(x),\end{aligned} \tag{39}$$

where (a) and (b) result from the application of (35) to the sequences PH and $\tilde{\Psi}$ respectively. ■

Finally, we also have the following corollary.

Corollary 7: Let Ψ_1 and Ψ_2 be *any* two sequences of scans such that $\hat{H}_{\Psi_1}(x) = \hat{H}_{\Psi_2}(x)$ (in particular, if both Ψ_1 and Ψ_2 are finite state sequences of scans they result in the same finite state compressibility). Then,

$$|L_{\Psi_1}(x) - L_{\Psi_2}(x)| \leq 2\epsilon_l \tag{40}$$

for any bounded loss function $l : \{0,1\} \times [0,1] \to \mathbb{R}$.

The proof of Corollary 7 is straightforward, using (35) for both Ψ_1 and Ψ_2 and the triangle inequality.

For a given sequence of scans Ψ, the set of scanning sequences Ψ' satisfying $\hat{H}_{\Psi}(x) = \hat{H}_{\Psi'}(x)$ is larger than one might initially think. For example, a close look at the definition of finite state compressibility given in [5] shows that the finite state encoders defined therein allow *limited scanning schemes*, as an encoder might read a large data set before its output for that data set is given. Thus, a legitimate finite state encoder in the sense of [5] may reorder the data in a block (of bounded length, as the number of states is bounded) before actually encoding it. Consequently, for any individual *sequence* x one can define several permutations having the same finite state compressibility. In the multidimensional scenario this sums up to saying that for each scanning sequence Ψ there exist several different scanning sequences Ψ' for which $H_{\Psi}(x) = \hat{H}_{\Psi'}(x)$.

Case Study: Hamming Loss. The bound in Theorem 5 is valid for any bounded loss function $l : \{0,1\} \times [0,1] \to \mathbb{R}$. When l is the Hamming loss, the resulting bound is

$$L_{PH}^{Hamming}(x) \leq L_{\Psi}^{Hamming}(x) + 0.16, \tag{41}$$

for any other finite state sequence of scans $\{\Psi\}_n$.

In [7], Feder, Merhav and Gutman proved that for any next-state function $g \in G_s$, where G_s is the set of all possible next state functions with s states, and for any sequence x_1^n,

$$\begin{aligned}\mu(g, x_1^n) &\leq \frac{1}{2}\rho(g, x_1^n),\\ \mu(g, x_1^n) &\geq h_b^{-1}\left(\rho(g, x_1^n)\right),\end{aligned} \tag{42}$$

where $\mu(g, \cdot)$ ($\rho(g, \cdot)$) is the best possible prediction (compression) performance when the next state function is g. Using the results of [1], it is possible to show [3] that

$$\begin{aligned}&\min_{g\in G_s} \mu(g, \Psi_{PH}(x_{V_n})) - \min_{g\in G_s} \mu(g, \Psi(x_{V_n}))\\ &\quad\leq \frac{1}{2}\min_{g\in G_s} \rho(g, \Psi_{PH}(x_{V_n}))\\ &\qquad - h^{-1}\left(\min_{g\in G_s} \rho(g, \Psi_{PH}(x_{V_n})) - \epsilon_{n,s}\right),\end{aligned} \tag{43}$$

for any finite-state scan Ψ_n, where $\epsilon_{n,s}$ satisfies $\lim_{s\to\infty} \limsup_{n\to\infty} \epsilon_{n,s} = 0$. Taking the limits $\limsup_{n\to\infty}$ and then $s \to \infty$ implies that in using the Peano-Hilbert scan for scanning and prediction under Hamming loss one losses no more than $\frac{1}{2}\rho(x) - h^{-1}(\rho(x))$ with respect to *any finite-state scan* Ψ, where $\rho(x)$ is the image's FS compressibility. The maximum possible loss is 0.16, similar to the bound given in Theorem 5, yet this value is achieved only when the image's FS compressibility is around 0.75 bits/symbol. For images which are highly compressible, for example, when $\rho < 0.1$, the resulting excess loss is smaller than 0.04.

V. Conclusion

In this paper, we considered the scenario of scanning and predicting a stationary strongly mixing random field, and showed that there exists a scandiction scheme which is independent of the field's distribution, yet almost surely asymptotically achieves the same performance as if this distribution was known. Namely, the existence of an algorithm which is universal in the almost sure sense was established.

We then discussed the scenario where the data array is an individual image, and showed that if the Peano-Hilbert scanning order is used, accompanied by an optimal predictor, the excess loss compared to optimal finite state scandiction is bounded from above by a small constant, which depends only on the loss function. This result marks the Peano-Hilbert scan as advantageous under any bounded loss function, extending the results of [1].

Future work may aim at extending the results for the individual image scenario, seeking universal scandictors in this case as well, for as large scandictor sets as possible.

References

[1] A. Lempel and J. Ziv, "Compression of two-dimensional data," *IEEE Trans. Inform. Theory*, vol. IT-32, no. 1, pp. 2–8, January 1986.

[2] N. Merhav and T. Weissman, "Scanning and prediction in multidimensional data arrays," *IEEE Trans. Inform. Theory*, vol. 49, no. 1, pp. 65–82, January 2003.

[3] A. Cohen, N. Merhav, and T. Weissman, "Universal scanning and sequential decision making for multi-dimensional data, part I: the noiseless case," *In preperation*, 2006.

[4] A. A. Tempelman, "Ergodic theorems for general dynamical systems," *Trans. Moscow Math. Soc.*, pp. 94–132, 1972.

[5] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 530–536, September 1978.

[6] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[7] M. Feder, N. Merhav, and M. Gutman, "Universal prediction of individual sequences," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1258–1270, July 1992.