Scanning and Sequential Decision Making for Multi-Dimensional Data - Part I: the Noiseless Case[*]

Asaf Cohen[†], Neri Merhav[†] and Tsachy Weissman[‡]

May 7, 2007

Abstract

We investigate the problem of scanning and prediction ("scandiction", for short) of multi-dimensional data arrays. This problem arises in several aspects of image and video processing, such as predictive coding, for example, where an image is compressed by coding the error sequence resulting from scandicting it. Thus, it is natural to ask what is the optimal method to scan and predict a given image, what is the resulting minimum prediction loss, and whether there exist specific scandiction schemes which are universal in some sense.

Specifically, we investigate the following problems: First, modeling the data array as a random field, we wish to examine whether there exists a scandiction scheme which is independent of the field's distribution, yet asymptotically achieves the same performance as if this distribution was known. This question is answered in the affirmative for the set of all spatially stationary random fields and under mild conditions on the loss function. We then discuss the scenario where a non-optimal scanning order is used, yet accompanied by an optimal predictor, and derive bounds on the excess loss compared to optimal scanning and prediction.

This paper is the first part of a two-part paper on sequential decision making for multi-dimensional data. It deals with clean, noiseless data arrays. The second part deals with noisy

1

data arrays, namely, with the case where the decision maker observes only a noisy version of the data, yet it is judged with respect to the original, clean data.

Index Terms-Universal scanning, Scandiction, Sequential decision making, Multi-dimensional data, Random Field, Individual image.

1   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, $1 \leq t \leq |B|$, 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_1 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.

The problem of sequentially predicting the next outcome of a one-dimensional sequence (or any data array with a fixed, predefined, order), x_t, based on the previously observed outcomes, $x_1, x_2, \ldots, x_{t-1}$, is well studied. The problem of prediction in multidimensional data arrays (or when reordering of the data is allowed), however, has received far less attention. Apart from the on-line strategies for the sequential prediction of the data, the fundamental problem of *scanning* it should be considered. We refer to the former problem as the *prediction* problem, where no reordering of the data is allowed, and to the latter as the *scandiction* problem.

The scandiction problem mainly arises in image compression, where various methods of predictive coding are used (e.g., [1]). In this case, the encoder may be given the freedom to choose the actual *path* over which it traverses the image, and thus it is natural to ask which path is optimal in the sense of minimal cumulative prediction loss (which may result in maximal compression). The scanning problem also arises in other areas of image processing, such as one-dimensional wavelet [2] or median [3] processing of images, where one seeks a space-filling curve which facilitates the one-dimensional signal processing of multidimensional data, digital halftoning [4], where a space filling curve is sought in order to minimize the effect of false contours, and pattern recognition [5], where it is shown that under certain conditions, the Bayes risk as well as the optimal decision rule are unchanged if instead of the original multidimensional classification problem one transforms the data using a measure-preserving space-filling curve and solves a simpler one-dimensional problem. More applications

can be found in multidimensional data query [6], [7] and indexing [8], where multidimensional data is stored on a one-dimensional storage device, hence a locality-preserving space-filling curve is sought in order to minimize the number of continuous read operations required to access a multidimensional object, and rendering of three-dimensional graphics [9], [10], where a rendering sequence which minimizes the number of cache misses is required.

The above applications have already been considered in the literature, and the benefits of not-trivial scanning orders have been proved (see [11], or [12] and [13] which we discuss later). Yet, the scandiction problem may have applications that go beyond image scanning alone. For example, consider a robot processing various types of products in a warehouse. The robot identifies a product using a bar-code or an RFID, and processes it accordingly. If the robot could predict the next product to be processed, and prepare for that operation while commuting to the product (e.g., prepare an appropriate writing-head and a message to be written), then the total processing time could be smaller compared to preparing for the operation only after identifying the product. Since different sites in the warehouse my be correlated in terms of the various products stored in them, it is natural to ask what is the optimal path to scan the entire warehouse in order to achieve minimum prediction error and thus minimal processing time.

In [14], a specific scanning method was suggested by Lempel and Ziv for the lossless compression of multidimensional data. It was shown that the application of the incremental parsing algorithm of [15] on the one dimensional sequence resulting from the *Peano-Hilbert* scan yields a universal compression algorithm with respect to all finite-state *scanning and encoding machines*. These results where later extended in [16] to the probabilistic setting, where it was shown that this algorithm is also universal for any stationary Markov random field [17]. Using the universal quantization algorithm of [18], the existence of a universal rate-distortion encoder was also established. Additional results regarding lossy compression of random fields (via pattern matching) were given in [19] and [20]. For example, in [20], Kontoyiannis considered a lossy encoder which encodes the random field by searching for a D-closest match in a given database, and then describing the position in the database.

While the algorithm suggested in [14] is asymptotically optimal, it may not be the optimal compression algorithm for real life images of sizes such as 256×256 or 512×512. In [12], Memon *et. al.* considered image compression with a codebook of block scans. Therein, the authors sought a scan which minimizes the zero order entropy of the *difference image*, namely, that of the sequence of differences between each pixel and its preceding pixel along the scan.

3

Since this problem is computationally expensive, the authors aimed for a suboptimal scan which minimizes the sum of absolute differences. This scan can be seen as a minimum spanning tree of a graph whose vertices are the pixels in the image and whose edges weights represent the differences (in gray levels) between each pixel and its adjacent neighbors. Although the optimal spanning tree can be computed in linear time, encoding it may yield a total bit rate which is higher than that achieved with an ordinary raster scan. Thus, the authors suggested to use a *codebook of scans*, and encode each block in the image using the best scan in the codebook, in the sense of minimizing the total loss.

Lossless compression of images was also discussed by Dafner *et. al.* in [13]. In this work, a context-based scan which minimizes the number of edge crossing in the image was presented. Similarly to [12], a graph was defined and the optimal scan was represented through a minimal spanning tree. Due to the bit rate required to encode the scan itself the results fall short behind [14] for two-dimensional data, yet they are favorable when compared to applying the algorithm in [14] *to each frame* in a three-dimensional data (assuming the context-based scans for each frame in the algorithm of [13] are similar).

Note that although the criterion chosen by Memon *et. al.* in [12], or by Dafner *et. al.* in [13], which is to minimize the sum of cumulative (first order) prediction errors or edge crossings, is similar to the criterion defined in this work, there are two important differences. First, the weights of the edges of the graph should be computed before the computation of the optimal (or suboptimal) scan begins, namely, the algorithm is not sequential in the sense of scanning and prediction in one pass. Second, the weights of the edges can only represent prediction errors of first order predictors (i.e., context of length one), since the prediction error for longer context depends on the scan itself - which has not been computed yet. In the context of lossless image coding it is also important to mention the work of Memon *et. al.* in [21], where common scanning techniques (such as raster scan, Peano-Hilbert and random scan) were compared in terms of minimal cumulative conditional entropy given a *finite* context (note that for unlimited context the cumulative conditional entropy does not depend on the scanning order, as will be elaborated on later). The image model was assumed to be an isotropic Gaussian random filed. Surprisingly, the results of [21] show that context-based compression techniques based on limited context may not gain by using Hilbert scan over raster scan. Note that under a different criterion, cumulative *squared* prediction error, the raster scan is indeed optimal for Gaussian fields, as it was shown later in [22], which we discuss next.

The results of [14] and [16] considered a specific, data independent scan of the data set.

Furthermore, even in the works of Memon *et. al.* [12] or Dafner *et. al.* [13], where data dependent scanning was considered, only limited prediction methods (mainly, first order predictors) were discussed, and the criterion used was minimal total bit rate of the encoded image. However, for a general predictor, loss function and random field (or individual image), it is not clear what is the optimal scan. This more general scenario was discussed in [22], where Merhav and Weissman formally defined the notion of a *scandictor*, a scheme for both scanning and prediction, as well as that of *scandictability*, the best expected performance on a data array. The main result in [22] is the fact that if a stochastic field can be represented autoregressively (under a specific scan Ψ) with a maximum-entropy innovation process, then it is optimally scandicted in the way it was created (i.e., by the specific scan Ψ and its corresponding optimal predictor).

While defining the yardstick for analyzing the scandiction problem, the work in [22] leaves many open challenges. As the topic of prediction is rich and includes elegant solutions to various problems, seeking analogous results in the scandiction scenario offers plentiful research objectives.

In Section 3, we consider the case where one strives to compete with a finite set $\mathcal{F}$ of scandictors. Specifically, assume that there exists a probability measure Q which governs the data array. Of course, given the probability measure Q and the scandictor set, one can compute the optimal scandictor in the set (in some sense which will be defined later). However, we are interested in a universal scandictor, which scans the data independently of Q, and yet achieves essentially the same performance as the best scandictor in $\mathcal{F}$ (see [23] for a complete survey of universal prediction). The reasoning behind the actual choice of the scandictor set $\mathcal{F}$ is similar to that common in the filtering and prediction literature (e.g., [24] and [25]). On the one hand, it should be large enough to cover a wide variety of random fields, in the sense that for each random field in the set, at least one scandictor is sufficiently close to the optimal scandictor corresponding to that random field. On the other hand, it should be small enough to compete with, at an acceptable cost of redundancy.

At first sight, in order to compete successfully with a finite set of scandictors, i.e., construct a universal scandictor, one may try to use known algorithms for learning with expert advice, e.g., the *exponential weighting* algorithm suggested in [26] or the work which followed it. In this algorithm, each expert is assigned a weight according to its past performance. By decreasing the weight of poorly performing experts, hence preferring the ones proved to perform well thus far, one is able to compete with the best expert, having neither any *a priori* knowledge

on the input sequence nor which expert will perform the best. However, in the scandiction problem, as each of the experts may use a different scanning strategy, at a given point in time each scanner might be at a different site, with different sites as its past. Thus, it is not at all guaranteed that one can alternate from one expert to the other. The problem is even more involved when the data is an individual image, as no statistical properties of the data can be used to facilitate the design or analysis of an algorithm. In fact, the first result in Section 3 is a negative one, stating that indeed in the individual image scenario (or under expected minimum loss in the stochastic scenario), it is not possible to successfully compete with any two scandictors on any individual image. This negative result shows that the scandiction problem is fundamentally different and more challenging than the previously studied problems, such as prediction and compression, where competition with an arbitrary finite set of schemes in the individual sequence setting is well known to be an attainable goal. However, in Theorem 4 of Section 3, we show that for the class of spatially stationary random fields, and subject to mild conditions on the prediction loss function, one can compete with any finite set of scandictors (under minimum expected loss). Furthermore, in Theorem 8, our main result in this section, we introduce a universal scandictor for any spatially stationary random field. Section 3 also includes almost surely analogues of the above theorems for mixing random fields and basic results on cases where universal scandiction of individual images is possible.

In Section 4, we derive upper bounds on the excess loss incurred when non-optimal scanners are used, with optimal prediction schemes. Namely, we consider the scenario where one cannot use a universal scandictor (or the optimal scan for a given random field), and instead uses an arbitrary scanning order, accompanied by the optimal predictor for that scan. In a sense, the results of Section 4 can be used to assess the sensitivity of the scandiction performance to the scanning order. Furthermore, in Section 4 we also 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. Section 5 includes a few concluding remarks and open problems.

In [27], the second part of this two-part paper, we consider sequential decision making for *noisy* data arrays. Namely, the decision maker observes a noisy version of the data, yet, it is judged with respect to the clean data. As the clean data is not available, two distinct cases are interesting to consider. The first, *scanning and filtering*, is when Y_{Ψ_t} is available in the estimation of X_{Ψ_t}, i.e., F_t depends on Y_{Ψ_1} to Y_{Ψ_t}, where $\{Y\}$ is the noisy data. The second, noisy scandiction, is when the noisy observation at the current site is not available

to the decision maker. In both cases, the decision maker cannot evaluate its performance precisely, as $l(x_{\Psi_t}, F_t)$ cannot be computed. Yet, many of the results for noisy scandiction are extendable from the noiseless case, similarly as results for noisy prediction were extended from results for noiseless prediction [28]. The scanning and filtering problem, however, poses new challenges and requires the use of new tools and techniques. Thus, in [27], we formally define the best achievable performance in these cases, derive bounds on the excess loss when non-optimal scanners are used and present universal algorithms. A special emphasis is given on the cases of binary random fields corrupted by a binary memoryless channel and real-valued fields corrupted by Gaussian noise.

2   Problem Formulation

The following notation will be used throughout this paper.[1] Let A denote the alphabet, which is either discrete or the real line. 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. Let $\mathcal{R}_\square$ be the set of all rectangles of the form $V = \mathbb{Z}^2 \cap ([m_1, m_2] \times [n_1, n_2])$. As a special case, denote by V_n the square $\{0, \ldots, n-1\} \times \{0, \ldots, n-1\}$. For $V \subset \mathbb{Z}^2$, let the interior diameter of V be

$$R(V) \overset{\triangle}{=} \sup\{r : \exists c \text{ s.t. } B(c, r) \subseteq V\}, \tag{1}$$

where $B(c, r)$ is a closed ball (under the l_1-norm) of radius r centered at c. Throughout, $\log(\cdot)$ will denote the natural logarithm, and entropies will be measured in nats.

Definition 1 ([22]). 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

---

[1]For easy reference, we try to follow the notation of [22] whenever possible.
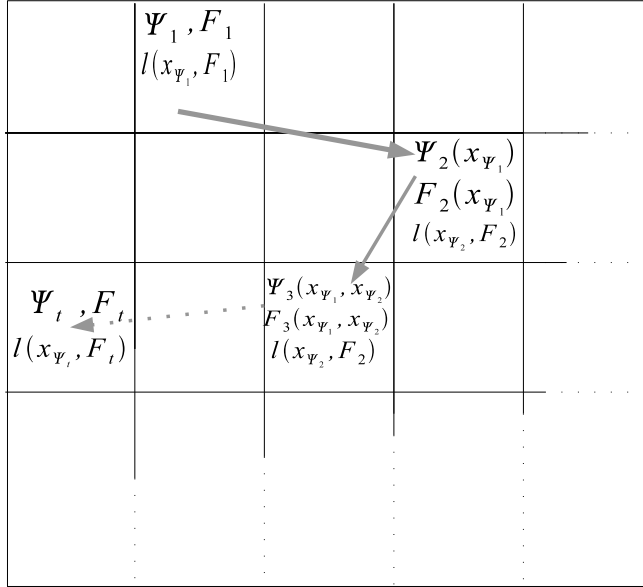
Figure 1: A graphical representation of the scandiction process. A scandictor (Ψ, F) first chooses an initial site Ψ_1. It then gives its prediction for the value at that site, F_1. After observing the true value at Ψ_1, it suffers a loss $l(x_{\Psi_1}, F_1)$, chooses the next site to be visited, $\Psi_2(x_{\Psi_1})$, gives its prediction for the value at that site, $F_2(x_{\Psi_1})$, and so on.

be visited at time t, with the property that

$$\left\{ \Psi_1, \Psi_2(x_{\Psi_1}), \Psi_3(x_{\Psi_1}, x_{\Psi_2}) \ldots, \Psi_{|B|} \left(x_{\Psi_1}, \ldots, x_{\Psi_{|B|-1}} \right) \right\} = B, \quad \forall x \in A^B. \tag{2}$$

- $\{F_t\}_{t=1}^{|B|}$ 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 the 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. At this point, it is important to note that scandictors for *infinite* data arrays are not considered in this paper. Definition 1, and the results to follow, consider only scandictors for finite sets of sites, ones which can be viewed merely as a reordering of the sites in a finite set B. We will consider, though, the limit as the size of the array tends to infinity. A scandictor, such that there exists a finite set of sites B, for which there is no deterministic finite point in time by which all sites in B are scanned, is not included in the scope of Definition 1. Figure 1 includes a graphical representation of the scandiction process.

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

$$L_{(\Psi,F)}(x_{V_n}) = \sum_{t=1}^{|V_n|} l\left(x_{\Psi_t}, F_t(x_{\Psi_1}, \ldots, x_{\Psi_{t-1}}) \right), \tag{3}$$

8

where $l : A \times D \to [0, \infty)$ is a given loss function. Throughout this paper, 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

$$U(l, Q_B) = \inf_{(\Psi, F) \in \mathcal{S}(B)} E_{Q_B} \frac{1}{|B|} L_{(\Psi, F)}(X_B), \tag{4}$$

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

$$U(l, Q) = \lim_{n \to \infty} U(l, Q_{V_n}). \tag{5}$$

By [22, Theorem 1], the limit in (5) exists for any $Q \in \mathcal{M}_S(\Omega)$ and, in fact, for any sequence $\{B_n\}$ of elements of $\mathcal{R}_\square$ for which $R(B_n) \to \infty$ we have

$$U(l, Q) = \lim_{n \to \infty} U(l, Q_{B_n}) = \inf_{B \in \mathcal{R}_\square} U(l, Q_B). \tag{6}$$

2.1 Finite-Set Scandictability

It will be constructive to refer to the *finite set* scandictability as well. Let $\mathcal{F} = \{\mathcal{F}_n\}$ be a sequence of finite sets of scandictors, where for each n, $|\mathcal{F}_n| = \lambda < \infty$, and the scandictors in $\mathcal{F}_n$ are defined for the finite set of sites V_n. A possible scenario is one in which one has a set of "scandiction rules", each of which defines a unique scanner for each n, but all these scanners comply with the same rule. In this case, $\mathcal{F} = \{\mathcal{F}_n\}$ can also be viewed as a finite set $\mathcal{F}$ which includes sequences of scandictors. For example, $|\mathcal{F}_n| = 2$ for all n, where for each n, $\mathcal{F}_n$ includes one scandictor which scans the data row-wise and one which scans the data column-wise. We may also consider cases in which $|\mathcal{F}_n|$ increases with n (but remains finite for every finite n). For $Q \in \mathcal{M}_S(\Omega)$ and $\mathcal{F} = \{\mathcal{F}_n\}$, we thus define the finite set scandictability of Q as the limit

$$U_{\mathcal{F}}(l, Q) \triangleq \lim_{n \to \infty} \min_{(\Psi, F) \in \mathcal{F}_n} E_{Q_{V_n}} \frac{1}{|V_n|} L_{(\Psi, F)}(X_{V_n}), \tag{7}$$

if it exists. Observe that the sub-additivity property of the scandictability as defined in [22], which was fundamental for the existence of the limit in (5), does not carry over verbatim to finite set scandictability. This is for the following reason. Suppose $(\Psi, F) \in \mathcal{S}$ is the optimal scandictor for X_V and $(\Psi', F') \in \mathcal{S}$ is optimal for X_U (assume $V \cap U = \emptyset$). When scanning $X_{V \cup U}$, one may not be able to apply (Ψ, F) for X_V and then (Ψ', F') for X_U, as this scandictor might not be in $\mathcal{S}$. Hence, we seek a universal scheme which competes successfully (in a sense soon to be defined) with a sequence of finite sets of scandictors $\{\mathcal{F}_n\}$, even when the limit in (7) does not exist.

3   Universal Scandiction

The problem of universal prediction is well studied, with various solutions to both the stochastic setting as well as the individual. In this section, we study the problem of universal *scandiction*. Notwithstanding strongly related to its prediction analogue, we first show that this problem is fundamentally different in several aspects, mainly due to the enormous degree of freedom in choosing the scanning order. Particularly, we first give a negative result, stating that while in the prediction problem it is possible to compete with any finite number of predictors and on every individual sequence, in the scandiction problem one cannot even compete with any two scandictors on a given individual data array. Nevertheless, we show that in the setting of stationary random fields, and under the minimum expected loss criterion, it is possible to compete with any finite set of scandictors. We then show that the set of *finite-state scandictors* is capable of achieving the scandictability of any spatially stationary source. In Theorem 8, our main result in this section, we give a universal algorithm which achieves the scandictability of any spatially stationary source.

3.1   A Negative Result on Scandiction

Assume both the alphabet A and the prediction space D are $[0, 1]$. Let l be any *non-degenerated* loss function, in the sense that prediction of a bernoulli sequence under it results in a positive expected loss. As an example, squared or absolute error can be kept in mind, though the result below applies to many other loss functions. The following theorem asserts that in the individual image scenario, it is not possible to compete successfully with any two arbitrary scandictors (it is possible, though, to compete with *some* scandictor sets, as proved in Section 3.6).

Theorem 2. *Let $A = D = [0, 1]$ and assume l is a non-degenerated loss function. There exist two scandictors $(\Psi, F)_1$ and $(\Psi, F)_2$ for V_n, such that for any scandictor (Ψ, F) for V_n there exists x_{V_n} for which*

$$L_{(\Psi,F)}(x_{V_n}) - \min\{L_{(\Psi,F)_1}(x_{V_n}), L_{(\Psi,F)_2}(x_{V_n})\} = \Theta(|V_n|). \tag{8}$$

In words, there exist two scandictors such that for any third scandictor, there exists an individual image for which the redundancy when competing with the two scandictors does not vanish. Theorem 2 marks a fundamental difference between the case where reordering of the data is allowed, e.g., scanning of multidimensional data or even reordering of one-dimensional

data, and the case where there is one natural order for the data. For example, using the exponential weighting algorithm discussed earlier, it is easy to show that in the prediction problem (i.e., with no scanning), it is possible to compete with *any finite set* of predictors under the above alphabets and loss functions. Thus, although the scandiction problem is strongly related to its prediction analogue, the numerous scanning possibilities result in a substantially richer and more challenging problem.

Theorem 2 is a direct application of the lemma below.

Lemma 3. *Let $A = D = [0, 1]$ and assume l is a non-degenerated loss function. There exist a random field X_{V_n} and two scandictors $(\Psi, F)_1$ and $(\Psi, F)_2$ for V_n, such that for any scandictor (Ψ, F) for V_n,*

$$EL_{(\Psi,F)}(X_{V_n}) - E\min\{L_{(\Psi,F)_1}(X_{V_n}), L_{(\Psi,F)_2}(X_{V_n})\} = \Theta(|V_n|). \tag{9}$$

Lemma 3 gives another perspective on the difference between the scandiction and prediction scenarios. The lemma asserts that when ordering of the data is allowed, one cannot achieve a vanishing redundancy with respect to the *expected value of the minimum* among a set of scandictors. This should be compared to the prediction scenario (no reordering), where one can compete successfully not only with respect to the minimum of the expected losses of all the predictors, but also with respect to the expected value of the minimum (for example, see [29, Corollary 1]). The main result of this section, however, is that for any stationary random field and under mild conditions on the loss function, one can compete successfully with any finite set of scandictors when the performance criterion is the *minimum expected loss*.

Proof. (Lemma 3) Let Y_{V_n} be a random field such that $Y(1, 1)$ is distributed uniformly on $[0, 1]$, and $Y_{V_n} \setminus Y(1, 1) = Y(1, 2), \ldots, Y(1, n), Y(2, 1), Y(2, 2), \ldots, Y(n, n)$ are simply the first $n^2 - 1$ bits in the binary representation of $Y(1, 1)$ (ordered row-wise). Note that $Y_{V_n} \setminus Y(1, 1)$ are i.i.d. unbiased bits, yet conditioned on $Y(1, 1)$, they are deterministic and known. Assume now that X_{V_n} is a random cyclic shift of Y_{V_n}, in the same row-wise order Y_{V_n} was created.

For concreteness, we assume the squared error loss function. In this case, it is easy to identify the constant of the $\Theta(\cdot)$ expression in (8). However, the computations below are easily generalized to other non-degenerated loss functions. We first show that the expected cumulative squared error of any scandictor on X_{V_n} is at least $(n^2 + 1)/8$, as the expected number of steps until the real valued site is located is $(n^2 + 1)/2$, with a loss of $1/4$ until that time. More specifically, let J be the random number of cyclic shifts, that is, J is uniformly distributed on $\{0, 1, \ldots, n^2 - 1\}$. For fixed j, let G be the random index such that Ψ_G is the

real-valued X (i.e., G is the time the real valued random variable is located by the scanner Ψ). Let ϕ_s denote the Bayes envelope associated with the squared error loss, i.e.,

$$\phi_s(p) = \min_{q \in [0,1]}[(1-p)q^2 + p(q-1)^2]. \tag{10}$$

For any scandictor (Ψ, F), we have,

$$
\begin{aligned}
EL_{(\Psi,F)}(X_{V_n}) &= E_J E\left\{ \sum_{i=1}^{n^2} \left(X_{\Psi_i} - F_i(X_{\Psi_1}^{\Psi_{i-1}}) \right)^2 \middle| j \right\} \\
&= E_J E\left\{ \sum_{i=1}^{G} \left(X_{\Psi_i} - F_i(X_{\Psi_1}^{\Psi_{i-1}}) \right)^2 + \sum_{i=G+1}^{n^2} \left(X_{\Psi_i} - F_i(X_{\Psi_1}^{\Psi_{i-1}}) \right)^2 \middle| j \right\} \\
&\geq E_J E\left\{ \sum_{i=1}^{G} \left(X_{\Psi_i} - F_i(X_{\Psi_1}^{\Psi_{i-1}}) \right)^2 \middle| j \right\} \\
&\geq E_J E\left\{ \sum_{i=1}^{G} \phi_s \left(P(X_{\Psi_i} | X_{\Psi_1}^{\Psi_{i-1}}) \right) \middle| j \right\} \\
&= E_J E\left\{ G\phi_s \left(P(X_{\Psi_1} = 1) \right) \middle| j \right\} \\
&= \frac{1}{4} E\{G\} \\
&= \frac{n^2 + 1}{8}. \tag{11}
\end{aligned}
$$

On the other hand, consider the expected *minimum of the losses* of the following two scandictors: $(\Psi, F)_1$ which scandicts X_{V_n} row-wise from $X(1,1)$ to $X(n,n)$, and $(\Psi, F)_2$ which scandicts X_{V_n} row-wise from $X(n,n)$ to $X(1,1)$. Using the same method as in (11), it is possible to show that this expected loss is *smaller* than $n^2/16 + o(n^2)$, as the expected number of steps until *the first* locates the real-valued site is $(n^2 + 1)^2/(4n^2)$, after which zero loss is incurred. This is since once the real-valued site is located, the rest of the values can be calculated by the predictor by cyclic shifting the binary representation of the real-valued pixel. This completes the proof. $\square$

Proof. (Theorem 2) By Lemma 3, there exists a stochastic setting under which the expected minimum of the losses of two scandictors is smaller than the expected loss of any single scandictor. Thus, for any scandictor there exists an individual image on which it cannot compete successfully with the two scandictors. $\square$

12

3.2 Universal Scandiction With Respect to Arbitrary Finite Sets

As mentioned in Section 1, straightforward implementation of the exponential weighting algorithm is not feasible, since one may not be able to alternate from one expert to the other at wish. However, the exponential weighting algorithm was found useful in several lossy source coding works such as Linder and Lugosi [30], Weissman and Merhav [31], Gyorgy *et. al.* [32] and the derivation of sequential strategies for loss functions with memory [33], all of which confronted a similar problem. A common method used in these works, is the alternation of experts only once every block of input symbols, necessary to bear the price of this change (e.g., transmitting the description of the chosen quantizer [30]-[32]). Thus, although the difficulties in these examples differ from those we confront here, the solution suggested therein, which is to persist on using the same expert for a significantly long block of data before alternating it, was found useful in our universal scanning problem.

Particularly, we divide the data array into smaller blocks and alternate scandictors only each time a new block of data is to be scanned. Unlike the case of sequential prediction dealt with in [33], here the scandictors must be *restarted* each time a new block is scanned, as it is not at all guaranteed that all the scandictors scan the data in the same (or any) block-wise order (i.e., it is not guaranteed that a scandictor for V_n divides the array to sub-blocks of size $m \times m$ and scans each of them separately). Hence, in order to prove that it is possible to compete with the best scandictor at each stage n, we go through two phases. In the first, we prove that an exponential weighting algorithm may be used to compete with the best scandictor among those operating in a block-wise order. This part of the proof will refer to any given data array (deterministic scenario). In the second phase, we use the stationarity of the random field to prove that a block-wise scandictor may perform essentially as well as one scanning the data array as a whole. The following theorem stands at the basis of our results, establishing the existence of a universal scandictor which competes successfully with any finite set of scandictors.

Theorem 4. *Let X be a stationary random field with a probability measure Q. Let $\mathcal{F} = \{\mathcal{F}_n\}$ be an arbitrary sequence of scandictor sets, where $\mathcal{F}_n$ is a set of scandictors for V_n and $|\mathcal{F}_n| = \lambda < \infty$ for all n. Then, there exists a sequence of scandictors $\{(\hat{\Psi}, \hat{F})_n\}$, where $(\hat{\Psi}, \hat{F})_n$ is a scandictor for V_n, independent of Q, for which*

$$\liminf_{n \to \infty} E_{Q_{V_n}} E \frac{1}{|V_n|} L_{(\hat{\Psi}, \hat{F})_n}(X_{V_n}) \leq \liminf_{n \to \infty} \min_{(\Psi, F) \in \mathcal{F}_n} E_{Q_{V_n}} \frac{1}{|V_n|} L_{(\Psi, F)}(X_{V_n}) \qquad (12)$$

13

for any $Q \in \mathcal{M}_S(\Omega)$, where the inner expectation in the l.h.s. of (12) is due to the possible randomization in $(\hat{\Psi}, \hat{F})_n$.

Before we prove Theorem 4, let us discuss an "individual image" type of result, which will later be the basis of the proof. Let x_{V_n} denote an individual $n \times n$ data array. For $m < n$, define $K \triangleq \lceil \frac{n}{m} \rceil - 1$. We 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, $0 \leq i \leq (K + 1)^2 - 1$ 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 (continuous) raster scan from the upper left corner. That is, the first line of blocks is scanned left to right, the second line is scanned right to left, and so on. We will refer to this scan simply as "raster scan".

As mentioned, the suggested algorithm scans the data in x_{V_n} block-wise, that is, it does not apply any of the scandictors in $\mathcal{F}_n$, only scandictors from $\mathcal{F}_m$. Omitting m for convenience, 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. Since we assumed the scandictors are capable of scanning only square blocks, 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

$$P_i\left(j | \{L_{j,i}\}_{j=1}^{\lambda}\right) = \frac{e^{-\eta L_{j,i}}}{\sum_{j=1}^{\lambda} e^{-\eta L_{j,i}}}, \tag{13}$$

where $\lambda = |\mathcal{F}_m|$. We offer the following algorithm for a block-wise scan of the data array x. For each $0 \leq i \leq (K + 1)^2 - 1$, 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. Namely, the universal scandictor $(\hat{\Psi}, \hat{F})_n$, promised by Theorem 4, is the one which divides the data to blocks, performs a raster scan of the data block-wise, and uses the above algorithm to decide which scandictor out of $\mathcal{F}_m$ to use for each block.

It is clear that both the block size and the number of blocks should tend to infinity with n in order to achieve meaningful results. Thus, we require the following: a. $m = m(n)$ tends to infinity, but strictly slower than n, i.e., $m(n) = o(n)$. b. $m(n)$ is an integer-valued monotonically increasing function, such that for each $K \in \mathbb{Z}$ there exists n such that $m(n) = K$. The results are summarized in the following two propositions, the first of which asserts that for $m(n) = o(n)$, vanishing redundancy is indeed possible, while the second asserts that under slightly stronger requirements on $m(n)$, this is also true in the a.s. sense (with respect to the

14

random selection of the scandictors in the algorithm).

Proposition 5. *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, for any x_{V_n},*

$$\bar{L}_{alg}(x_{V_n}) - L_{min}(x_{V_n}) \leq m(n)(n + m(n))\sqrt{\log \lambda} \frac{l_{max}}{\sqrt{2}}. \tag{14}$$

Proposition 6. *Assume $m(n) = o\left(n^{1/3}\right)$. Then, for any image x_{V_n}, the difference between the normalized cumulative loss of the proposed algorithm and that of the best scandictor in $\mathcal{F}_m$, operating block-wise, converges to 0 with probability 1 with respect to the randomized scandictor selection of the algorithm.*

The proofs of Propositions 5 and 6 are rather technical and are based on the very same methods used in [34] and [33]. See Appendices A.1 and A.2 for the details.

On the more technical side, note that the suggested algorithm has "finite horizon," that is, one has to know the size of the image in order to divide it to blocks, and only then can the exponential weighting algorithm be used. It is possible to extend the algorithm to infinite horizon. The essence of this generalization is in dividing the infinite image into blocks of exponentially growing size[2], and to apply the finite horizon algorithm for each block. We may now proceed to the proof of Theorem 4.

Proof of Theorem 4. Since the result of Proposition 5 applies to any individual data array, it certainly applies after taking the expectation with respect to Q. Therefore,

$$E_{Q_{V_n}} \frac{1}{n^2} \bar{L}_{alg} - E_{Q_{V_n}} \frac{1}{n^2} L_{min} \leq \frac{m(n)}{n} l_{max} \sqrt{2 \log \lambda}. \tag{15}$$

However, remember that we are not interested in competing with $E_{Q_{V_n}} \frac{1}{n^2} L_{min}$, as this is the performance of the best *block-wise* scandictor. We wish to compete with the best scandictor

---

[2]For example, take four blocks of size $l \times l$, then three of size $2l \times 2l$, and so on.

operating on the *entire* data array X_{V_n}, that is, on the whole image of size $n \times n$. We have

$$
\begin{aligned}
E_{Q_{V_n}} \frac{1}{n^2} L_{min} &= E_{Q_{V_n}} \frac{1}{n^2} \min_{(\Psi,F)_j \in \mathcal{F}_{m(n)}} \sum_{i=1}^{(K+1)^2} L_j(X^i) \\
&\leq \min_{(\Psi,F)_j \in \mathcal{F}_{m(n)}} E_{Q_{V_n}} \frac{1}{n^2} \sum_{i=1}^{(K+1)^2} L_j(X^i) \\
&\stackrel{(a)}{\leq} \min_{(\Psi,F)_j \in \mathcal{F}_{m(n)}} \frac{1}{n^2} \cdot \left[K^2 E_{Q_{V_n}} L_j(X^0) \right. \\
&\quad \left. + 2Km(n)(n - Km(n))l_{max} + (n - Km(n))^2 l_{max} \right] \\
&\leq \min_{(\Psi,F)_j \in \mathcal{F}_{m(n)}} E_{Q_{V_n}} \frac{1}{m(n)^2} L_j(X^0) + 2\frac{m(n)}{n} l_{max}, \quad (16)
\end{aligned}
$$

where (a) follows from the stationarity of Q, the assumption that each $(\Psi, F)_j$ operates in the same manner on each $m(n) \times m(n)$ block, no matter what its coordinates are, and the fact that each $(\Psi, F)_j$ may incur maximal loss on non-square rectangles. From (15) and (16), we have

$$
\begin{aligned}
E_{Q_{V_n}} \frac{1}{n^2} \bar{L}_{alg} &\leq E_{Q_{V_n}} \frac{1}{m(n)^2} L_{j^*(m(n))}(X^0) + 2\frac{m(n)}{n} l_{max} + \frac{m(n)}{n} l_{max} \sqrt{2\log\lambda} \\
&= E_{Q_{V_n}} \frac{1}{m(n)^2} L_{j^*(m(n))}(X^0) + O\left(\frac{m(n)}{n} \sqrt{\log\lambda} \right) \quad (17)
\end{aligned}
$$

where $(\Psi, F)_{j^*(m(n))}$ is the scandictor achieving the minimum in (16). Finally, by our assumptions on $\{m(n)\}$, we have

$$
\begin{aligned}
\inf_{k \geq n} \left\{ E_{Q_{V_k}} \frac{1}{k^2} \bar{L}_{alg} \right\} &\leq \inf_{k \geq n} \left\{ E_{Q_{V_k}} \frac{1}{m(k)^2} L_{j^*(m(k))}(X^0) + \frac{m(k)}{k} l_{max}(2 + \sqrt{2\log\lambda}) \right\} \\
&\leq \inf_{k \geq n} \left\{ E_{Q_{V_k}} \frac{1}{m(k)^2} L_{j^*(m(k))}(X^0) \right\} + \frac{m(n)}{n} l_{max}(2 + \sqrt{2\log\lambda}) \\
&\leq \inf_{k \geq n} \left\{ E_{Q_{V_k}} \frac{1}{k^2} L_{j^*(k)}(X_{V_k}) \right\} + \frac{m(n)}{n} l_{max}(2 + \sqrt{2\log\lambda}). \quad (18)
\end{aligned}
$$

Taking the limit as $n \to \infty$ and using the fact that $m(k)/k \to 0$ together with the arbitrariness of k, gives:

$$
\liminf_{n \to \infty} E_{Q_{V_n}} \frac{1}{n^2} \bar{L}_{alg} \leq \liminf_{n \to \infty} E_{Q_{V_n}} \frac{1}{n^2} L_{j^*(n)}(X_{V_n}), \quad (19)
$$

which completes the proof of (12). $\qquad \square$

It is evident from (14) and (18) that although the results of Theorem 4 and Proposition 5 are formulated for fixed $\lambda < \infty$ (the cardinality of the scandictor set), these results hold for the more general case of $\lambda = \lambda(n)$, as long as the redundancy vanishes, i.e., as long as $m(n) = o(n)$ and $\lambda(n)$ is such that $\frac{m(n)}{n} \sqrt{\log\lambda} \to 0$ when $n \to \infty$. The requirement that

$\lambda(n) = o\left(e^{\frac{n^2}{m(n)^2}}\right)$ still allows very large scandictor sets, especially when $m(n)$ grows slowly with n. Furthermore, it is evident from equation (17) that whenever the redundancy vanishes, the statement of Theorem 4 is valid with $\limsup$ as well ,i.e.,

$$\limsup_{n \to \infty} E_{Q_{V_n}} E \frac{1}{|V_n|} L_{(\hat{\Psi}, \hat{F})_n}(X_{V_n}) \leq \limsup_{n \to \infty} \min_{(\Psi, F) \in \mathcal{F}_n} E_{Q_{V_n}} \frac{1}{|V_n|} L_{(\Psi, F)}(X_{V_n}). \tag{20}$$

3.3 Finite-State Scandiction

Consider now the set of finite-state scandictors, very similar to the set of finite-state encoders described in [14]. At time $t = 1$, a *finite-state scandictor* starts at an arbitrary initial site Ψ_1, with an arbitrary initial state $s_0 \in S$ and gives $F(s_0)$ as its prediction for x_{Ψ_1}. Only then it observes x_{Ψ_1}. After observing x_{Ψ_i}, it computes its next state, s_i, according to $s_i = g(s_{i-1}, x_{\Psi_i})$ and advances to the next site, $x_{\Psi_{i+1}}$, according to $\Psi_{i+1} = \Psi_i + d(s_i)$, where $g : S \times A \mapsto S$ is the next state function and $d : S \mapsto B$ is the displacement function, $B \subset \mathbb{Z}^2$ denoting a fixed finite set of possible relative displacements. It then gives its prediction $F(s_i)$ to the value $x_{\Psi_{i+1}}$. Similarly to [14], we assume the alphabet A includes an additional "End of File" (EoF) symbol to mark the image edges. The following lemma and the theorem which follows establish the fact that the set of finite-state scandictors is indeed rich enough to achieve the scandictability of any stationary source, yet not too rich to compete with.

Lemma 7. *Let $\mathcal{F}_S = \{(\Psi, F)_j\}$ be the set of all finite-state scandictors with at most S states. Then, for any $Q \in \mathcal{M}_S(\Omega)$,*

$$\lim_{S \to \infty} U_{\mathcal{F}_S}(l, Q) = U(l, Q). \tag{21}$$

That is, the scandictability of any spatially stationary source is asymptotically achieved with finite-state scandictors.

Proof. Take $B = V_m$ and let $(\bar{\Psi}, \bar{F})_m$ be the achiever of the infimum in (4). That is,

$$E_{Q_{V_m}} \frac{1}{m^2} L_{(\bar{\Psi}, \bar{F})_m}(X_{V_m}) \leq \inf_{(\Psi, F) \in \mathcal{S}(V_m)} E_{Q_{V_m}} \frac{1}{m^2} L_{(\Psi, F)}(X_{V_m}). \tag{22}$$

Since V_m is a rectangle of size $m \times m$, the scandictor $(\bar{\Psi}, \bar{F})_m$ is certainly implementable with a finite-state machine having $S(m) < \infty$ states. In other words, since V_m is finite, *any* scanning rule $\Psi_t : A^{t-1} \mapsto B$ and *any* prediction rule $F_t : A^{t-1} \mapsto A$ can be implemented with a finite-state machine having at most $\tilde{S}(m) = A^{m^2} \times m^2$ states, where in a straightforward implementation A^{m^2} states are required to account for all possible inputs and m^2 states are required to implement a counter.

Now, for an $n \times n$ image (assuming now that m divides n, as dealing with the general case can be done in the exact same way as (16)), we take $(\bar{\Psi}', \bar{F}')_n$ to be the scandictor which scans the image in the block-by-block raster scan described earlier, applying $(\bar{\Psi}, \bar{F})_m$ to each $m \times m$ block. Namely, $\bar{\Psi}'$ scans all the blocks in the first m lines left-to-right until it reaches an EoF symbol, then moves down m lines, scans all the blocks right-to-left until an EoF is reached, and so on. The predictor $\bar{F}'$ simply implements $\bar{F}$ for each block separately, i.e., it resets to its initial values at the beginning of each block. It is clear that the scanner $\bar{\Psi}'$ is implementable with a finite-state machine having $S(m) = \tilde{S}(m) + 2 < \infty$ states and thus $(\bar{\Psi}', \bar{F}') \in \mathcal{F}_{S(m)}$.

From the stationarity of Q, we have

$$
\begin{aligned}
\inf_{(\Psi,F)\in\mathcal{S}(V_n)} E_{Q_{V_n}} \frac{1}{n^2} L_{(\Psi,F)}(X_{V_n}) \quad &\leq \quad \min_{(\Psi,F)\in\mathcal{F}_{S(m)}} E_{Q_{V_n}} \frac{1}{n^2} L_{(\Psi,F)}(X_{V_n}) \\
&\leq \quad E_{Q_{V_n}} \frac{1}{n^2} L_{(\bar{\Psi}',\bar{F}')_n}(X_{V_n}) \\
&= \quad E_{Q_{V_m}} \frac{1}{m^2} L_{(\bar{\Psi},\bar{F})_m}(X_{V_m}) \\
&\leq \quad \inf_{(\Psi,F)\in\mathcal{S}(V_m)} E_{Q_{V_m}} \frac{1}{m^2} L_{(\Psi,F)}(X_{V_m}). \quad (23)
\end{aligned}
$$

Taking the limits $\limsup_{n\to\infty}$ and $\liminf_{n\to\infty}$, by (6), we have

$$
\begin{aligned}
U(l,Q) \quad &\leq \quad \limsup_{n\to\infty} \min_{(\Psi,F)\in\mathcal{F}_{S(m)}} E_{Q_{V_n}} \frac{1}{n^2} L_{(\Psi,F)}(X_{V_n}) \\
&\leq \quad \inf_{(\Psi,F)\in\mathcal{S}(V_m)} E_{Q_{V_m}} \frac{1}{m^2} L_{(\Psi,F)}(X_{V_m}) \quad (24)
\end{aligned}
$$

and

$$
\begin{aligned}
U(l,Q) \quad &\leq \quad \liminf_{n\to\infty} \min_{(\Psi,F)\in\mathcal{F}_{S(m)}} E_{Q_{V_n}} \frac{1}{n^2} L_{(\Psi,F)}(X_{V_n}) \\
&\leq \quad \inf_{(\Psi,F)\in\mathcal{S}(V_m)} E_{Q_{V_m}} \frac{1}{m^2} L_{(\Psi,F)}(X_{V_m}). \quad (25)
\end{aligned}
$$

The proof is completed (including the existence of the limit in the l.h.s. of (21)) by taking m to infinity, applying (6), and remembering that $U_{\mathcal{F}_S}(l,Q)$ is monotone in S, thus the convergence of the sub-sequence $\{U_{\mathcal{F}_{S(m)}}(l,Q)\}_{m=1}^{\infty}$ implies the convergence of the sequence $\{U_{\mathcal{F}_S}(l,Q)\}_{S=1}^{\infty}$). $\qquad\square$

In words, Lemma 7 asserts that for any m, finite-state machines attain the $m \times m$ Bayesian scandictability for any stationary random field. Note that the reason such results are accomplishable with FSMs is their ability to scan the entire data, block by block, with a machine having no more than $S(m)$ states, regardless of the size of the complete data array. The number of the states depends only on the block size.

3.4 A Universal Scandictor for Any Stationary Random Field

We now show that a universal scandictor which competes successfully with all finite-state machines of the form given in the proof of Lemma 7, does exist and can, in fact, be implemented using the exponential weighting algorithm. In order to show that we assume that the alphabet A is finite and the prediction space D is either finite or bounded (such as the $|D| - 1$ simplex of probability measures on D). In the latter case we further assume that $l(x, F)$ is Lipschitz in its second argument for all x, i.e, there exists a constant c such that for all x, F and ϵ we have $|l(x, F) - l(x, F + \epsilon)| \leq c|\epsilon|$. The following theorem establishes, under the above assumptions, the existence of a universal scandictor for all stationary random fields.

Theorem 8. *Let X be a stationary random field over a finite alphabet A and a probability measure Q. Let the prediction space D be either finite or bounded (with $l(x, F)$ then being Lipschitz in its second argument). Then, there exists a sequence of scandictors $\{(\Psi, F)_n\}$, independent of Q, for which*

$$\lim_{n \to \infty} E_{Q_{V_n}} E \frac{1}{|V_n|} L_{(\Psi, F)_n}(X_{V_n}) = U(l, Q) \tag{26}$$

for any $Q \in \mathcal{M}_S(\Omega)$, where the inner expectation in the l.h.s. of (26) is due to the possible randomization in $(\Psi, F)_n$.

Proof. Assume first that the range D of the predictors $\{F_t\}$ is finite. Consider the exponential weighting algorithm described in the proof of Theorem 4, where at each $m(n) \times m(n)$ block the algorithm computes the cumulative loss of *every* possible scandictor for an $m(n) \times m(n)$ block, then chooses the best scandictor (according to the exponential weighting regime described therein) as its output for the next block. By (17), we have

$$E_{Q_{V_n}} \frac{1}{n^2} \bar{L}_{alg} \leq \min_{(\Psi, F) \in \mathcal{S}(V_{m(n)})} E_{Q_{V_{m(n)}}} \frac{1}{m(n)^2} L_{(\Psi, F)}(X^0) + O\left(\frac{m(n)}{n} \sqrt{\log \lambda}\right), \tag{27}$$

where $\mathcal{S}(V_{m(n)})$ is the set of all possible scandictors on $m(n) \times m(n)$ and λ is the size of that set. Since $\lambda = \lambda(m(n))$, all that is left to check is that the $O\left(\frac{m(n)}{n} \sqrt{\log \lambda}\right)$ expression indeed decays to zero as n tends to infinity.

Indeed, the number of possible scanners for a field B over an alphabet A is

$$\left| \left\{ \{\Psi_t\}_{t=1}^{|B|}, \Psi_t : A^{t-1} \mapsto B \right\} \right| = \prod_{k=0}^{|B|} (|B| - k)^{|A|^k}$$

$$\leq (|B|!)^{|A|^{|B|}}, \tag{28}$$

19

while the number of possible predictors is

$$\left|\left\{\{F_t\}_{t=1}^{|B|}, F_t : A^{t-1} \mapsto D\right\}\right| = \prod_{k=1}^{|B|} |D|^{|A|^k}$$

$$\leq |D|^{|B||A|^{|B|-1}}. \tag{29}$$

Thus, using the Stirling approximation, $\log k! \approx k \log k$, in the sense that $\lim_{k \to \infty} \frac{\log k!}{k \log k} = 1$, we have

$$\begin{aligned}
\frac{m(n)}{n}\sqrt{\log \lambda} &\leq \frac{m(n)}{n}\sqrt{\log\left[(m(n)^2!)^{|A|^{m(n)^2}}|D|^{m(n)^2|A|^{m(n)^2-1}}\right]} \\
&\approx \frac{m(n)}{n}\sqrt{2|A|^{m(n)^2}m(n)^2 \log m(n) + m(n)^2|A|^{m(n)^2-1}\log|D|} \\
&\approx \frac{m(n)^2}{n}\sqrt{|A|^{m(n)^2}\log m(n)}, \tag{30}
\end{aligned}$$

which decays to zero as $n \to \infty$ for any $m(n) = o(\sqrt{\log n})$. Namely, for $m(n) = o(\sqrt{\log n})$, equation (27) results in

$$\liminf_{n \to \infty} E_{Q_{V_n}}\frac{1}{n^2}\bar{L}_{alg} \leq \liminf_{n \to \infty} \min_{(\Psi,F) \in \mathcal{S}(V_{m(n)})} E_{Q_{V_{m(n)}}}\frac{1}{m(n)^2}L_{(\Psi,F)}(X^0), \tag{31}$$

and

$$\limsup_{n \to \infty} E_{Q_{V_n}}\frac{1}{n^2}\bar{L}_{alg} \leq \limsup_{n \to \infty} \min_{(\Psi,F) \in \mathcal{S}(V_{m(n)})} E_{Q_{V_{m(n)}}}\frac{1}{m(n)^2}L_{(\Psi,F)}(X^0). \tag{32}$$

Since $m(n) \to \infty$ as $n \to \infty$, by [22] the limit $\lim_{n \to \infty} \min_{(\Psi,F) \in \mathcal{S}(V_{m(n)})} E_{Q_{V_{m(n)}}}\frac{1}{m(n)^2}L_{(\Psi,F)}(X^0)$ exists and equals the scandictability of the source, $U(l,Q)$. However, by definition, $U(l,Q)$ is the best achievable scandiction performance for the source Q, hence,

$$\liminf_{n \to \infty} E_{Q_{V_n}}\frac{1}{n^2}\bar{L}_{alg} \geq U(l,Q), \tag{33}$$

which results in

$$\lim_{n \to \infty} E_{Q_{V_n}}\frac{1}{n^2}\bar{L}_{alg} = U(l,Q). \tag{34}$$

For the case of infinite (but bounded) range D, similarly to [25], we use the fact that the loss function l is Lipschitz and take an ϵ-approximation of D. We thus have

$$\begin{aligned}
E_{Q_{V_n}}\frac{1}{n^2}\bar{L}_{alg} \leq \min_{(\Psi,F) \in \mathcal{S}(V_{m(n)})} E_{Q_{V_{m(n)}}}\frac{1}{m(n)^2}L_{(\Psi,F)}(X^0) \\
+ cm(n)^2\epsilon(m(n)) + O\left(\frac{m(n)}{n}\sqrt{\log\lambda}\right) \tag{35}
\end{aligned}$$

for some constant c. Choosing $\epsilon(m(n)) = \frac{1}{m(n)^4}$ results in $|D| = \frac{1}{2}m(n)^4$, hence $\frac{m(n)}{n}\sqrt{\log\lambda}$ still decays to zero for any $m(n) = O(\sqrt{\log n})$ and (34) is still valid. $\qquad\square$

Note that the proof of Theorem 8 does not use the well established theory of universal prediction. Instead, the exponential weighting algorithm is used for all possible scans (within a block) as well as *all possible predictors*. This is since important parts of the work on prediction in the probabilistic scenario include some assumption on the stationarity of the measure governing the process, such as stationarity or asymptotically mean stationarity [35].[3] In the scandiction scenario, however, the properties of the output sequence are not easy to determine, and it is possible, in general, that the output sequence is not stationary or ergodic even if the input data array is. Thus, although under certain assumptions, one can use a single universal predictor, applied to any scan in a certain set of scans, this is not the case in general.

3.5 Universal Scandiction for Mixing Random Fields

The proof of Theorem 4 established the universality of $(\hat{\Psi}, \hat{F})_n$ under the expected cumulative loss criterion. In order to establish its universality in the Q-a.s. sense, we examine the conditions on the measure Q such that the following equality holds.

$$\lim_{n \to \infty} \frac{1}{K^2} \sum_{i=1}^{K^2} L_j(x^i) = E_{Q_{V_m}} L_j(X^0) \quad Q-a.s. \tag{36}$$

To this end, we briefly review the conditions for the individual ergodic theorem for general dynamical systems given in [37], specialized for $\mathbb{Z}^2$. Let $\{A_n\}$ be a sequence of subsets of $\mathbb{Z}^2$. For each n, the set A_n is the set of sites over which the arithmetical average is taken. Let $A \triangle B$ denote the symmetric difference between the sets A and B, $A \cup B \smallsetminus A \cap B$, and remember that $\tau_i(x)_j = x_{j+i}$.

Condition 1 ([37, $E1'$]). For all $i \in \mathbb{Z}^2$,

$$\lim_{n \to \infty} \frac{|A_n \triangle \tau_i(A_n)|}{|A_n|} = 0. \tag{37}$$

Condition 2 ([37, $E3''$]). There exists a constant $C_1 < \infty$ such that for all n,

$$|k : k = i - j, \quad i, j \in A_n| \le C_1 |A_n|. \tag{38}$$

Condition 3 ([37, $E4$]). There exists a sequence of measurable sets $\{M_n\}$ such that,

$$\liminf_{n \to \infty} \frac{|k : k = i + j, \quad i \in A_n, j \in M_n|}{|M_n|} = C_2 < \infty. \tag{39}$$

By [37, Theorem 6.1$'$], if the sequence $\{A_n\}$ satisfies conditions 1-3, then, for any stationary random field X with $E|X_0| < \infty$, we have,

$$\lim_{n \to \infty} \frac{1}{|A_n|} \sum_{i \in A_n} X_i = E\{X_0 | \mathcal{I}\} \quad Q-a.s., \tag{40}$$

---

[3]An important exception is the Kalman filter [36, Section 7.7].

where Q is the measure governing X and $\mathcal{I}$ is the σ-algebra of invariant sets of Ω, that is,

$$A \in \mathcal{I} \text{ iff } \tau_i(A) = A \quad \text{for all } i \in \mathbb{Z}^2. \tag{41}$$

If Q is ergodic, namely, for each $A \in \mathcal{I}$, $Q(A) \in \{0,1\}$, then $E\{X_0|\mathcal{I}\}$ is deterministic and equals EX_0.

Clearly, since $L_j(x^i)$ depends on a set of m^2 sites, with the average in taken over the sets $A_n = \{i : i = m \cdot j, j \in V_K\}$, (36) may not hold, even if Q is ergodic, as, for example, Condition 1 is not satisfied.[4] These two obstacles can be removed by defining an alternative random field, $\tilde{X}$, over the set of sites $m \cdot \mathbb{Z}^2 = \{j : j = m \cdot i, i \in \mathbb{Z}^2\}$, where $\tilde{X}_i$ equals $L_j(X^k)$ and X^k is the corresponding $m \times m$ block of X. Note that since the loss function $l(\cdot)$ is bounded and m is finite, $E|\tilde{X}_0| < \infty$. It is not hard to see that conditions 1-3 are now satisfied (with the new space being $m \cdot \mathbb{Z}^2$). However, for $E\{\tilde{X}_0|\mathcal{I}_m\}$ to be deterministic, where $\mathcal{I}_m$ is the σ-algebra of m-invariant sets,

$$A \in \mathcal{I}_m \text{ iff } \tau_j(A) = A \quad \text{for all } j = i \cdot m, i \in \mathbb{Z}^2, \tag{42}$$

it is required that $\mathcal{I}_m$ is the trivial σ-algebra. In other words, block ergodicity of Q is required.

We now show that if the measure Q is strongly mixing, then it is block-ergodic for any finite block size. For $A, B \in \mathbb{Z}^2$, define

$$\alpha^Q(A, B) = \sup\{|Q(U \cap V) - Q(U)Q(V)|, U \in \sigma(X_A), V \in \sigma(X_B)\}, \tag{43}$$

where $\sigma(X_B)$ is the smallest sigma algebra generated by X_B. Let $\alpha^Q_{a,b}(k)$ denote the strong mixing coefficient [38, Sec. 1.7] of the random field Q

$$\alpha^Q_{a,b}(k) = \sup\{\alpha^Q(A, B), |A| \leq a, |B| \leq b, d(A, B) \geq k\}, \tag{44}$$

where d is a metric on $\mathbb{Z}^2$ and $d(A, B)$ is the distance between the closest points, i.e., $d(A, B) = \min_{i \in A, j \in B} d(i, j)$. Assume now that Q is strongly mixing in the sense that for all $a, b \in \mathbb{N} \cup \{\infty\}$, $\alpha^Q_{a,b}(k) \to 0$ as $k \to \infty$. It is easy to see that $Q(A) \in \{0,1\}$ for all $A \in \mathcal{I}_m$. Indeed,

$$\lim_{d(i,0) \to \infty} |Q(\tau_{i \cdot m}(A) \cap A) - Q(\tau_{i \cdot m}(A))Q(A)| = 0, \tag{45}$$

however, since A is m-invariant, $\tau_{i \cdot m}(A) = A$ and thus $Q(A) = Q(A)^2$. Hence Q is m-block ergodic for each m (i.e., totally ergodic).

The following theorem asserts that under the assumption that the random field Q is strongly mixing, the results of Theorem 4 apply in the a.s. sense as well.

---

[4]In fact, Tempelman's work [37] also includes slightly weaker conditions, but neither are satisfied in the current setting.

Theorem 9. *Let X be a stationary strongly mixing random field with a probability measure Q. Let $\mathcal{F} = \{\mathcal{F}_n\}$ be a sequence of finite sets of scandictors and assume that $U_{\mathcal{F}}(l, Q)$ exists. Then, if the universal algorithm suggested in the proof of Theorem 4 uses a fixed block size m, we have*

$$\liminf_{n \to \infty} \frac{1}{|V_n|} L_{alg}(X_{V_n}) \leq U_{\mathcal{F}}(l, Q) + \delta(m) \quad Q - a.s. \tag{46}$$

for any such Q and some $\delta(m)$ such that $\delta(m) \to 0$ as $m \to \infty$.

Proof. For each x_{V_n}, we have,

$$
\begin{aligned}
\frac{1}{|V_n|} L_{min}(x_{V_n}) &= \frac{1}{|V_n|} \min_{(\Psi, F)_j \in \mathcal{F}_m} \sum_{i=1}^{(K+1)^2} L_j(x^i) \\
&\leq \frac{1}{|V_n|} \left(\min_{(\Psi, F)_j \in \mathcal{F}_m} \sum_{i=1}^{K^2} L_j(x^i) + 2Km(n - Km)l_{max} + (n - Km)^2 l_{max} \right) \\
&\leq \frac{1}{|V_m|} \min_{(\Psi, F)_j \in \mathcal{F}_m} \frac{1}{K^2} \sum_{i=1}^{K^2} L_j(x^i) + 2\frac{m}{n} l_{max}.
\end{aligned}
\tag{47}
$$

By Proposition 5,

$$\frac{1}{|V_n|} \bar{L}_{alg}(x_{V_n}) \leq \frac{1}{|V_n|} L_{min}(x_{V_n}) + \frac{m(n+m)}{|V_n|} \sqrt{\log \lambda} \frac{l_{max}}{\sqrt{2}}. \tag{48}$$

Thus,

$$
\begin{aligned}
\liminf_{n \to \infty} \frac{1}{|V_n|} \bar{L}_{alg}(x_{V_n}) &\leq \liminf_{n \to \infty} \frac{1}{|V_n|} L_{min}(x_{V_n}) \\
&\leq \frac{1}{|V_m|} \liminf_{n \to \infty} \min_{(\Psi, F)_j \in \mathcal{F}_m} \frac{1}{K^2} \sum_{i=1}^{K^2} L_j(x^i) \\
&\leq \frac{1}{|V_m|} \min_{(\Psi, F)_j \in \mathcal{F}_m} \liminf_{n \to \infty} \frac{1}{K^2} \sum_{i=1}^{K^2} L_j(x^i).
\end{aligned}
\tag{49}
$$

Since $K \to \infty$ as $n \to \infty$, by the block ergodicity of Q and the fact that for finite m and each $(\Psi, F)_j \in \mathcal{F}_m$, $L_j(X)$ is a bounded function, it follows that

$$\liminf_{n \to \infty} \frac{1}{K^2} \sum_{i=1}^{K^2} L_j(x^i) = E_{Q_{V_m}} L_j(X^0) \quad Q - a.s. \tag{50}$$

Finally, since $U_{\mathcal{F}}(l, Q)$ exists, there exists $\delta(m)$ such that $\delta(m) \to 0$ as $m \to \infty$ and we have

$$\liminf_{n \to \infty} \frac{1}{|V_n|} \bar{L}_{alg}(x_{V_n}) \leq U_{\mathcal{F}}(l, Q) + \delta(m) \quad Q - a.s. \tag{51}$$

The fact that $\bar{L}_{alg}(x_{V_n})$ converges to $L_{alg}(x_{V_n})$ a.s. is clear from the proof of Proposition 6. $\quad \square$

Very similar to Theorem 9, we also have the following corollary.

Corollary 10. *Let X be a stationary strongly mixing random field over a finite alphabet A and a probability measure Q. Let the prediction space D be either finite or bounded (with $l(x, F)$ then being Lipschitz in its second argument). Then, there exists a sequence of scandictors $\{(\Psi, F)_n\}$, independent of Q, for which*

$$\liminf_{n \to \infty} \frac{1}{|V_n|} L_{alg}(X_{V_n}) \leq U(l, Q) + \delta(m) \quad Q - a.s. \tag{52}$$

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, $Q - a.s.$

3.6 Universal Scandiction for Individual Images

The proofs of Theorems 4 and 9 relied on the stationarity, or the stationarity and mixing property, of the random field X (respectively). In the proof of Theorem 4, we used the fact that the cumulative loss of any scandictor (Ψ, F) on a given block of data has the same expected value as that on any other block. In the proof of Theorem 9, on the other hand, the fact that the Cesaro mean of the losses on finite blocks converges to a single value, the expected cumulative loss, was used.

When x is an individual image, however, the cumulative loss of the suggested algorithm may be higher than that of the best scandictor in the scandictors set since restarting a scandictor at the beginning of each block may result in arbitrarily larger loss compared to the cumulative loss when the scandictor scans the entire data. Compared to the prediction problem, in the scandiction scenario, if the scanner is arbitrary, then different starting conditions may yield different scans (i.e., a different reordering of the data) and thus arbitrarily different cumulative loss, even if the predictor attached to it is very simple, e.g., a Markov predictor. It is expected, however, that when the scandictors have some structure, it will be possible to compete with finite sets of scandictors in the individual image scenario.

In this subsection, we suggest a basic scenario under which universal scandiction of individual images is possible. Further research in this area is required, though, in order to identify larger sets of scandictors under which universality is achievable. As mentioned earlier, since the exponential weighting algorithm used in the proofs of Theorems 4 and 9 applied only *block-wise* scandictors, i.e., scandictors which scan every block of the data separately from all other blocks, stationarity or stationarity and ergodicity *of the data* were required in order to prove its convergence. Here, since the data is an individual image, we impose restrictions on the families of scandictors in order to achieve meaningful results (this reasoning is analogous

to that described in [23, Section I-B] for the prediction problem). The first restriction is that the scanners with which we compete are such that the actual path taken by each scanner when it is applied in a block-wise order has some kind of an overlap (in a sense which will be defined later) with the path taken when it is applied to the whole image. The second restriction is that the predictors are Markovian of finite order (i.e., the prediction depends only on the last k symbols seen, for some finite k). Note that the first restriction does not restrict us to compete only with scandictors which operate in a block-wise order, only requires that the excess loss induced when the scandictors operate in a block-wise order, compared to operating on the entire image, is not too large, if, in addition, the predictor is Markovian.

The following definition, and the results which follow, make the above requirements precise. For two scanners Ψ and Ψ' for the data array x_B, define $N_{B,K}(x_B, \Psi, \Psi')$ as the number of sites in B such that their immediate past (context of unit length) under Ψ is contained in the context of length K under Ψ', namely,

$$N_{B,K}(x_B, \Psi, \Psi') = \left| \left\{ 1 \le i \le |B| : \exists_{1 \le j \le |B|, k \le K} \quad (\Psi_i, \Psi_{i-1}) = (\Psi'_j, \Psi'_{j-k}) \right\} \right|. \qquad (53)$$

Note that in the above definition, a "context" of size w for a site in B refers to the *set of w sites* which precede it in the discussed scan, and not their actual values. When $\{\Psi_n\}$ is a sequence of scanners, where Ψ_n is a scanner for V_n, it will be interesting to consider the number of sites in $B \subset V_{n_2}$, where B is an $n_1 \times n_1$ rectangle, $n_1 \le n_2$, such that their immediate past under Ψ_{n_2} (applied to V_{n_2}) is contained in the context of length K under Ψ_{n_1} (applied to B), that is

$$N_{B,K}(x_B, \Psi_{n_2}, \Psi_{n_1}) = \left| \left\{ 1 \le i \le |B| : \exists_{1 \le j \le |V_{n_2}|, k \le K} \quad (\Psi_{n_2, i}, \Psi_{n_2, i-1}) = (\Psi_{n_1, j}, \Psi_{n_1, j-k}) \right\} \right|, \qquad (54)$$

where $\Psi_{n.,i}$ is the i'th site the scanner $\Psi_{n.}$ visits. The following proposition is proved in Appendix A.3.

Proposition 11. *Consider two scanners Ψ and Ψ' for B such that for any individual image x_B we have*

$$\frac{N_{B,K}(x_B, \Psi, \Psi')}{|B|} = 1 - o(|B|). \qquad (55)$$

Then, for any x_B,

$$L_{(\Psi', F^{Kw, opt})}(x_B) \le L_{(\Psi, F^{w, opt})}(x_B) + o(|B|)(K+1)^{w-1} l_{max}, \qquad (56)$$

where for each scandictor $(\Psi, F^{w, opt})$, $F^{w, opt}$ denotes the optimal w-order Markov predictor for the scan Ψ.

Note that in order to satisfy the condition in (55) for *any* array x_B, it is likely (but not a compulsory) that both Ψ and Ψ' are data-independent scans. However, they need not be identical. If, for example, Ψ is a raster scan from left to right, and Ψ' applies the same left to right scan, but with a different ordering of the rows, then the condition is satisfied for any x_B.

The result of Proposition 11 yields the following corollary, which gives sufficient conditions on the scandictors sets under which a universal scandictor for any individual image exists. The proof can be found in Appendix A.4.

Corollary 12. *Let* $\{\mathcal{F}_n\}$, $|\mathcal{F}_n| = \lambda < \infty$, *be a sequence of scandictor sets, where* $\mathcal{F}_n = \{(\Psi_n^1, F^1), (\Psi_n^2, F^2), \ldots, (\Psi_n^\lambda, F^\lambda)\}$ *is a set of scandictors for* V_n. *Assume that the predictors are Markov of finite order* w, *the prediction space* D *is finite, and that there exists* $m(n) = o(n)$ *(yet* $m(n) \to \infty$ *as* $n \to \infty$*) such that for all* $1 \leq i \leq \lambda$, n, *and* x_{V_n} *we have*

$$\frac{N_{B_{m(n)},K}\left(x_{B_{m(n)}}, \Psi_n^i, \Psi_{m(n)}^i\right)}{m(n)^2} = 1 - o\left(m(n)^2\right), \tag{57}$$

where $B_{m(n)}$ *is any one of the* $\left\lfloor \frac{n}{m(n)} \right\rfloor^2$ *sub-blocks of size* $m(n) \times m(n)$ *of* V_n. *Then, there exists a sequence of scandictors* $\{(\hat{\Psi}, \hat{F})_n\}$ *such that for any image* x

$$\liminf_{n \to \infty} E \frac{1}{|V_n|} L_{(\hat{\Psi}, \hat{F})_n}(x_{V_n}) \leq \liminf_{n \to \infty} \min_{(\Psi,F) \in \mathcal{F}_n} \frac{1}{|V_n|} L_{(\Psi,F)}(x_{V_n}) \tag{58}$$

where the expectation in the l.h.s. of (58) is due to the possible randomization in $(\hat{\Psi}, \hat{F})_n$.

Although the condition in (57) is limiting, and may not be met by many data-dependent scans, Corollary 12 still answers on the affirmative the following basic question: do there exist scandictor sets for which one can find a universal scandictor in the individual image scenario? For example, by Corollary 12, if the scandictor set includes all raster-type scans (e.g., left-to-right, right-to-left, up-down, down-up, diagonal, etc.), accompanied with Markov predictors of finite order, then there exists a universal scandictor whose asymptotic normalized cumulative loss is less or equal than that of the best scandictor in the set, for any individual image x. The condition in (57) is also satisfied for some well-known "self-similar" space filling curves, such as the Sierpinski or Lebesgue curves [39].

4 Bounds on the Excess Scandiction Loss for Non-Optimal Scanners

While the results of Section 3 establish the existence of a universal scandictor for all stationary random fields and bounded loss function (under the terms of Theorem 8), it is interesting to

investigate, from both practical and theoretical reasons, what is the excess scandiction loss when non-optimal scanners are used. I.e., in this section we answer the following question: Suppose that, for practical reasons for example, one uses a non-optimal scanner, accompanied with the optimal predictor for that scan. How large is the excess loss incurred by this scheme with respect to optimal scandiction?

For the sake of simplicity, we consider the scenario of predicting the next outcome of a binary source, with $D = [0, 1]$ as the prediction space. Hence, $l : \{0, 1\} \times [0, 1] \to \mathbb{R}$ is the loss function. Furthermore, we assume deterministic scanner (though data-dependent, of course). The generalization to randomized scanners is cumbersome but straightforward.

Let ϕ_l denote the Bayes envelope associated with l, i.e.,

$$\phi_l(p) = \min_{q \in [0,1]} [(1 - p)l(0, q) + pl(1, q)]. \tag{59}$$

We further define

$$\epsilon_l = \min_{\alpha, \beta} \max_{0 \le p \le 1} |\alpha h_b(p) + \beta - \phi_l(p)|, \tag{60}$$

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, the expected instantaneous loss equals the conditional entropy, hence the expected cumulative loss coincides with the entropy, which is invariant to the scan, and we have $\epsilon_l = 0$. To wit, the scan is inconsequential under log loss.

Although the definitions of $\phi_l(p)$ and ϵ_l refer to the binary scenario, the results below (Theorem 13 and Propositions 14 and 15) hold for larger alphabets, with ϵ_l defined as in (60), with the maximum ranging over the simplex of all distributions on the alphabet, and $h(p)$ (replacing $h_b(p)$) and $\phi_l(p)$ denoting the entropy and Bayes envelope of the distribution p, respectively.

Let Ψ be any (possibly data dependent) scan, and let $E_{Q_B} \frac{1}{|B|} L_{(\Psi, F^{opt})}(X_B)$ denote the expected normalized cumulative loss in scandicting X_B with the scan Ψ and the optimal predictor for that scan, under the loss function l. Remembering that $U(l, Q_B)$ denotes the scandictability of X_B w.r.t the loss function l, namely, $U(l, Q_B) = \inf_{\Psi} E_{Q_B} \frac{1}{|B|} L_{(\Psi, F^{opt})}(X_B)$, our main result in this section is the following.

Theorem 13. *Let X_B be an arbitrarily distributed binary field. Then, for any scan Ψ,*

$$\left| E_{Q_B} \frac{1}{|B|} L_{(\Psi, F^{opt})}(X_B) - U(l, Q_B) \right| \le 2\epsilon_l. \tag{61}$$

That is, the excess loss incurred by applying *any scanner* Ψ, accompanied with the optimal predictor for that scan, with respect to optimal scandiction is not larger than $2\epsilon_l$.

To prove Theorem 13, we first introduce a prediction result (i.e., with no data reordering) on the error in estimating the cumulative loss of a predictor under a loss function l with the best affine function of the entropy. We then generalize this result to the multi-dimensional case.

Proposition 14. *Let X^n be an arbitrarily distributed binary n-tuple and let $EL_l^{opt}(X^n)$ denote the expected cumulative loss in predicting X^n with the optimal distribution-dependent scheme for the loss function l. Then,*

$$\left| \alpha_l \frac{1}{n} H(X^n) + \beta_l - \frac{1}{n} EL_l^{opt}(X^n) \right| \leq \epsilon_l, \tag{62}$$

where α_l and β_l are the achievers of the minimum in (60).

Proof. Let α_l and β_l be the achievers of the minimum in (60). We have,

$$\left| \alpha_l \frac{1}{n} H(X^n) + \beta_l - \frac{1}{n} EL_l^{opt}(X^n) \right|$$

$$= \left| \frac{1}{n} \sum_{t=1}^{n} \sum_{x^t} P(x^{t-1}) \right.$$

$$\left. \left[-\alpha_l P(x_t|x^{t-1}) \log P(x_t|x^{t-1}) + P(x_t|x^{t-1})\beta_l - P(x_t|x^{t-1})l(x_t, F_t^{opt}(x^{t-1})) \right] \right|$$

$$\stackrel{(a)}{=} \left| \frac{1}{n} \sum_{t=1}^{n} \sum_{x^{t-1}} P(x^{t-1}) \left[\alpha_l h_b(P(\cdot|x^{t-1})) + \beta_l - \phi_l(P(\cdot|x^{t-1})) \right] \right|$$

$$\leq \frac{1}{n} \sum_{t=1}^{n} \sum_{x^{t-1}} P(x^{t-1}) \left| \alpha_l h_b(P(\cdot|x^{t-1})) + \beta_l - \phi_l(P(\cdot|x^{t-1})) \right|$$

$$\leq \frac{1}{n} \sum_{t=1}^{n} \sum_{x^{t-1}} P(x^{t-1}) \max_p \left| \alpha_l h_b(p) + \beta_l - \phi_l(p) \right|$$

$$= \max_p \left| \alpha_l h_b(p) + \beta_l - \phi_l(p) \right|$$

$$= \epsilon_l, \tag{63}$$

where (a) is by the definition of $\phi_l(\cdot)$ and the optimality of F_t^{opt} with respect to l. $\qquad \square$

The following proposition is the generalization of Proposition 14 to the multi-dimensional case.

Proposition 15. *Let X_B be an arbitrarily distributed binary random field. Then, for any scan Ψ,*

$$\left| \alpha_l \frac{1}{|B|} H(X_B) + \beta_l - E_{Q_B} \frac{1}{|B|} L_{(\Psi, F^{opt})}(X_B) \right| \leq \epsilon_l, \tag{64}$$

where α_l and β_l are the achievers of the minimum in (60).

For data-independent scans, the proof follows the proof of Proposition 14 verbatim by applying it to the reordered $|B|$-tuple $X_{\Psi_1}, \ldots, X_{\Psi_{|B|}}$ and remembering that $H(X_B) = H(X_{\Psi_1}, \ldots, X_{\Psi_{|B|}})$. For data-dependent scans, the proof is similar, but requires more caution.

Proof of Proposition 15. Let α_l and β_l be the achievers of the minimum in (60). For a given data array x_B, $\Psi_1, \Psi_2(x_{\Psi_1}), \ldots, \Psi_{|B|}(x^{\Psi_{|B|-1}})$ are fixed, and merely reflect a reordering of x_B as a $|B|$-tuple. Thus,

$$\left| \alpha_l \frac{1}{|B|} H(X_B) + \beta_l - E_{Q_B} \frac{1}{|B|} L_{(\Psi, F^{opt})}(X_B) \right|$$

$$= \left| \frac{1}{|B|} \sum_{x_B} \left(-\alpha_l P(x_B) \log P(x_B) - P(x_B) \sum_{t=1}^{|B|} l(x_{\Psi_t}, F_t^{opt}(x^{\Psi_{t-1}})) \right) + \beta_l \right|$$

$$= \left| \frac{1}{|B|} \sum_{x_B} \left(-\alpha_l P(x_B) \sum_{t=1}^{|B|} \log P(x_{\Psi_t}|x^{\Psi_{t-1}}) - P(x_B) \sum_{t=1}^{|B|} l(x_{\Psi_t}, F_t^{opt}(x^{\Psi_{t-1}})) \right) + \beta_l \right|$$

$$= \left| \frac{1}{|B|} \sum_{t=1}^{|B|} \sum_{x_B} P(x_B) \left(-\alpha_l \log P(x_{\Psi_t}|x^{\Psi_{t-1}}) + \beta_l - l(x_{\Psi_t}, F_t^{opt}(x^{\Psi_{t-1}})) \right) \right|. \tag{65}$$

Fix $t = t_0$ in the sum over t. Consider all data arrays x_B such that for a specific scanner Ψ we have

$$\left\{ \Psi_1, \Psi_2(x_{\Psi_1}), \ldots, \Psi_{t_0-1}(x_{\Psi_1}, \ldots, x_{\Psi_{t_0-2}}) \right\} = I(\Psi), \tag{66}$$

where $I(\Psi) \subset B$ is a fixed set of sites, and $(x_{\Psi_1}, \ldots, x_{\Psi_{t_0-1}}) = \underline{a}$, for some $\underline{a} \in \{0,1\}^{t_0-1}$. In this case, $\Psi_t(x^{\Psi_{t_0-1}})$ is also fixed, and since the term in the parentheses of (65) depends only on I, $\underline{a}$ and $x_{\Psi_{t_0}}$, we have

$$\sum_{x^n} P(x^n) \left(-\alpha_l \log P(x_{\Psi_{t_0}}|x^{\Psi_{t_0-1}}) + \beta_l - l(x_{\Psi_{t_0}}, F_{t_0}^{opt}(x^{\Psi_{t_0-1}})) \right)$$

$$= \sum_{I,\underline{a}} P(x_I = \underline{a}) \sum_{x_{\Psi_{t_0}} \in \{0,1\}} P(x_{\Psi_{t_0}}|x_I = \underline{a}) \left(-\alpha_l \log P(x_{\Psi_{t_0}}|x_I = \underline{a}) + \beta_l - l(x_{\Psi_{t_0}}, F_{t_0}^{opt}(\underline{a})) \right).$$

$$\tag{67}$$

Consequently,

$$\left| \alpha_l \frac{1}{|B|} H(X_B) + \beta_l - E_{Q_B} \frac{1}{|B|} L_{(\Psi, F^{opt})}(X_B) \right|$$

$$\leq \frac{1}{|B|} \sum_{t=1}^{|B|} \sum_{I, \underline{a}} P(x_I = \underline{a})$$

$$\left| \sum_{x_{\Psi_t} \in \{0,1\}} P(x_{\Psi_t} | x_I = \underline{a}) \Big(-\alpha_l \log P(x_{\Psi_t} | x_I = \underline{a}) + \beta_l - l(x_{\Psi_t}, F_t^{opt}(\underline{a})) \Big) \right|$$

$$= \frac{1}{|B|} \sum_{t=1}^{|B|} \sum_{I, \underline{a}} P(x_I = \underline{a}) \left| \alpha_l h_b(P(\cdot | x_I = \underline{a})) + \beta_l - \phi_l(P(\cdot | x_I = \underline{a})) \right|$$

$$\leq \frac{1}{|B|} \sum_{t=1}^{|B|} \sum_{I, \underline{a}} P(x_I = \underline{a}) \max_p |\alpha_l h_b(p) + \beta_l - \phi_l(p)|$$

$$= \max_p |\alpha_l h_b(p) + \beta_l - \phi_l(p)|$$

$$= \epsilon_l. \tag{68}$$

$\square$

It is now easy to see why Theorem 13 holds.

Proof of Theorem 13. The proof is a direct application of Proposition 15, as for any scan Ψ,

$$\left| E_{Q_B} \frac{1}{|B|} L_{(\Psi, F^{opt})}(X_B) - U(l, Q_B) \right|$$

$$\leq \left| \frac{\alpha_l}{|B|} H(X_B) + \beta_l - E_{Q_B} \frac{1}{|B|} L_{(\Psi, F^{opt})}(X_B) \right| + \left| \frac{\alpha_l}{|B|} H(X_B) + \beta_l - U(l, Q_B) \right|$$

$$\leq 2\epsilon_l. \tag{69}$$

$\square$

At this point, a few remarks are in order. For the bound in Theorem 13 to be tight, the following conditions should be met. First, equality is required in (64) for both the scan Ψ and the optimal scan (which achieves $U(l, Q_B)$). It is not hard to see that for a given scan Ψ, equality in (64) is achieved if and only if $P(\cdot | x^{\Psi_{t-1}}) = p$ for all $x^{\Psi_{t-1}}$, where p is a maximizer of (60). However, for (61) to be tight, it is also required that

$$\frac{\alpha_l}{|B|} H(X_B) + \beta_l - E_{Q_B} \frac{1}{|B|} L_{(\Psi, F^{opt})}(X_B) = -\frac{\alpha_l}{|B|} H(X_B) - \beta_l + U(l, Q_B), \tag{70}$$

so the triangle inequality is held with equality. Namely, it is required that under the scan Ψ, for example, $P(\cdot | x^{\Psi_{t-1}}) = p$ for all $x^{\Psi_{t-1}}$, where p is such that $\alpha_l h_b(p) + \beta_l - \phi_l(p) = \epsilon_l$, yet under the optimal scan, say Ψ', $P(\cdot | x^{\Psi'_{t-1}}) = p'$ for all $x^{\Psi'_{t-1}}$, where p' is such that

$\alpha_l h_b(p') + \beta_l - \phi_l(p') = -\epsilon_l$. Clearly this is not always the case, and thus, generally, the bound in Theorem 13 is not tight. Indeed, although under a different setting (individual images), in subsection 4.1 we derive a tighter upper bound on the excess loss for the specific case of Hamming loss. Using this bound, it is easy to see that the 0.16 bound given here (as $\epsilon_l = 0.08$ for Hamming loss) is only a worst case, and typically much tighter bounds on the excess loss apply, depending on the image compressibility. For example, consider a 1st order symmetric Markov chain with transition probability 1/4. Scanning this source in the trivial (sequential) order results in an error rate of 1/4. By [22], this is indeed the optimal scanning order for this source, as it can be represented as an autoregressive process whose innovation process has a maximum entropy distribution with respect to the Hamming distance. The "odds-then-evens" scan[5], however, which was proved useful for this source but with larger transition probabilities (larger than 1/2, [22]), results in an error rate of 5/16, which is 1/16 away from the optimum. It is not hard to show that different transition probabilities result in lower excess loss.

4.1 Individual Images and the Peano-Hilbert Scan

In this subsection, we seek analogous results for the individual image scenario. Namely, the data array x_B has no stochastic model. A scandictor (Ψ, F), in this case, wishes to minimize the cumulative loss over x_{V_n}, that is, $L_{(\Psi,F)}(x_{V_n})$ as defined in (3).

In this setting, although one can easily define an empirical probability measure, the invariance of the entropy $H(X^n)$ to the reordering of the components, which stood at the heart of Theorem 13, does not hold for *any* reordering (scan) and any finite n. Thus, we limit the possible set of scanners to that of the finite state machines discussed earlier. Moreover, in the sequel, we do not bound the difference in the scandiction losses of any two scandictors from that set, only that between the Peano-Hilbert scan (which is asymptotically optimal for compression of individual images [14]) and any other finite state scanner (both accompanied with an optimal Markov predictor), or between two scans (finite state or not) for which the FS compressibility of the resulting sequence is the same.

We start with several definitions. 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 $s \in \{0,1\}^{k+1}$

---

[5]An "odds-then-evens" scanner for a one-dimensional vector x_1^n, first scans all the sites with an odd index, in an ascending order, then all the sites with an even index.

define the empirical distribution of order $k + 1$ as

$$\hat{P}_{\Psi_B}^{k+1}(s) = \frac{1}{|B| - k} \left| \{ k < i \le |B| : x_{i-k}^i = s \} \right|. \tag{71}$$

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

$$\hat{P}_{\Psi_B}^{k+1}(s') = \hat{P}_{\Psi_B}^{k+1}([s',0]) + \hat{P}_{\Psi_B}^{k+1}([s',1]) \tag{72}$$

and

$$\hat{P}_{\Psi_B}^{k+1}(x|s') = \frac{\hat{P}_{\Psi_B}^{k+1}([s',x])}{\hat{P}_{\Psi_B}^{k+1}(s')}, \tag{73}$$

where $0/0$ is defined as $1/2$ and $[\cdot, \cdot]$ denotes string concatenation.[6] Let $\hat{H}_{\Psi_B}^{k+1}(X|X^k)$ be the empirical conditional entropy of order k, i.e.,

$$\hat{H}_{\Psi_B}^{k+1}(X|X^k) = - \sum_{s \in \{0,1\}^k} \hat{P}_{\Psi_B}^{k+1}(s) \sum_{x \in \{0,1\}} \hat{P}_{\Psi_B}^{k+1}(x|s) \log \hat{P}_{\Psi_B}^{k+1}(x|s). \tag{74}$$

Finally, denote by $F^{k,opt}$ the optimal k-th order Markov predictor, in the sense that it minimizes the expected loss with respect to $\hat{P}_{\Psi_B}^{k+1}(\cdot|\cdot)$ and $x_1^{|B|}$. The following proposition is the individual image analogue of Proposition 15.

Proposition 16. *Let x_B be any data array. Let $\frac{1}{|B|} L_{(\Psi_B, F^{k,opt})}(x_B)$ denote the normalized cumulative loss of the scandictor $(\Psi_B, F^{k,opt})$, where Ψ_B is any (data dependent) scan and $F^{k,opt}$ is the optimal k-th order Markov predictor with respect to Ψ_B and l. Then,*

$$\left| \alpha_l \hat{H}_{\Psi_B}^{k+1}(X|X^k) + \beta_l - \frac{1}{|B|} L_{(\Psi_B, F^{k,opt})}(x_B) \right| \le \epsilon_l + \frac{k l_{max}}{|B|}, \tag{75}$$

where α_l and β_l are the achievers of the minimum in (60).

Since x_B is an individual image, $x_1^{|B|} = \Psi_B(x_B)$ is fixed. In that sense, the proof resembles that of Proposition 14 and we write x_t for the value of x at the t-th site Ψ_B visits. On the other hand, since the order of the predictor, k, is fixed, we can use $\hat{P}_{\Psi_B}^{k+1}(\cdot)$ and avoid the summation over the time index t. The complete details can be found in Appendix A.5.

The bound in Proposition 16 differs from the one in Proposition 15 for two reasons. First, it is only asymptotic due to the $O(k/|B|)$ term. Second, the empirical entropy $\hat{H}_{\Psi_B}^{k+1}(X|X^k)$ is not invariant to the scanning order. This is a profound difference between the random and

---

[6]Note that defining $\hat{P}_{\Psi_B}^{k+1}(x|s')$, $s' \in \{0,1\}^k$ as $\frac{\hat{P}_{\Psi_B}^{k+1}([s',x])}{\hat{P}_{\Psi_B}^k(s')}$ is not consistent since generally $\hat{P}_{\Psi_B}^k(s') \ne \hat{P}_{\Psi_B}^{k+1}([s',0]) + \hat{P}_{\Psi_B}^{k+1}([s',1])$.

the individual settings, and, in fact, is at the heart of [14]. In the random setting, the chain rule for entropies implies invariance of the entropy rate to the scanning order. This fact does not hold for a k-th order empirical distribution of an individual image, hence the usage of the Peano-Hilbert scanning order.[7] Consequently, we cannot directly compare between any two scans. Nevertheless, Proposition 16 has the following two interesting applications, given by Proposition 17 and Corollary 18.

For $\Psi = \{\Psi_n\}$, where Ψ_n is a scan for V_n, and an infinite individual image x define

$$L_\Psi^k(x) = \limsup_{n \to \infty} \frac{1}{|V_n|} L_{(\Psi_n, F^{k,opt})}(x_{V_n}) \tag{76}$$

and

$$L_\Psi(x) = \lim_{k \to \infty} L_\Psi^k(x). \tag{77}$$

Proposition 17 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.

Proposition 17. *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 $l : \{0, 1\} \times [0, 1] \to \mathbb{R}$,*

$$L_{PH}(x) \le L_\Psi(x) + 2\epsilon_l. \tag{78}$$

Before we prove Proposition 17, define the asymptotic k-th order empirical conditional entropy under $\{\Psi_n\}$ as

$$\hat{H}_\Psi^{k+1}(x) = \limsup_{n \to \infty} \hat{H}_{\Psi_n}^{k+1}(X|X^k) \tag{79}$$

and further define

$$\hat{H}_\Psi(x) = \lim_{k \to \infty} \hat{H}_\Psi^{k+1}(x). \tag{80}$$

The existence of $\hat{H}_\Psi(x)$ is established later in the proof of Proposition 17, where it is also shown that this limit equals $\lim_{k \to \infty} \limsup_{n \to \infty} \frac{1}{k} \hat{H}_{\Psi_n}^k(X^k)$. By [15, Theorem 3], the latter limit is no other than the asymptotic finite state compressibility of x under the sequence of scans Ψ, namely,

$$\begin{aligned} \lim_{k \to \infty} \limsup_{n \to \infty} \frac{1}{k} \hat{H}_{\Psi_n}^k(X^k) &= \rho(\Psi(x)) \\ &= \lim_{s \to \infty} \limsup_{n \to \infty} \rho_{E(s)}(\Psi_n(x_{V_n})), \end{aligned} \tag{81}$$

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 [15, eq. (1)-(4)]. We may now introduce the following corollary.

---

[7]Yet, the Peano-Hilbert is by no means the *only* optimal scan. We Elaborate on this issue later in this section.

Corollary 18. *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,*

$$|L_{\Psi_1}(x) - L_{\Psi_2}(x)| \leq 2\epsilon_l. \tag{82}$$

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

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 [15] 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 [15] 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 $H_{\Psi}(x) = \hat{H}_{\Psi'}(x)$.

Proof of Proposition 17. For each n, Ψ_n is a scanner for V_n. Thus, by Proposition 16, we have

$$\left| \alpha_l \hat{H}_{\Psi_n}^{k+1}(X|X^k) + \beta_l - \frac{1}{|V_n|} L_{(\Psi_n, F^{k,opt})}(x_{V_n}) \right| \leq \epsilon_l + \frac{kl_{max}}{|V_n|}, \tag{83}$$

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

$$\left| \alpha_l \limsup_{n \rightarrow \infty} \hat{H}_{\Psi_n}^{k+1}(X|X^k) + \beta_l - L_{\Psi}^k(x) \right| \leq \epsilon_l. \tag{84}$$

For a stationary source, it is well known (e.g., [40, Theorem 4.2.1]) that $\lim_{k \rightarrow \infty} H(X_k|X_1^{k-1})$ exists and in fact

$$\lim_{k \rightarrow \infty} H(X_k|X_1^{k-1}) = \lim_{k \rightarrow \infty} \frac{1}{k} H(X_1^k). \tag{85}$$

To this end, we show that the same holds for empirical entropies. We start by showing that $\limsup_{n \rightarrow \infty} \hat{H}_{\Psi_n}^{k+1}(X|X^k)$ is a decreasing sequence in k. Since conditioning reduces the entropy, it is clear that $\hat{H}_{\Psi_n}^{k+1}(X|X^k) \leq \hat{H}_{\Psi_n}^{k+1}(X|X^{k-1})$, where both are calculated using $\hat{P}_{\Psi_n}^{k+1}(\cdot)$. However, the above may not be true when $\hat{H}_{\Psi_n}^{k+1}(X|X^{k-1})$ is replaced by $\hat{H}_{\Psi_n}^k(X|X^{k-1})$, as the later is calculated using $\hat{P}_{\Psi_n}^k(\cdot)$. 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

$$\hat{P}_{\Psi_n}^{k+1}(s) - \frac{k+1-j}{|V_n| - k} \leq \hat{P}_{\Psi_n}^j(s) \leq \hat{P}_{\Psi_n}^{k+1}(s) + \frac{k+1-j}{|V_n| - k}. \tag{86}$$

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

$$
\begin{aligned}
\limsup_{n\to\infty} \hat{H}_{\Psi_n}^{k+1}(X|X^k) &\leq \limsup_{n\to\infty} \hat{H}_{\Psi_n}^{k+1}(X|X^{k-1}) \\
&= \limsup_{n\to\infty} \hat{H}_{\Psi_n}^{k}(X|X^{k-1}),
\end{aligned}
\tag{87}
$$

hence $\limsup_{n\to\infty} \hat{H}_{\Psi_n}^{k}(X|X^{k-1})$ is decreasing in k. Since it is a non negative sequence, $\hat{H}_\Psi(x)$ as defined in (80) exists and we have

$$
\left| \alpha_l \hat{H}_\Psi(x) + \beta_l - L_\Psi(x) \right| \leq \epsilon_l.
\tag{88}
$$

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 [14] can be applied. The method is similar to that in [40, Theorem 4.2.1]), with an adequate handling of empirical entropies. By (86),

$$
\begin{aligned}
\limsup_{n\to\infty} \frac{1}{k} \hat{H}_{\Psi_n}^{k}(X^k) &= \limsup_{n\to\infty} \frac{1}{k} \sum_{i=1}^{k} \hat{H}_{\Psi_n}^{k}(X_i|X_1^{i-1}) \\
&= \limsup_{n\to\infty} \frac{1}{k} \sum_{i=1}^{k} \hat{H}_{\Psi_n}^{i}(X_i|X_1^{i-1}).
\end{aligned}
\tag{89}
$$

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

$$
\begin{aligned}
\hat{H}_\Psi(x) &= \lim_{k\to\infty} \limsup_{n\to\infty} \frac{1}{k} \hat{H}_{\Psi_n}^{k}(X^k) \\
&= \rho(\Psi(x)).
\end{aligned}
\tag{90}
$$

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 [14, eq. (4)]. For any other sequence of finite state scans $\tilde{\Psi}$ we have

$$
\begin{aligned}
\hat{H}_{PH}(x) &\leq \rho(x) \\
&\leq \hat{H}_{\tilde{\Psi}}(x),
\end{aligned}
\tag{91}
$$

where the first inequality is by [14, eq. (9) and (16)] and the second is straightforward from the definition of $\rho(x)$. Finally,

$$
\begin{aligned}
L_{PH}(x) &\overset{(a)}{\leq} \epsilon_l + \beta_l + \alpha_l \hat{H}_{PH}(x) \\
&\leq \epsilon_l + \beta_l + \alpha_l \hat{H}_{\tilde{\Psi}}(x) \\
&\overset{(b)}{\leq} 2\epsilon_l + L_{\tilde{\Psi}}(x),
\end{aligned}
\tag{92}
$$

where (a) and (b) result from the application of (88) to the sequences PH and $\tilde{\Psi}$ respectively.

$\square$

The proof of Corollary 18 is straightforward, using (88) for both Ψ_1 and Ψ_2 and the triangle inequality.

4.1.1 Hamming Loss

The bound in Proposition 17 is valid for any loss function $l : \{0,1\} \times [0,1] \to \mathbb{R}$. When l is the Hamming loss, the resulting bound is

$$L_{PH}^{Hamming}(x) \leq L_{\Psi}^{Hamming}(x) + 0.16, \tag{93}$$

for any other finite state sequence of scans, namely, a uniform bound, regardless of the compressibility of x. However, using known bounds on the predictability of a sequence (under Hamming loss) in terms of its compressibility can yield a tighter bound.

In [41], 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

$$\begin{aligned}
\mu(g, x_1^n) &\leq \frac{1}{2}\rho(g, x_1^n), \\
\mu(g, x_1^n) &\geq h_b^{-1}\left(\rho(g, x_1^n)\right),
\end{aligned} \tag{94}$$

where $\mu(g, \cdot)$ $(\rho(g, \cdot))$ is the best possible prediction (compression) performance when the next state function is g. Consequently, for *any* two finite-state scans Ψ_n^1 and Ψ_n^2 for x_{V_n},

$$\begin{aligned}
&\min_{g \in G_s} \mu(g, \Psi_n^1(x_{V_n})) - \min_{g \in G_s} \mu(g, \Psi_n^2(x_{V_n})) \\
&\qquad \leq \min_{g \in G_s} \frac{1}{2}\rho(g, \Psi_n^1(x_{V_n})) - \min_{g \in G_s} h_b^{-1}\left(\rho(g, \Psi_n^2(x_{V_n}))\right) \\
&\qquad = \frac{1}{2}\min_{g \in G_s} \rho(g, \Psi_n^1(x_{V_n})) - h_b^{-1}\left(\min_{g \in G_s} \rho(g, \Psi_n^2(x_{V_n}))\right).
\end{aligned} \tag{95}$$

Taking Ψ_n^1 to be the Peano-Hilbert scan, the results of [14] imply that

$$\min_{g \in G_s} \rho(g, \Psi_{PH}(x_{V_n})) \leq \min_{g \in G_s} \rho(g, \Psi_n(x_{V_n})) + \epsilon_{n,s} \tag{96}$$

for any finite-state scan Ψ_n, where $\epsilon_{n,s}$ satisfies $\lim_{s \to \infty} \limsup_{n \to \infty} \epsilon_{n,s} = 0$. Hence,

$$\begin{aligned}
&\min_{g \in G_s} \mu(g, \Psi_{PH}(x_{V_n})) - \min_{g \in G_s} \mu(g, \Psi(x_{V_n})) \\
&\qquad \leq \frac{1}{2}\min_{g \in G_s} \rho(g, \Psi_{PH}(x_{V_n})) - h_b^{-1}\left(\min_{g \in G_s} \rho(g, \Psi_{PH}(x_{V_n})) - \epsilon_{n,s}\right). \tag{97}
\end{aligned}$$

Taking the limits $\limsup_{n \to \infty}$ and then $s \to \infty$ implies the following proposition.

Proposition 19. *Let x be any individual image. Let PH denote the Peano-Hilbert sequence of scans. Then, under the Hamming loss function, for any sequence of finite state scans Ψ we have*

$$L_{PH}(x) \le L_\Psi(x) + \frac{1}{2}\rho(x) - h_b^{-1}(\rho(x)), \tag{98}$$

where $\rho(x)$ is the compressibility of the individual image x.

In other words, the specific scandictor composed of the Peano-Hilbert scan followed by the optimal predictor, adheres to the same asymptotic bounds (on predictability in terms of the compressibility) as the *best* finite-state scandictor. Figure 2 plots the function $\frac{1}{2}\rho - h_b^{-1}(\rho)$. The maximum possible loss is 0.16, similar to the bound given in Proposition 17, 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.
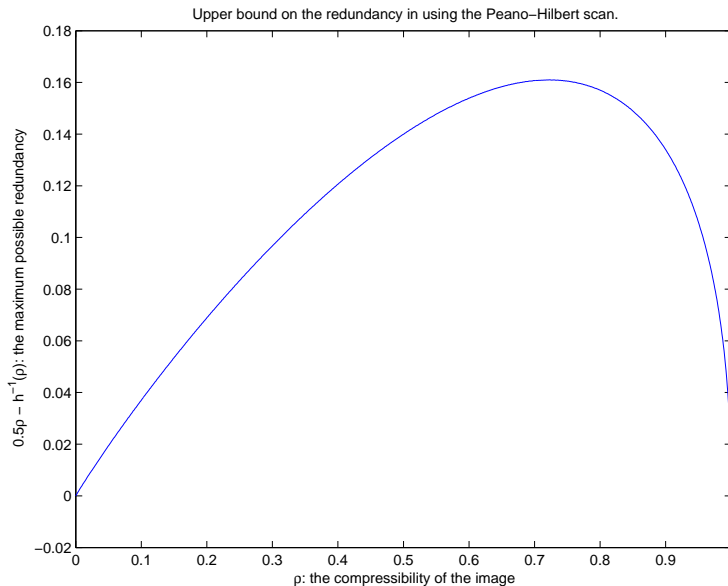

Figure 2: A plot of $\frac{1}{2}\rho - h^{-1}(\rho)$. The maximum redundancy is not higher than 0.16 in worst case, but will be much lower for more compressible arrays.

5   Conclusion

In this paper, we formally defined finite set scandictability, and showed that there exists a universal algorithm which successfully competes with any finite set of scandictors when the random field is stationary. Moreover, the existence of a universal algorithm which achieves the

scandictability of any spatially stationary random field was established. We then considered the scenario where non-optimal scanners are used, and derived a bound on the excess loss in that case, compared to optimal scandiction.

It is clear that the scandiction problem is even more intricate than its prediction analogue. For instance, very basic results in the prediction scenario do not apply to the scandiction case in a straightforward way, and, in fact, are still open problems. To name a few, consider the case of universal scandiction of individual images, briefly discussed in Section 3.6. Although the question whether there exists a universal scandictor which competes successfully with any finite set of scandictors on any individual image was answered negatively in Section 3.1, it is interesting to discover interesting sets of scandictors for which universal scandiction *is possible*. The sequential prediction literature also includes an elegant result [41] on the asymptotic equivalence between finite state and Markov predictors. We conjecture that this equivalence does not hold in the multi-dimensional scenario for any individual image. Finally, the very basic problem of determining the optimal scandictor for a given random field X with a *known probability measure Q*, is still unsolved in the general case.

It is also interesting to consider the problems of scanning and prediction, as well as filtering, in a noisy environment. These problems are intimately related to various problems in communications and image processing, such as filtering and denoising of images and video. As mentioned in Section 1, these problems are the subject of [27].

A Appendixes

A.1 Proof of Proposition 5

For the sake of simplicity, we suppress the dependence of $m(n)$ in n. Define $W_i = \sum_{j=1}^{\lambda} e^{-\eta L_{j,i}}$. We have

$$
\begin{aligned}
\log \frac{W_{(K+1)^2}}{W_0} &= \log \sum_{j=1}^{\lambda} e^{-\eta L_{j,(K+1)^2}} - \log \lambda \\
&\geq \log \max_j e^{-\eta L_{j,(K+1)^2}} - \log \lambda \\
&= -\eta \min_j L_{j,(K+1)^2} - \log \lambda \\
&= -\eta L_{min} - \log \lambda.
\end{aligned}
\tag{99}
$$

Moreover,

$$
\begin{aligned}
\log \frac{W_{i+1}}{W_i} &= \log \frac{\sum_{j=1}^{\lambda} e^{-\eta(L_{j,i}+L_j(x^i))}}{\sum_{j'=1}^{\lambda} e^{-\eta L_{j',i}}} \\
&= \log \sum_{j=1}^{\lambda} P_i\left(j|\{L_{j,i}\}_{j=1}^{\lambda}\right) e^{-\eta L_j(x^i)} \\
&\leq -\eta \sum_{j=1}^{\lambda} P_i\left(j|\{L_{j,i}\}_{j=1}^{\lambda}\right) L_j(x^i) + \frac{m^4 l_{max}^2 \eta^2}{8},
\end{aligned}
\tag{100}
$$

where the last inequality follows from the extension to Hoeffding's inequality given in [33] and the fact that $-\eta L_j(x^i)$ is in the range $[-\eta m^2 l_{max}, 0]$. Thus,

$$
\begin{aligned}
\log \frac{W_{(K+1)^2}}{W_0} &= \sum_{i=0}^{(K+1)^2-1} \log \frac{W_{i+1}}{W_i} \\
&\leq -\eta \sum_{i=0}^{(K+1)^2-1} \sum_{j=1}^{\lambda} P_i\left(j|\{L_{j,i}\}_{j=1}^{\lambda}\right) L_j(x^i) + \frac{m^4 l_{max}^2 \eta^2 (K+1)^2}{8} \\
&= -\eta \bar{L}_{alg} + \frac{m^4 l_{max}^2 \eta^2 (K+1)^2}{8}.
\end{aligned}
\tag{101}
$$

Finally, from (99) and (101), we have

$$
\begin{aligned}
\bar{L}_{alg} - L_{min} &\leq \frac{\log \lambda}{\eta} + \frac{m^4 l_{max}^2 \eta (K+1)^2}{8} \\
&\leq \frac{\log \lambda}{\eta} + \frac{m^2 l_{max}^2 \eta (n+m)^2}{8}.
\end{aligned}
\tag{102}
$$

The bound in (14) easily follows after optimizing the right hand side of (102) with respect to η.

A.2 Proof of Proposition 6

Let $\delta(n)$ be some sequence satisfying $\delta(n) \to 0$ as $n \to \infty$. Define the sets

$$
A_n = \left\{ \omega : \frac{L_{alg}(x_{V_n}) - L_{min}(x_{V_n})}{n^2} > \delta(n^2) \right\},
\tag{103}
$$

where (Ω, P) is the probability space. We wish to show that

$$
P\left(\limsup_{n \to \infty} A_n\right) = 0,
\tag{104}
$$

that is, $P(A_n \text{ i.o.}) = 0$. Let $(\Psi, F)_k$ be the scandictor chosen by the algorithm for the $k+1$ block, x^k. Define

$$
Z_k = L_{(\Psi,F)_k}(x^k) - E\left\{ L_{(\Psi,F)_k}(x^k)|\{L_{j,k}\}_{j=1}^{\lambda} \right\},
\tag{105}
$$

where the expectation is with respect to $P_k\left(j|\{L_{j,k}\}_{j=1}^{\lambda}\right)$. Namely, the actual randomization in Z_k is in the *choice* of $(\Psi,F)_k$. Thus, $\{Z_k\}$ are clearly independent, and adhere to the following Chernoff-like bound [33, eq. 33]

$$P\left(\sum_{k=1}^{(K+1)^2} Z_k \geq (K+1)^2\epsilon\right) \leq \exp\left\{-\frac{2(K+1)^2\epsilon^2}{(m^2 l_{max})^2}\right\} \tag{106}$$

for any $\epsilon > 0$. Note that

$$\sum_{k=1}^{(K+1)^2} Z_k = L_{alg}(x_{V_n}) - \bar{L}_{alg}(x_{V_n}), \tag{107}$$

thus, together with eq. (14), we have

$$P\left(L_{alg}(x_{V_n}) - L_{min}(x_{V_n}) \geq (K+1)^2\epsilon + m(n+m)\sqrt{\log\lambda}\frac{l_{max}}{\sqrt{2}}\right)$$
$$\leq \exp\left\{-\frac{2(K+1)^2\epsilon^2}{(m^2 l_{max})^2}\right\}. \tag{108}$$

Set

$$\delta(n) = \frac{(K+1)^2\epsilon + m(n+m)\sqrt{\log\lambda}\frac{l_{max}}{\sqrt{2}}}{n^2}. \tag{109}$$

Clearly $\delta(n) \to 0$ as $n \to \infty$ for any $m(n) = o(n)$ satisfying $m(n) \to \infty$. For the summability of the r.h.s. of (108) we further require that $m(n) = o\left(n^{1/3}\right)$. The proposition then follows directly by applying the Borel-Cantelli lemma.

A.3 Proof of Proposition 11

We show, by induction on w, that the number of sites in B for which the context of size w (in terms of sites in B) under the scan Ψ is not contained in the context of size Kw under the scan Ψ' is at most $o(|B|)(K+1)^{w-1}$. This proves the proposition, as the cumulative loss of $(\Psi',F^{Kw,opt})$ is no larger than $o(|B|)(K+1)^{w-1}l_{max}$ on these sites, and is at least as small as that of $(\Psi,F^{w,opt})$ on all the rest $|B| - o(|B|)(K+1)^{w-1}$ sites.

For $w = 1$ this is indeed so, by our assumption on Ψ and Ψ' - i.e., (55). We say that a site in B satisfies the context-condition with length $w - 1$ if its context of size $i - 1$, $1 < i \leq w$, under the scan Ψ is contained in its context of size $K(i-1)$ under the scan Ψ'. Assume that the number of sites in B which do not satisfy the context-condition with length $w - 1$ is at most $o(|B|)(K+1)^{w-2}$. We wish to lower bound the number of sites in B for which the context-condition with length w is satisfied. A sufficient condition is that the context-condition with length $w - 1$ is satisfied for both the site itself and its immediate past under Ψ. If the context-condition with length $w - 1$ is satisfied for a site, its immediate past under

Ψ is contained in its past of length K under Ψ'. Thus, if the context-condition of length $w-1$ is satisfied for a given site, and for all K preceding sites under Ψ', then it is also satisfied for length w. In other words, each site in B which does not satisfy the context-condition with length $w-1$ results in at most $K+1$ sites (itself and K more sites) which do not satisfy the context-condition with length w. Hence, if our inductive assumption is satisfied for $w-1$, the number of sites in B which do not satisfy the context-condition with length w is at most $o(|B|)(K+1)^{w-2}(K+1)$, which completes the proof.

A.4 Proof of Proposition 12

The proof is a direct application of Propositions 5 and 11. For each n, define the scandictors set

$$
\begin{aligned}
\tilde{\mathcal{F}}_n \;=\; \Big\{ &(\Psi_n^1, F^{Kw,1}), (\Psi_n^1, F^{Kw,2}), \dots, (\Psi_n^1, F^{Kw,|D|^{|A|^{Kw}}}), \dots \\
&(\Psi_n^2, F^{Kw,1}), (\Psi_n^2, F^{Kw,2}), \dots, (\Psi_n^2, F^{Kw,|D|^{|A|^{Kw}}}), \dots \\
&\dots \\
&(\Psi_n^\lambda, F^{Kw,1}), (\Psi_n^\lambda, F^{Kw,2}), \dots, (\Psi_n^\lambda, F^{Kw,|D|^{|A|^{Kw}}}) \Big\},
\end{aligned}
\tag{110}
$$

where $\{F^{Kw,i}\}_{i=1}^{|D|^{|A|^{Kw}}}$ is the set of all Markov predictors of order Kw.[8] Applying the results of Proposition 5 to $\{\tilde{\mathcal{F}}_n\}$, we have, for any image x and all n,

$$
EL_{(\hat\Psi,\hat F)_n}(x_{V_n}) - \min_{(\Psi,F)\in\tilde{\mathcal{F}}_{m(n)}} L_{(\Psi,F)}(x_{V_n}) \le m(n)\,(n+m(n))\,\sqrt{\log \lambda |D|^{|A|^{Kw}}}\,\frac{l_{max}}{\sqrt{2}},
\tag{111}
$$

where $\min_{(\Psi,F)\in\mathcal{F}_{m(n)}} L_{(\Psi,F)}(x_{V_n})$ is the cumulative loss of the best scandictor in $\mathcal{F}_{m(n)}$ operating block-wise on x_{V_n}. However, by Proposition 11, for any $1\le i\le \lambda$, x and n,

$$
\min_{1\le j\le |D|^{|A|^{Kw}}} EL_{(\Psi_{m(n)}^i, F^{Kw,j})}(x_{V_n}) \le EL_{(\Psi_n^i, F^i)}(x_{V_n}) + o\left(m(n)^2\right)(K+1)^{w-1} l_{max} \left\lfloor \frac{n}{m(n)} \right\rfloor^2.
\tag{112}
$$

Note that

$$
\begin{aligned}
\min_{(\Psi,F)\in\tilde{\mathcal{F}}_{m(n)}} L_{(\Psi,F)}(x_{V_n}) \;=\;& \min_{1\le i\le \lambda}\; \min_{1\le j\le |D|^{|A|^{Kw}}} EL_{(\Psi_{m(n)}^i, F^{Kw,j})}(x_{V_n}) \\
\le\;& \min_{1\le i\le \lambda} \left\{ EL_{(\Psi_n^i, F^i)}(x_{V_n}) + o\left(m(n)^2\right)(K+1)^{w-1} l_{max} \left\lfloor \frac{n}{m(n)} \right\rfloor^2 \right\} \\
=\;& \min_{(\Psi,F)\in\mathcal{F}_n} EL_{(\Psi,F)}(x_{V_n}) + o\left(m(n)^2\right)(K+1)^{w-1} l_{max} \left\lfloor \frac{n}{m(n)} \right\rfloor^2.
\end{aligned}
\tag{113}
$$

---

[8] Alternatively, one can use *one* universal predictor which competes successfully with all the Markov predictors of that order.

Thus, together with (111), we have

$$EL_{(\hat{\Psi},\hat{F})_n}(x_{V_n}) - \min_{(\Psi,F)\in\mathcal{F}_n} EL_{(\Psi,F)}(x_{V_n})$$

$$\leq m(n)\left(n+m(n)\right)\sqrt{\log\lambda|D|^{|A|Kw}}\frac{l_{max}}{\sqrt{2}} + o\left(m(n)^2\right)(K+1)^{w-1}l_{max}\left\lfloor\frac{n}{m(n)}\right\rfloor^2, \quad (114)$$

which completes the proof since $|D|, |A|, K$ and w are finite.

A.5   Proof of Proposition 16

Similar to the proof of Proposition 14, we have,

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

$$= \left|\alpha_l \hat{H}_{\Psi_B}^{k+1}(X|X^k) + \beta_l - \frac{1}{|B|}\left(\sum_{t=1}^{k}l(x_t, F^{k,opt}(x_1^{t-1})) + \sum_{t=k+1}^{|B|}l(x_t, F^{k,opt}(x_{t-k}^{t-1}))\right)\right|$$

$$= \left|\alpha_l \hat{H}_{\Psi_B}^{k+1}(X|X^k) + \beta_l - \right.$$

$$\left.\frac{1}{|B|}\sum_{t=1}^{k}l(x_t, F^{k,opt}(x_1^{t-1})) - \left(1-\frac{k}{|B|}\right)\frac{1}{|B|-k}\sum_{t=k+1}^{|B|}l(x_t, F^{k,opt}(x_{t-k}^{t-1}))\right|$$

$$\leq \left|\alpha_l \hat{H}_{\Psi_B}^{k+1}(X|X^k) + \beta_l - \frac{1}{|B|-k}\sum_{t=k+1}^{|B|}l(x_t, F^{k,opt}(x_{t-k}^{t-1}))\right| + \frac{kl_{max}}{|B|}. \quad (115)$$

Since the order of the predictor is fixed, we can use the definition of $\hat{P}_{\Psi_B}^{k+1}(s)$ ans sum over $s\in\{0,1\}^{k+1}$ instead of t. Thus,

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

$$\leq \left|\alpha_l \hat{H}_{\Psi_B}^{k+1}(X|X^k) + \beta_l - \sum_{s\in\{0,1\}^{k+1}}\hat{P}_{\Psi_B}^{k+1}(s)l(s_{k+1}, F^{k,opt}(s_1^k))\right| + \frac{kl_{max}}{|B|}$$

$$= \left|\sum_{s'\in\{0,1\}^k}\hat{P}_{\Psi_B}^{k+1}(s')\sum_{x\in\{0,1\}}\hat{P}_{\Psi_B}^{k+1}(x|s')\left(-\alpha_l\log\hat{P}_{\Psi_B}^{k+1}(x|s') + \beta_l - l(x, F^{k,opt}(s'))\right)\right|$$

$$+ \frac{kl_{max}}{|B|}$$

$$= \left|\sum_{s'\in\{0,1\}^k}\hat{P}_{\Psi_B}^{k+1}(s')\left(\alpha_l h_b(\hat{P}_{\Psi_B}^{k+1}(\cdot|s')) + \beta_l - \phi_l(\hat{P}_{\Psi_B}^{k+1}(\cdot|s'))\right)\right| + \frac{kl_{max}}{|B|}$$

$$\leq \sum_{s'\in\{0,1\}^k}\hat{P}_{\Psi_B}^{k+1}(s')\max_{p}|\alpha_l h_b(p) + \beta_l - \phi_l(p)| + \frac{kl_{max}}{|B|}$$

$$= \epsilon_l + \frac{kl_{max}}{|B|}. \quad (116)$$

References

[1] M.J. Weinberger, G. Seroussi, and G. Sapiro, "LOCO-I: A low complexity, context-based, lossless image compression algorithm," *Proc. IEEE Data Compression Conf.*, pp. 140–149, 1996.

[2] C.-H. Lamarque and F. Robert, "Image analysis using space-filling curves and 1D wavelet bases," *Pattern Recognition*, vol. 29, no. 8, pp. 1309–1322, 1996.

[3] A. Krzyzak, E. Rafajlowicz, and E. Skubalska-Rafajlowicz, "Clipped median and space filling curves in image filtering," *Nonlinear Analysis*, vol. 47, pp. 303–314, 2001.

[4] L. Velho and J. M. Gomes, "Digital halftoning with space filling curves," *Computer Graphics*, vol. 25, no. 4, pp. 81–90, July 1991.

[5] E. Skubalska-Rafajlowicz, "Pattern recognition algorithms based on space-filling curves and orthogonal expansions," *IEEE Trans. Inform. Theory*, vol. 47, no. 5, pp. 1915–1927, July 2001.

[6] T. Asano, D. Ranjan, T. Roos, E. Welzl, and P. Widmayer, "Space-filling curves and their use in the design of geometric data structures," *Theoretical Computer Science*, vol. 181, pp. 3–15, 1997.

[7] K.-L. Chung, Y.-H. Tsai, and F.-C. Hu, "Space-filling approach for fast window query on compressed images," *IEEE Trans. img. processing*, vol. 9, no. 12, pp. 2109–2116, 2000.

[8] B. Moon, H. V. Jagadish, C. Faloutsos, and J. H. Saltz, "Analysis of the clustering properties of the Hilbert space-filling curve," *IEEE Trans. Knowledge and Data Engineering*, vol. 13, no. 1, pp. 124–141, January/February 2001.

[9] A. Bogomjakov and C. Gotsman, "Universal rendering sequences for transparent vertex caching of progressive meshes," *Computer Graphics Forum*, vol. 21, no. 2, pp. 137–148, 2002.

[10] R. Niedermeier, K. Reinhardt, and P. Sanders, "Towards optimal locality in mesh-indexings," *Discrete Applied Mathematics*, vol. 117, pp. 211–237, 2002.

[11] H. Tang, S.-I. Kamata, K. Tsuneyoshi, and M.-A. Kobayashi, "Lossless image compression via multi-scanning and adaptive linear prediction," *IEEE Asia-Pacific Conference on Circuits and Systems*, pp. 81–84, December 2004.

[12] N. D. Memon, K. Sayood, and S. S. Magliveras, "Lossless image compression with a codebook of block scans," *IEEE Journal on Selected Areas In Communications*, vol. 13, no. 1, pp. 24–30, January 1995.

[13] R. Dafner, D. Cohen-Or, and Y. Matias, "Context-based space filling curves," *EURO-GRAPHICS*, vol. 19, no. 3, 2000.

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

[15] 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.

[16] T. Weissman and S. Mannor, "On universal compression of multi-dimensional data arrays using self-similar curves," in *Proc. 38th Annu. Allerton Conf. Communication, Control, and Computing*, October 2000, vol. I, pp. 470–479.

[17] Z. Ye and T. Berger, *Information measures for discrete random fields*, Science Press, Beijing, 1998.

[18] J. Ziv, "On universal quantization," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 344–347, May 1985.

[19] A. Dembo and I. Kontoyiannis, "Source coding, large deviations, and approximate pattern matching," *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1590–1615, June 2002.

[20] I. Kontoyiannis, "Pattern matching and lossy data compression on random fields," *IEEE Trans. Inform. Theory*, vol. 49, pp. 1047–1051, April 2003.

[21] N. Memon, D. L. Neuhoff, and S. Shende, "An analysis of some common scanning techniques for lossless image coding," *IEEE Trans. on Image Processing*, vol. 9, no. 11, pp. 1837–1848, November 2000.

[22] 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.

[23] N. Merhav and M. Feder, "Universal prediction," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2124–2147, October 1998.

[24] L. Gyorfi, G. Lugosi, and G. Morvai, "A simple randomized algorithm for sequential prediction of ergodic time series," *IEEE Trans. Inform. Theory*, vol. 45, no. 7, pp. 2642–2650, November 1999.

[25] T. Weissman and N. Merhav, "Universal prediction of random binary sequences in a noisy environment," *Ann. App. Prob.*, vol. 14, no. 1, pp. 54–89, February 2004.

[26] V. G. Vovk, "Aggregating strategies," *Proc. 3rd Annu. Workshop Computational Learning Theory,* San Mateo, CA, pp. 372–383, 1990.

[27] A. Cohen, T. Weissman, and N. Merhav, "Scanning and sequential decision making for multi-dimensional data - part II: the noisy case," 2007, in preparation.

[28] T. Weissman and N. Merhav, "Universal prediction of individual binary sequences in the presence of noise," *IEEE Trans. Inform. Theory*, vol. 47, pp. 2151–2173, September 2001.

[29] T. Weissman, E. Ordentlich, M. Weinberger, A. Somekh-Baruch, and N. Merhav, "Universal filtering via prediction," *IEEE Trans. Inform. Theory*, vol. 53, no. 4, pp. 1253–1264, April 2007.

[30] T. Linder and G. Lugosi, "A zero-delay sequential scheme for lossy coding of individual sequences," *IEEE Trans. Inform. Theory*, vol. 47, no. 6, pp. 2533–2538, September 2001.

[31] T. Weissman and N. Merhav, "On limited-delay lossy coding and filtering of individual sequences," *IEEE Trans. Inform. Theory*, vol. 48, no. 3, pp. 721–733, March 2002.

[32] A. Gyorgy, T. Linder, and G. Lugosi, "Efficient adaptive algorithms and minimax bounds for zero-delay lossy source coding," *IEEE Trans. Signal Processing*, vol. 52, no. 8, pp. 2337–2347, August 2004.

[33] N. Merhav, E. Ordentlich, G. Seroussi, and M. J. Weinberger, "On sequential strategies for loss functions with memory," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1947–1958, July 2002.

[34] D. Haussler, J. Kivinen, and M. K. Warmuth, "Sequential prediction of individual sequences under general loss functions," *IEEE Trans. on Information Theory*, vol. 44, no. 5, pp. 1906–1925, 1998.

[35] R. M. Gray and J. C. Kieffer, "Asymptotically mean stationary measures," *Ann. Prob.*, vol. 8, no. 5, pp. 962–973, 1980.

[36] B. Porat, *Digital Processing of Random Signals*, Prentice-Hall, New Jersey, 1994.

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

[38] X. Guyon, *Ramdom Fields on a Network*, Springer-Verlag, 1995.

[39] H. Sagan, *Space-Filling Curves*, Springer-Verlag, New York, 1994.

[40] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.

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