

Scanning and Sequential Decision Making for Multidimensional Data

Asaf Cohen

Department of Electrical Engineering
Technion - Israel Institute of Technology
Haifa 32000, ISRAEL
Email: soofsoof@tx.technion.ac.il

Neri Merhav

Department of Electrical Engineering
Technion - Israel Institute of Technology
Haifa 32000, ISRAEL
Email: merhav@ee.technion.ac.il

Tsachy Weissman

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

Abstract—We investigate several problems in scanning of multidimensional data arrays, such as universal scanning and prediction (“scandiction”, for short), and scandiction of noisy data arrays. These problems arise in several aspects of image and video processing, such as predictive coding, filtering and denoising. In predictive coding of images, for example, an image is compressed by coding the prediction 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 if there exist specific scandiction schemes which are universal in some sense.

More specifically, we investigate the following problems: First, given a random field, we 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 a bound on the excess loss compared to optimal scandiction. For individual data arrays, where we show that universal scandictors with respect to arbitrary finite scandictor sets do not exist, we show that the Peano-Hilbert scan has a uniformly small redundancy compared to optimal finite state scandiction.

Finally, we examine the scenario where the random field is corrupted by noise, but the scanning and prediction (or filtering) scheme is judged with respect to the underlying noiseless field. A special emphasis is given to the interesting scenarios of binary random fields communicated through binary symmetric channels and Gaussian random fields corrupted by additive white Gaussian noise.

I. INTRODUCTION

Consider the problem of sequentially scanning and predicting a multidimensional data array, while minimizing a given loss function. Particularly, at each time instant t , $1 \leq t \leq |B|$, 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 scandiction problem mainly arises in image compression, where various methods of predictive coding are used (e.g., the LOCO algorithm [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 several other areas, such as one-dimensional wavelet processing of images [2], where one seeks a space-filling curve which facilitates the one-dimensional signal processing of multidimensional data, and pattern recognition [3], 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 [4] and indexing [5], 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 [6], where a rendering sequence which minimizes the number of cache misses is required.

While the problem of sequentially predicting the next outcome of a one-dimensional sequence, based on the previously observed outcomes, is well-studied, the problem of prediction in multidimensional data arrays 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.

In [7], 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 [8] 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 were later extended in [9] to the probabilistic setting, where it was shown that this algorithm is also universal for any stationary Markov random field, and the existence of a universal rate-distortion encoder was established. Additional results regarding lossy compression of random fields (via pattern matching) were given in [10] and [11]. In [11], for example, 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 [7] 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. 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 [7] for two-dimensional data, yet they are favorable when compared to applying the algorithm in [7] *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 [14], 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 field. Surprisingly, the results of [14] 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, the cumulative *squared* prediction error, the raster scan is indeed optimal for Gaussian fields, as it was shown later in [15], which we discuss next.

The results of [7] and [9] 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 [15], 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 [15] 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).

The work in [15] defined the yardstick for analyzing scanning and prediction in multidimensional arrays. However, many challenges were left open. As the topic of prediction in one-dimensional arrays is rich and includes elegant solutions to various prediction problems, seeking analogous results in the multidimensional case offers plentiful research objectives.

The outline of this paper is as follows: In Section II, we give a precise formulation of the scandiction problem. In Section III, we first show that there does not exist any universal scandictor which can compete successfully (i.e., with a vanishing redundancy) with any two scandictors on any individual image. However, we show that it is possible to compete successfully with any finite set of scandictors on any stationary random field, and, in fact, there exists a universal scandictor which achieves the scandictability of any stationary random field. In Section IV, we consider strongly mixing random fields, and show that the results of Section III apply in the stronger a.s. (almost surely) sense as well. In Section V, we derive bounds on the performance of non-optimal scanners. Specifically, we assume that, due to implementation constraints, for example, one cannot use the optimal scanner for a given data array, or the universal scandictor discussed previously, and is forced to use an arbitrary scanning order. In such a scenario, it is important to understand what is the excess loss incurred, compared to optimal scanning and prediction. Section V includes upper bounds on this excess loss, which are valid for arbitrary distributed random fields, thus helping the designers of practical scanning algorithms in intelligently allocating computational and storage resources. In Section VI we return to the individual image scenario, and show that the Peano-Hilbert scan has a uniformly small redundancy (excess loss) with respect to the set of all finite state scandictors. Finally, in Section VII, we consider the scenario where the scandictor observes a noisy version of the data, yet, it is judged with respect to the clean data. We formally define scandiction

(or scanning and filtering - “scantering”, for short) of noisy data arrays, derive explicit expressions for the best achievable performance, and construct universal algorithms which achieve this value. Therein, we see that while many of the results for noisy scandiction are extendable from the noiseless case, the scantering problem, however, poses new challenges and requires the use of new tools and techniques.

II. PROBLEM FORMULATION

The following notation will be used throughout this paper. 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, they are formulated for $d = 2$. A probability measure Q on Ω is stationary if it is invariant under translations τ_i , $i \in \mathbb{Z}^2$ (i.e., shift invariant). Denote by $\mathcal{M}(\Omega)$ and $\mathcal{M}_S(\Omega)$ the spaces of all probability measures and stationary probability measures on Ω , respectively. Elements of $\mathcal{M}(\Omega)$, *random fields*, will be denoted by upper case letters while elements of Ω , *individual data arrays*, will be denoted by the corresponding lower case.

Let \mathcal{V} denote the set of all finite subsets of \mathbb{Z}^2 . For $V \in \mathcal{V}$, denote by X_V the restrictions of the data array X to V . For $i \in \mathbb{Z}^2$, X_i is the random variable corresponding to X at site i . Denote by V_n the square $\{0, \dots, n-1\} \times \{0, \dots, n-1\}$.

Definition 1 ([15]): A *scandictor* for the finite set of sites $B \in \mathcal{V}$ is the following pair (Ψ, F) :

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

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

- $\{F_t\}_{t=1}^{|B|}$ is a sequence of measurable predictors, $F_t : A^{t-1} \mapsto D$ determining the prediction for the site visited at time t based on the observations at past visited sites, where 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. Note that definition 1 considers only scandictors for a finite set of sites. We will consider, though, the limit as the cardinality of the set tends to infinity.

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

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

where $l : A \times D \rightarrow [0, \infty)$ 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

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

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

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

whenever the limit exists. By [15, Theorem 1], the limit in (4) exists for any $Q \in \mathcal{M}_S(\Omega)$.

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 , yet all these scanners comply with the same rule. In this case, $\mathcal{F} = \{\mathcal{F}_n\}$ can also be viewed as one finite set \mathcal{F} which includes sequences of scandictors. We may also consider cases in which $|\mathcal{F}_n|$ increases with n (but finite for 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

$$U_{\mathcal{F}}(l, Q) \triangleq \lim_{n \rightarrow \infty} \min_{(\Psi, F) \in \mathcal{F}_n} E_{Q_{V_n}} \frac{1}{|V_n|} L_{(\Psi, F)}(X_{V_n}), \quad (5)$$

if it exists.

III. UNIVERSAL SCANDICTION

The problem of universal prediction in the one-dimensional scenario is well studied, with various solutions to both the stochastic setting as well as the individual (see [16] for a complete survey from an information theoretic point of view). 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 [17] 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. The following theorem asserts that indeed, 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 [18, Section 3.3] and elaborated on in Section VI).

Theorem 2: Let $A = [0, 1]$ and assume l is the squared error 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

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

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 one-dimensional scenario (i.e., with no scanning), it is possible to compete with *any finite set* of predictors under the alphabet $[0, 1]$ and squared error loss. Thus, although the scandiction problem is strongly related to its one-dimensional analogue, the numerous scanning possibilities result in a substantially richer and more challenging problem.

Proof Outline (Theorem 2): We prove Theorem 2 by showing that 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, and, thus, for any scandictor there exists an individual image on which it cannot compete successfully with the two scandictors.

Let Y_{V_n} be a random field such that $Y(1, 1)$ is distributed uniformly on $[0, 1]$, and $Y_{V_n} \setminus Y(1, 1) = Y(1, 2), \dots, Y(1, n), Y(2, 1), Y(2, 2), \dots, Y(n, n)$ 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. The expected cumulative squared error loss of any scandictor on X_{V_n} is $(n^2 - 1)/8$, 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. However, the expected *minimum of the losses* of two different scandictors, one which scandicts X_{V_n} row-wise from $X(1, 1)$ to $X(n, n)$, and one which scandicts X_{V_n} row-wise from $X(n, n)$ to $X(1, 1)$, 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. ■

The obstacle faced when seeking universal scandictors is in some way similar to the one faced in [19], where the consideration of a loss function with memory prevented the alternation of experts each time instant, or to source coding problems such as [20], where the price (in terms of bits transferred) might be too high to bear if experts are alternated too frequently. The difficulties in these examples differ from those we confront here. Yet, 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 scandiction problem.

Particularly, in order to establish the existence of a universal scandictor, we propose the following algorithm. Let x_{V_n} be the $n \times n$ data array to be scandicted. For $m < n$, define $K \triangleq \lceil \frac{n}{m} \rceil - 1$. 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. This scanning order is irrelevant in this case, so we assume from now on that it is a (continuous) raster

scan from the upper left corner. Let $\mathcal{F} = \{\mathcal{F}_n\}$ be the sequence of scandictor sets. 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

$$P_i(j|\{L_{j,i}\}_{j=1}^\lambda) = \frac{e^{-\eta L_{j,i}}}{\sum_{j=1}^\lambda e^{-\eta L_{j,i}}}, \quad (6)$$

where $\lambda = |\mathcal{F}_m|$. For each $0 \leq i \leq (K + 1)^2 - 1$, after scanning i blocks of data, the algorithm computes $P_i(j|\{L_{j,i}\}_{j=1}^\lambda)$ 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.

The following two propositions, whose complete proofs are given in [18], form the foundations for the universality results.

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

$$\bar{L}_{alg}(x_{V_n}) - L_{min}(x_{V_n}) \leq m(n + m)\sqrt{\log \lambda} \frac{l_{max}}{\sqrt{2}}. \quad (7)$$

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

Based on the result given in Proposition 3, the next theorem asserts the existence of a universal scandictor which competes successfully with any finite set of scandictors.

Theorem 5: 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 $(\hat{\Psi}, \hat{F})_n$, independent of Q , for which

$$\begin{aligned} & \liminf_{n \rightarrow \infty} E_{Q_{V_n}} E_{\frac{1}{|V_n|}} L_{(\hat{\Psi}, \hat{F})_n}(X_{V_n}) \\ & \leq \liminf_{n \rightarrow \infty} \min_{(\Psi, F) \in \mathcal{F}_n} E_{Q_{V_n}} \frac{1}{|V_n|} L_{(\Psi, F)}(X_{V_n}) \end{aligned} \quad (8)$$

for any $Q \in \mathcal{M}_S(\Omega)$, where the inner expectation in the l.h.s. of (8) is due to the possible randomization in $(\hat{\Psi}, \hat{F})_n$.

Proof Outline: Taking the expectation of (7) with respect to X_{V_n} , then taking $n \rightarrow \infty$, results in a vanishing redundancy whenever the block size, m , is $o(n)$. However, this only means that the universal scandictor suggested competes successfully with the best scandictor in \mathcal{F}_m , operating on X_{V_n} in a block-wise order. Yet, using the spatial stationarity of X ,

it is not hard to show that the performance of any block-wise scandictor approaches that of the non-constrained one (operating on the entire data array) if $m = \omega(1)$, that is, $\lim_{n \rightarrow \infty} m(n) = \infty$. Thus, using the suggested algorithm with $m = o(n)$ but $m = \omega(1)$ achieves (8). ■

A. Finite-State Scandiction

Consider now the set of finite-state scandictors, very similar to the set of finite-state encoders described in [7]. 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 [7], 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 6: Let $\mathcal{F}_\nu = \{(\Psi, F)_j\}$ be the set of all finite-state scandictors with at most ν states. Then, for any $Q \in \mathcal{M}_S(\Omega)$, $\lim_{\nu \rightarrow \infty} U_{\mathcal{F}_\nu}(l, Q) = U(l, Q)$. That is, the scandictability of any spatially stationary source is asymptotically achieved with finite-state scandictors.

Proof Outline: For a finite block size m , every scandictor operating on that block can be implemented using a finite state machine with $\nu(m) < \infty$ states (of course, $\lim_{m \rightarrow \infty} \nu(m) = \infty$, yet $\nu(m)$ is independent of n). Thus, every scandictor operating on X_{V_n} block-wise, with a block size m , can be implemented using a finite state machine with $\nu(m) + 2$ states. Taking $n \rightarrow \infty$, it is clear that finite state machines with $\nu(m) + 2$ states achieve the m -th order scandictability of any spatially stationary random field (namely, the best achievable performance among block-wise scandictors with block size m). The lemma follows by taking the number of states to infinity as well. ■

Assume now that both the source alphabet A and the prediction alphabet D are finite. The following theorem asserts, under the above assumption, the existence of a universal scandictor for all stationary random fields.

Theorem 7: Let X be a stationary random field over a finite alphabet A and a probability measure Q . Assume that the prediction alphabet D is finite. Then, there exists a sequence of scandictors $(\Psi, F)_n$, independent of Q , for which

$$\lim_{n \rightarrow \infty} E_{Q_{V_n}} E \frac{1}{|V_n|} L_{(\Psi, F)_n}(X_{V_n}) = U(l, Q) \quad (9)$$

for any $Q \in \mathcal{M}_S(\Omega)$, where the inner expectation in the l.h.s. of (9) is due to the possible randomization in $(\Psi, F)_n$.

Theorem 7 is proved using the same method used in Lemma 6, with the additional effort required to show that the set of all

block-wise scandictors with block size $m(n)$ is not too large to compete with, using the appropriate choice of $m(n)$. The complete proof is in [18].

IV. UNIVERSAL SCANDICTION FOR MIXING RANDOM FIELDS

In Section III, we established the existence of a universal scandictor with respect to any finite set of scandictors (Theorem 5), and that of a universal scandictor achieving the scandictability of any stationary random field (Theorem 7), both under the expected cumulative loss criterion. Apparently, if the underlying random field adheres to a decaying memory condition, stronger convergence results can be achieved, as given by the next two theorems, whose proofs are in [18].

We start with several definitions. For $A, B \in \mathbb{Z}^2$, define

$$\alpha^Q(A, B) = \sup |Q(U \cap V) - Q(U)Q(V)|, \quad (10)$$

where the supremum is over all $U \in \sigma(X_A)$ and $V \in \sigma(X_B)$, and $\sigma(X_V)$ is the smallest sigma algebra generated by X_V . Let $\alpha_{a,b}^Q(k)$ denote the strong mixing coefficient of Q ,

$$\alpha_{a,b}^Q(k) = \sup \{ \alpha^Q(A, B), |A| \leq a, |B| \leq b, d(A, B) \geq k \}, \quad (11)$$

where d is a metric on \mathbb{Z}^2 and $d(A, B)$ is the distance between the closest points, i.e., $d(A, B) = \min_{i \in A, j \in B} d(i, j)$. A measure Q is strongly mixing if for all $a, b \in \mathbb{N} \cup \{\infty\}$, $\alpha_{a,b}^Q(k) \rightarrow 0$ as $k \rightarrow \infty$. It is not hard to show that if the measure Q is strongly mixing, then it is block-ergodic for any finite block size (i.e., totally ergodic).

The following theorem is the analogue to Theorem 5.

Theorem 8: 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 Section III uses a fixed block size m , we have

$$\liminf_{n \rightarrow \infty} \frac{1}{|V_n|} L_{alg}(X_{V_n}) \leq U_{\mathcal{F}}(l, Q) + \delta(m) \quad Q - a.s. \quad (12)$$

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

Similarly to Section III, a universal scandictor for the class of all stationary strongly mixing random fields can be constructed, whose performance converge Q-a.s. to the fields scandictability.

Theorem 9: Let X be a stationary strongly mixing random field with a probability measure Q . Then, there exists a sequence of scandictors $\{(\Psi, F)_n\}$, independent of Q , where $(\Psi, F)_n$ is a scandictor for V_n and operates in blocks of size $m \times m$, $m < n$, for which

$$\liminf_{n \rightarrow \infty} \frac{1}{|V_n|} L_{(\Psi, F)_n}(X_{V_n}) \leq U(l, Q) + \delta(m) \quad Q - a.s. \quad (13)$$

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

V. BOUNDS ON THE EXCESS SCANDICTION LOSS FOR NON-OPTIMAL SCANNERS

The results of Sections III and IV establish the existence of a universal scandictor for all stationary random fields and bounded loss function (under the terms of Theorem 7 or Theorem 9, respectively). However, it is clear that implementation of the universal algorithm suggested therein may be too complex in real-world applications. On the other hand, even if the probability measure governing the data is known completely, the cases where the optimal scandiction scheme is known are limited. In fact, only when the random field can be represented autoregressively (under a specific scan) with a maximum entropy innovation process, the optimal scandiction scheme is known [15]. While the results of [15] cover, for example, the interesting case of Gaussian fields and squared error loss, the problem of identifying the optimal scandiction scheme for general random fields and loss functions remains open.

Hence, it is interesting to investigate what is the excess scandiction loss when non-optimal scanners are used. 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, $l : \{0, 1\} \times [0, 1] \rightarrow \mathbb{R}$ is the loss function. Let ϕ_l denote the Bayes envelope associated with l , i.e.,

$$\phi_l(p) = \min_{q \in [0, 1]} [(1-p)l(0, q) + pl(1, q)]. \quad (14)$$

We further define

$$\epsilon_l = \min_{\alpha, \beta} \max_{0 \leq p \leq 1} |\alpha h_b(p) + \beta - \phi_l(p)|, \quad (15)$$

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

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, $U(l, Q_B) = \inf_{\Psi} E_{Q_B} \frac{1}{|B|} L_{(\Psi, F^{opt})}(X_B)$, we have the following result.

Theorem 10: Let X_B be an arbitrarily distributed binary field. Then, for any scan Ψ ,

$$\left| E_{Q_B} \frac{1}{|B|} L_{(\Psi, F^{opt})}(X_B) - U(l, Q_B) \right| \leq 2\epsilon_l. \quad (16)$$

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 $2\epsilon_l$.

Proof Outline (Theorem 10): In [18], we prove the following result: 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,

$$\left| \alpha_l \frac{1}{n} H(X^n) + \beta_l - \frac{1}{n} EL_l^{opt}(X^n) \right| \leq \epsilon_l, \quad (17)$$

where α_l and β_l are the achievers of the minimum in (15). That is, the difference between the best affine transform of the entropy of the random vector and the cumulative loss of the optimal predictor is bounded in by ϵ_l . Their key property of that bound is, of course, that even if the order of the random variables in the vector is permuted using any given scandictor, *the entropy of the vector remains the same*, and only the cumulative loss $E_l^{opt}(\tilde{X}^n)$ changes, where \tilde{X}^n is the permuted vector. Generalizing to data-dependent scans and random fields, and using the triangle inequality, Theorem 10 results. ■

At this point, it is important to understand why for logarithmic loss, $\epsilon_l = 0$, to wit, the scan is inconsequential under log loss. Indeed, when the loss function is the logarithmic loss, the expected instantaneous loss equals the conditional entropy, hence the expected cumulative loss coincides with the entropy, which is invariant to the scan. Namely, under the logarithmic loss function, *any scanner*, accompanied with the optimal predictor for that scan, results in optimal *scandiction* performance.

Finally, note that although the definitions of $\phi_l(p)$ and ϵ_l refer to the binary scenario, Theorem 10 holds for larger alphabets, with ϵ_l defined as in (15), 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.

VI. INDIVIDUAL IMAGES AND THE PEANO-HILBERT SCAN

Theorems 5 and 8 relied on the stationarity, or the stationarity and mixing property, of the random field X (respectively). When proving Theorem 5, 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 was used. When proving Theorem 8, 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. Hence, both results depend on the statistical properties of the source X .

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. This point was further emphasised in Theorem 2,

where it was shown that universal scandictors which compete with any two scandictors do not exist in the individual setting.

In [18, Section 3.3], we suggest a basic scenario under which universal scandiction of individual images is possible. However, identifying more sets of scandictors with which one can compete successfully on any individual image is an important open problem. For example, it is not clear whether one can compete successfully with all finite state scandictors.

Nevertheless, in the individual setting, one can expect to find a scanning scheme with uniformly small (but not vanishing) redundancy with respect to arbitrary scandictors set. In this section, we show that for the interesting class of all finite state scandictors, such a scan exists.

Consider the scenario of predicting the next outcome of a binary individual source, with $D = [0, 1]$ as the prediction space. Let Ψ_B be a scanner for the data array x_B . Let $x_1^{|B|}$ be the sequence resulting from scanning x_B with Ψ_B . Fix $k < |B|$ and for any $s \in \{0, 1\}^{k+1}$ define the empirical distribution of order $k+1$ as

$$\hat{P}_{\Psi_B}^{k+1}(s) = \frac{1}{|B| - k} \left| \{k < i \leq |B| : x_{i-k}^i = s\} \right|. \quad (18)$$

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

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

and

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

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

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

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

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

and $L_{\Psi}(x) = \lim_{k \rightarrow \infty} L_{\Psi}^k(x)$. Theorem 11 relates the asymptotic cumulative loss of any sequence of finite state scans Ψ to that resulting from the Peano-Hilbert sequence of scans, establishing the Peano-Hilbert sequence as an advantageous scanning order for any loss function.

Theorem 11: 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 $l : \{0, 1\} \times [0, 1] \rightarrow \mathbb{R}$,

$$L_{PH}(x) \leq L_{\Psi}(x) + 2\epsilon_l. \quad (23)$$

Theorem 11 relates the cumulative loss of a *single scan*, the Peano-Hilbert, to that of the cumulative loss of arbitrary finite state scan, unlike Theorem 10, which actually compares between any two scans. Furthermore, Theorem 11 is an asymptotic result, unlike Theorem 10, which is valid for any random field on a finite set of sites B . To understand these difficulties in the individual scenario, consider the following result, which relates the empirical conditional entropy to the cumulative loss under any scan Ψ_B , and is a key tool in proving Theorem 11.

Proposition 12: Let x_B be any data array. Then,

$$\left| \alpha_l \hat{H}_{\Psi_B}^{k+1}(X|X^k) + \beta_l - \frac{1}{|B|} L_{(\Psi_B, F^{k, opt})}(x_B) \right| \leq \epsilon_l + \frac{k l_{max}}{|B|}, \quad (24)$$

where α_l and β_l are the achievers of the minimum in (15). Proposition 12 is the individual setting analogue of equation (17). However, in the individual setting, the empirical conditional entropy $\hat{H}_{\Psi_B}^{k+1}(X|X^k)$ is not necessarily scan invariant. To see how Proposition 12 is utilized in order to prove Theorem 11, the following definitions are required.

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

$$\hat{H}_{\Psi}^{k+1}(x) = \limsup_{n \rightarrow \infty} \hat{H}_{\Psi_n}^{k+1}(X|X^k) \quad (25)$$

and further define $\hat{H}_{\Psi}(x) = \lim_{k \rightarrow \infty} \hat{H}_{\Psi}^{k+1}(x)$. The existence of $\hat{H}_{\Psi}(x)$ is established in [18], where it is also shown that this limit equals $\lim_{k \rightarrow \infty} \limsup_{n \rightarrow \infty} \frac{1}{k} \hat{H}_{\Psi_n}^k(X^k)$. By [8, Theorem 3], the later limit is no other than the asymptotic finite state compressibility of x under the sequence of scans Ψ , namely,

$$\begin{aligned} \lim_{k \rightarrow \infty} \limsup_{n \rightarrow \infty} \frac{1}{k} \hat{H}_{\Psi_n}^k(X^k) &= \rho(\Psi(x)) \\ &= \lim_{s \rightarrow \infty} \limsup_{n \rightarrow \infty} \rho_{E(s)}(\Psi_n(x_{V_n})), \end{aligned} \quad (26)$$

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 [8, eq. (1)-(4)]. Thus, $\hat{H}_{\Psi_B}^{k+1}(X|X^k)$, as the size of the data array and k tend to infinity, converges to $\rho(\Psi(x))$, which is the finite state compressibility of x . Since the finite state compressibility is achieved by the Peano-Hilbert scan, and *no other scan can achieve better compression*, we have (23).

Case Study: Hamming Loss. The bound in Theorem 11 is valid for any bounded loss function $l : \{0, 1\} \times [0, 1] \rightarrow \mathbb{R}$. When l is the Hamming loss, the resulting bound is

$$L_{PH}^{Hamming}(x) \leq L_{\Psi}^{Hamming}(x) + 0.16, \quad (27)$$

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

However, using the results of Feder, Merhav and Gutman in [21, Section 6], which lower and upper bound the finite state predictability (under Hamming loss) in terms of the finite state compressibility, it is possible to show [18] that using the Peano-Hilbert scan for scanning and prediction under Hamming loss one losses no more than $\frac{1}{2}\rho(x) - h^{-1}(\rho(x))$

with respect to *any finite-state scan* Ψ , where $\rho(x)$ is the image's FS compressibility. The maximum possible loss is 0.16, similar to the bound given in Theorem 11, 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.

VII. PREDICTION AND FILTERING OF NOISY DATA ARRAYS

In this section, we consider the scenario in which the decision maker has access only to noisy observations of the data. First, we consider the interesting scenario of filtering, where the predictor is replaced by a filter, which has access to the current (noisy) observation as well. In that case, a lower bound on the best achievable performance is given. For the practical cases of binary valued field observed through a binary symmetric channel and Gaussian field corrupted by additive white Gaussian noise, we bound the excess loss when a non-optimal scanner is used (with an optimal filter). We then return to the prediction scenario, characterize the noisy scandictability and the achieving scandictors in terms of the "clean" scandictability of the noisy data and give a bound on the excess loss when non optimal scanners are used. Results regarding the existence of universal scannerers and scandictors in the noisy setting can be found in [22].

We first formally define the noisy scenario. Let $\{(X_t, Y_t)\}_{t \in \mathbb{Z}^2}$ be a random field with components, (X_t, Y_t) , in $A \times N$, where N is the noisy observation alphabet. Here, $\{X_t\}_{t \in \mathbb{Z}^2}$ represents the clean signal and $\{Y_t\}_{t \in \mathbb{Z}^2}$ represents the noisy observations. We assume that the noisy observations are stochastically connected to the clean signal.

The notion of scandiction is similar to that of Section II, however, both the scanner and the predictor are allowed to access only the noisy signal $\{Y_t\}$. Namely, for any $B \in \mathcal{V}$, the scan, Ψ , is a sequence of measurable mappings $\{\Psi_t\}_{t=1}^{|B|}$, $\Psi_t : A^{t-1} \rightarrow B$, determining the next site in B according to the previously observed values $\{Y_i\}_{i \in \Psi_1, \dots, \Psi_{t-1}}$. The predictor, F , is a sequence of measurable mappings $\{F_t\}_{t=1}^{|B|}$, $F_t : A^{t-1} \rightarrow D$, determining the prediction for the value of x_{Ψ_t} according to the previously observed values $\{Y_i\}_{i \in \Psi_1, \dots, \Psi_{t-1}}$. The cumulative loss of a scandictor (Ψ, F) is given by $L_{(\Psi, F)}(x_B, y_B)$, the sum of the instantaneous losses over the array B , i.e.,

$$L_{(\Psi, F)}(x_B, y_B) = \sum_{t=1}^{|B|} l(x_{\Psi_t}, F_t(y_{\Psi_1}, \dots, y_{\Psi_{t-1}})). \quad (28)$$

The *noisy scandictability* is given by

$$\bar{U}(l, Q_B) = \inf_{(\Psi, F) \in \mathcal{S}(B)} E_{Q_B} \frac{1}{|B|} L_{(\Psi, F)}(X_B, Y_B), \quad (29)$$

and $\bar{U}(l, Q) = \lim_{n \rightarrow \infty} \bar{U}(l, Q_{V_n})$, if this limit exists.

In the important case where F_t is allowed to base its estimation on y_{Ψ_t} as well (scattering), we denote it by \tilde{F}_t ,

and we have

$$L_{(\Psi, \tilde{F})}(x_B, y_B) = \sum_{t=1}^{|B|} l(x_{\Psi_t}, \tilde{F}_t(y_{\Psi_1}, \dots, y_{\Psi_t})), \quad (30)$$

$$\tilde{U}(l, Q_B) = \inf_{(\Psi, \tilde{F})} E_{Q_B} \frac{1}{|B|} L_{(\Psi, \tilde{F})}(X_B, Y_B), \quad (31)$$

and $\tilde{U}(l, Q) = \lim_{n \rightarrow \infty} \tilde{U}(l, Q_{V_n})$, if this limit exists. Analogously to [15, Theorem 1], it can be shown that for any stationary random field both $\bar{U}(l, Q)$ and $\tilde{U}(l, Q)$ exist. The proofs of the results to follow can be found in [22].

A. Scattering of Noisy Binary Data Arrays

We assume an invertible memoryless channel, meaning the input distribution of a single symbol is uniquely determined given the output distribution. In this case, the associated Bayes envelope is $\phi_l(P) = \min_{f(\cdot)} El(X, f(Y))$, where P is the distribution of the channel output Y . Define $\zeta(d) = \max\{H(P) : \phi_l(P) \leq d\}$ and let $\bar{\zeta}(\cdot)$ be the upper concave envelope of $\zeta(\cdot)$. The following theorem is the direct analogue of the lower bounds in [15] for the filtering scenario. Note, however, that it holds for any finite n .

Theorem 13: Let Y_B be the output of an invertible memoryless channel whose input is X_B . Then, for any scanner (Ψ, \tilde{F}) we have

$$\bar{\zeta} \left(\frac{1}{|B|} E_{Q_B} L_{\Psi, \tilde{F}}(X_B, Y_B) \right) \geq \frac{1}{|B|} H(Y_B). \quad (32)$$

In particular, $\bar{\zeta}(\tilde{U}(l, Q_B)) \geq \frac{1}{|B|} H(Y_B)$, where $H(Y_B)$ is the entropy of Y_B .

Theorem 13 lower bounds the best possible scattering performance. Yet, there are scenarios where the function $\bar{\zeta}$ does not have a closed form expression [22]. Moreover, it may also be hard to evaluate the entropy rate of the noisy field Y_B . Thus, calculating explicit bounds on the scattering \tilde{U} , and, more importantly, identifying random fields and channels for which those bounds are achieved, is an interesting open problem.

Similarly to Section V, we proceed to derive a bound on the excess loss when a non-optimal scanning order is used (yet with the optimal filter for that scan). Define

$$f_\delta(p) = \min \left\{ \frac{p - \delta}{1 - 2\delta}, \delta \right\} \quad (33)$$

and

$$\epsilon_\delta = \min_{a, b} \max_{\delta \leq p \leq 1/2} |ah_b(p) + b - f_\delta(p)|. \quad (34)$$

Under these definitions, we have the following theorem.

Theorem 14: Let Y_B be the output of a binary symmetric channel with crossover probability δ whose input is X_B . Then, for any scanner (Ψ, F^{opt}) , where \tilde{F}^{opt} is the optimal filter for the scan Ψ , we have

$$\left| \frac{1}{|B|} E_{Q_B} L_{\Psi, \tilde{F}^{opt}}(X_B, Y_B) - \tilde{U}(l_H, Q_B) \right| \leq 2\epsilon_\delta. \quad (35)$$

Even without evaluating ϵ_δ explicitly, it is easy to see that the excess loss when using non optimal scanners is quite small in the filtering scenario. For example, for $\delta = 0.1$ and $\delta = 0.25$ we have $\epsilon_\delta < 0.035$ and $\epsilon_\delta < 0.03$ respectively, yielding a maximal loss of 0.07 or even 0.06. This should be compared to 0.16 in the prediction scenario (or even larger values in the noisy prediction scenario). The fact that the filtering problem is less sensitive to the scanning order is quite clear as the noisy observation of X_{Ψ_t} is available under any scan.

B. Scattering of Noisy Gaussian Data Arrays

Although stated for discrete valued fields, Theorem 13 applies to real valued fields as well, corrupted by real valued noise, with the appropriate replacement of entropy by differential entropy. To bound the excess scattering loss, however, new tools should be used.

The following bound results from a relation between the performance of *discrete time* filtering and *continuous time* filtering, together with the fundamental result of Duncan [23] on the relation between mutual information and causal minimal mean square error estimation. From now on we assume the loss function is the squared error loss, denoted l_s .

We start with several definitions. With a slight abuse of notations, let X be a Gaussian random variable, $X \sim \mathcal{N}(0, \sigma_X^2)$. Consider the following two estimation problems:

- Estimating X based on $Y = X + N$, where $N \sim \mathcal{N}(0, \sigma_N^2)$, independent of X .
- Causally estimating $X_t \equiv X$, $t \in [0, 1]$, based on Y_t , which is an AWGN-corrupted version of X_t , the Gaussian noise having a double-sided spectrum of height σ_N^2 .

To bound the sensitivity of the scattering performance, we consider the difference

$$\int_0^1 \text{Var}(X|Y^t) dt - \text{Var}(X|Y). \quad (36)$$

It is not hard to show [22] that

$$\int_0^1 \text{Var}(X|Y^t) dt - \text{Var}(X|Y) = \sigma_N^2 f\left(\frac{\sigma_X^2}{\sigma_N^2}\right), \quad (37)$$

where

$$f(x) = \ln(1+x) - \frac{x}{x+1}. \quad (38)$$

Theorem 15: Let X_{V_n} be a Gaussian random field with a constant marginal distribution satisfying $\text{Var}(X_i) = \sigma_X^2 < \infty$ for all $i \in V_n$. Let $Y_i = X_i + N_i$, where N_{V_n} is a white Gaussian noise of variance σ_N^2 . Then, for any two scans Ψ^1 and Ψ^2 , we have

$$\begin{aligned} \frac{1}{n^2} \left| EL_{(\Psi^1, \bar{F}^{opt})}(X_{V_n}, Y_{V_n}) - EL_{(\Psi^2, \bar{F}^{opt})}(X_{V_n}, Y_{V_n}) \right| \\ \leq \sigma_N^2 f\left(\frac{\sigma_X^2}{\sigma_N^2}\right). \end{aligned} \quad (39)$$

Proof Outline: The comparison between any two scans is made easy by bounding the normalized cumulative loss of *any scan*

Ψ in terms of a *scan invariant* entity, which is the mutual information.

By [23], for the continuous time model, $dY_t = X_t dt + dW_t$, where W_t is a standard Brownian motion, the causal cumulative squared estimation error equals the mutual information between $\{X\}$ and $\{Y\}$. However, we are interested in a discrete time model. We thus construct a *piecewise constant* process from the discrete time (one-dimensional) process resulting from the scan, and describe the cumulative scattering loss in terms of the continuous time causal estimation error and f (note that f simple quantifies a difference between continuous and discrete time filtering). Since the causal estimation error equals the mutual information, which is shown to be scan invariant, the theorem results by upper and lower bounding the cumulative scandiction loss using the appropriate upper and lower bounds on f . ■

At this point, a few remarks are in order. The bound in Theorem 15 is applicable only to Gaussian random fields corrupted by AWGN. Two generalizations, one applicable to arbitrarily distributed input fields, and one applicable to continuous valued input fields can be found in [22].

The bound in Theorem 15 has the form $\text{Var}(X_1) \frac{f(\text{SNR})}{\text{SNR}}$, and we have

$$\lim_{\text{SNR} \rightarrow 0^+} \frac{f(\text{SNR})}{\text{SNR}} = \lim_{\text{SNR} \rightarrow \infty} \frac{f(\text{SNR})}{\text{SNR}} = 0, \quad (40)$$

that is, the scan is inconsequential at very high or very low SNR. The function $\frac{f(\text{SNR})}{\text{SNR}}$ has a unique maximum of approximately 0.216, that is, the excess loss due to a suboptimal scan at any SNR is upper bounded by $0.216 \text{Var}(X_1)$. However, it is possible to show that at high SNR, the bound in Theorem 15 is far from being tight, and analyzing simple symbol-by-symbol filtering results in a smaller excess loss bound.

C. Scandiction of Noisy Data Arrays

Analysis of the best achievable performance and bounds on the excess loss is simpler in the noisy prediction scenario (compared to filtering), as one can use the same tools used in the clean setting, with respect to a *modified loss function*. In this section, we briefly state the noisy scandiction results in the binary and Gaussian settings.

Let l_H denote the Hamming loss function. Let Q_Y denote the marginal distribution of the noisy observations field Y . The following lemma relates $\bar{U}(l_H, Q)$ to $U(l_H, Q_Y)$, and bounds the excess loss similar to Section V.

Lemma 16: Let $\{(X_t, Y_t)\}_{t \in \mathbb{Z}^2}$ be a binary random field governed by a probability measure Q such that $\{Y_t\}$ is the output of a binary memoryless symmetric channel with cross over probability $\delta < 1/2$ and input $\{X_t\}$. Then,

$$\bar{U}(l_H, Q) = \frac{U(l_H, Q_Y) - \delta}{1 - 2\delta}, \quad (41)$$

and $\bar{U}(l_H, Q)$ is achieved by the scandictor which achieves $U(l_H, Q_Y)$. Furthermore, for any scandictor (Ψ, F^{opt}) , where Ψ is arbitrary and F^{opt} is the optimal predictor for Ψ , we

have

$$\left| \frac{1}{|B|} E_{Q_B} L_{\Psi, \text{Fopt}}(X_B, Y_B) - \bar{U}(l_H, Q_B) \right| \leq \frac{2\epsilon_{l_H}}{1 - 2\delta}. \quad (42)$$

The following lemma is the direct analogue for squared error.

Lemma 17: Let $\{(X_t, Y_t)\}_{t \in \mathbb{Z}^2}$ be a random field governed by a probability measure Q such that $Y_t = X_t + N_t$, where N_t , $t \in \mathbb{Z}^2$, are i.i.d. random variables with $\text{Var}(N_t) = \sigma_N^2 < \infty$. Let $\{B_n\}_{n \geq 1}$ be any sequence of elements in \mathcal{V} , satisfying $R(B_n) \rightarrow \infty$. Then $\bar{U}(l_s, Q) = U(l_s, Q_Y) - \sigma_N^2$. Furthermore, $\bar{U}(l_s, Q)$ is achieved by the scandictor which achieves $U(l_s, Q_Y)$.

Actually, Lemmas 16 and 17 are only scarcely related to scanning. They merely states that in the prediction of a process based on its noisy observations, under the assumptions stated above, the optimal predictor is a one which disregards the noise, and attempts to predict the next *noisy* outcome.

Let both X and N be Gaussian random fields, where the components of N are i.i.d. and independent of X . That is, Y is the output of an additive white Gaussian noise channel, with a Gaussian input X . In this scenario, similarly to the clean one, the noisy scandictability is known exactly and is given by a single letter expression.

A subset $S \subseteq \mathbb{Z}^2$ is called a *half plane* if it is closed to addition and satisfies $S \cup (-S) = \mathbb{Z}^2$ and $S \cap (-S) = \{0\}$. We can now state the following corollary, regarding the noisy scandictability in the Gaussian regime and squared error loss, which is a trivial application of Lemma 17 and the results of [15, Section IV].

Corollary 18: Under the terms of Lemma 17, assuming both X and N are Gaussian, the noisy scandictability of Q is given by $\bar{U}(l_s, Q) = \sigma_u^2(Y) - \sigma_N^2$, where $\sigma_u^2(Y)$ is the squared error of the best linear predictor for Y_0 given Y_i , $i \in S \setminus \{0\}$, for any half plane S . Furthermore, $\bar{U}(l_s, Q)$ is asymptotically achieved by a scandictor which scans (X_t, Y_t) according to the total order defined by any half-plane S and applies the corresponding best linear predictor for the next outcome of Y .

VIII. CONCLUSION

In this paper, we first established the existence of a universal algorithm which achieves the scandictability of any spatially stationary random field. 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. Finally, the noisy scenario, where the scandictor (or scanner) has access only to a noisy observation of the data, was discussed.

Despite the illusive similarity to the one-dimensional analogue, the multidimensional setting can be strikingly different. Thus, as multidimensional data is extensively used in various multimedia applications, investigating the key process of scanning the data is essential. Besides the open problems mentioned throughout the sequel, there are plentiful directions to pursue. Among them are applications such as image classification or scandiction and scantering with limited

memory or complexity resources. Finally, deriving simple universal algorithms, or even suggesting simple algorithms with uniformly small redundancy, is still an issue waiting to be resolved.

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] 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.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] A. Lempel and J. Ziv, "Compression of two-dimensional data," *IEEE Trans. Inform. Theory*, vol. IT-32, no. 1, pp. 2–8, January 1986.
- [8] 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.
- [9] 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.
- [10] 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.
- [11] I. Kontoyiannis, "Pattern matching and lossy data compression on random fields," *IEEE Trans. Inform. Theory*, vol. 49, pp. 1047–1051, April 2003.
- [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," *EUROGRAPHICS*, vol. 19, no. 3, 2000.
- [14] 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.
- [15] 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.
- [16] N. Merhav and M. Feder, "Universal prediction," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2124–2147, October 1998.
- [17] V. G. Vovk, "Aggregating strategies," *Proc. 3rd Annu. Workshop Computational Learning Theory*, San Mateo, CA, pp. 372–383, 1990.
- [18] A. Cohen, N. Merhav, and T. Weissman, "Scanning and sequential decision making for multi-dimensional data - part I: the noiseless case," to appear in *IEEE Trans. on Inform. Theory*.
- [19] 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.
- [20] 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.
- [21] M. Feder, N. Merhav, and M. Gutman, "Universal prediction of individual sequences," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1258–1270, July 1992.
- [22] A. Cohen, T. Weissman, and N. Merhav, "Scanning and sequential decision making for multi-dimensional data - part II: the noisy case," submitted to *IEEE Trans. Inform. Theory*, May 2007.
- [23] T. E. Duncan, "On calculation of mutual information," *SIAM Journal of Applied Mathematics*, vol. 19, pp. 215–220, July 1970.