Scanning and Prediction in Multi-Dimensional Data Arrays*

Neri Merhav Tsachy Weissman

August 3, 2001

Abstract

The problem of sequentially scanning and predicting data arranged in a multi-dimensional array is considered. We introduce the notion of a *scandictor*, which is any scheme for the sequential scanning and prediction of such multi-dimensional data. The *scandictability* of any finite (probabilistic) data-array is defined as the best achievable expected "scandiction" performance on that array. The scandictability of any (spatially) stationary random field on $\mathbb{Z}^m$ is defined as the limit of its scandictability on finite "boxes" (subsets of $\mathbb{Z}^m$), as their edges become large. Fundamental limitations on scandiction performance in both the probabilistic and the deterministic settings are characterized for the family of difference loss functions. These results are specialized for cases of particular interest: The scandictability of any stationary Gaussian field under the squared-error loss function is given a single-letter expression in terms of its spectral measure. For a family of binary Markov Random Fields (MRFs), the scandictability under the Hamming distortion measure is fully characterized. Finally, we consider the problem of scandiction under more than one loss function.

1 Introduction

The main motivation for this work comes from predictive coding, a compression technique used for encoding images, voice signals, video signals, and other types of data. The basic idea consists of scanning the data array, constructing a model of the data, employing a predictor corresponding to the model, and then encoding the difference between the prediction and the actual outcome. Examples for predictive coding include LPC-based voice coders (e.g., [1]) and image coders (e.g., [2]). The compression efficiency of such schemes naturally boils down to the efficiency of the prediction scheme employed. Now, assuming the predictive coding part of the scheme is fixed (i.e., the way that the difference between the predicted value and the actual outcome is encoded), a degree of freedom left for optimization (besides the prediction itself) is the order in which the data is scanned. In this work, we take a first step in addressing the question of the optimal strategy for scanning and prediction of data contained in a multi-dimensional array.

In typical prediction problems the data is most naturally assumed ordered as a one-dimensional time series. In such problems sequentiality usually dictates only one possibility for scanning the data, namely, the direction of the flow of time. However, when the dimension of the data array, d, is larger than 1 (e.g., in image and video coding applications [3, 4, 5, 2]) there is no natural direction

*Authors are with the Department of Electrical Engineering, Technion- Israel Institute of Technology, Haifa 32000, Israel. merhav@ee.technion.ac.il, tsachy@tx.technion.ac.il.

of the flow of time and the question of the optimal scheme for scanning and predicting the data arises naturally.

For a concrete treatment of this question we shall introduce the notion of a "scandictor", which is any scheme for the sequential scanning and prediction of data contained in a multi-dimensional array, or, more generally, data which is indexed by a set which may not be naturally and uniquely ordered. For example, suppose that the data is arranged in an $n_1 \times n_2$ rectangular grid, e.g., an image where the data represents gray-level values. A scandictor operates as follows: At each time unit $1 \leq t \leq n_1 \cdot n_2$, having observed the values of the grid at the $t-1$ sites visited thus far, the scandictor chooses the t-th site (out of the remaining $[n_1 \cdot n_2 - (t-1)]$ unobserved sites), makes a prediction F_t for the value x_t at that site, and is then allowed to observe that value. The loss at time t is given by a fixed loss function $l(F_t, x_t)$. The goal is to minimize the cumulative "scandiction" loss $\sum_{t=1}^{n_1 \cdot n_2} l(F_t, x_t)$.

For the probabilistic setting, we define the "scandictability" of a source as the limit of the expected average loss per symbol for large blocks, when using the optimal scandictor for these blocks. By a sub-additivity argument, this limit can be shown to exist for any (spatially) stationary source and be independent of the ratios between the edges of the "box" confining the array.

After introducing the notions of a scandictor and scandictability in a general setting, we shall focus in Section 3 on the case where the data, as well as the predictions, are real-valued, and loss is measured w.r.t. a difference loss function[1]. Two approaches for assessing fundamental limitations on scandiction performance will be developed. The first, in subsection 3.A, will be based on the observation that for any sufficiently well-behaved (smooth) scandictor, the map which takes the data-array into the sequence of scandiction errors is a volume-preserving injection. As will be seen, this observation leads to several converse results on scandiction performance in a probabilistic as well as an "individual-data-array" setting. The second approach, in subsection 3.B, is based on MDL-type lower bounds [6, 7, 8]. More specifically, we extend an idea, which was applied in [9, subsection III.A] in the context of universal prediction of probabilistic time series, to the case of scandiction of "individual" data-arrays. Given an arbitrary scandictor, the idea is to construct a probability distribution such that an MDL-type lower bound for this distribution leads to a lower bound on the loss of the scandictor. As will be seen, one of the merits of this approach is that it allows to dispose of the regularity (smoothness) assumption needed for the validity of the converse results in subsection 3.A.

In Section 4 we pay special attention to the stationary Gaussian field on $\mathbb{Z}^2$. The main probabilistic result of Section 3 is applied to this special case. The scandictability of any stationary Gaussian field under the squared-error loss function is given a single-letter expression in terms of its spectral measure. Specifically, it is shown to be given by the power of the innovation process

[1] I.e., when $l(F_t, x_t) = \rho(x_t - F_t)$ for some $\rho(\cdot)$.

corresponding to any half-plane. In particular, this is shown to imply that the scandictability of the stationary Gaussian field is (asymptotically, for large rectangular arrays) achieved with any scan which corresponds to a total order on $\mathbb{Z}^2$ induced by any half-plane, a notion which will be made precise.

In Section 5 we consider the case where the alphabet and the prediction space are identical and finite. Furthermore, in order to paraphrase the type of arguments employed in Section 3 in the context of $\mathbb{R}$-valued observations and predictions, we assume here that the alphabet forms a group so that the subtraction operation is well-defined and the loss function is of the form $l(F_t, x_t) = \rho(x_t - F_t)$. Results pertaining to the fundamental limitations on scandiction performance for this setting are derived analogously as in Section 3, where "volume-preservation" arguments are replaced by "cardinality-preservation" ones. These results are then specialized to the case of the Hamming distortion measure. For a large family of MRFs, namely, those that can be auto-regressively represented, the scandictability is fully characterized.

The scandiction problem that we consider seems to be inherently different from standard problems involving cumulative loss minimization of predictors. While the latter are usually concerned with various on-line prediction scenarios, in this framework we are interested, in parallel to the prediction strategy, in finding the best strategy for scanning the data. To the best of our knowledge, the problem of finding the best scanning strategy, as it is formulated in this work, has not been previously considered in the literature. The issue of choosing the order by which pixels are scanned has been shown to be a consequential one in the context of universal compression of multi-dimensional data arrays. This was demonstrated by employing the self-similar Peano-Hilbert scan [10, 11] in the "individual-image" setting of [12] (cf. also [13] for the stochastic setting). As we elaborate on in Section 2, however, in the context of (non-universal) compression of probabilistic multi-dimensional data or, equivalently, of scandiction under the logarithmic loss function, the scheme used for the scan is completely immaterial: the scandictability (which, in this case, coincides with the entropy) is achieved by any scan (provided, of-course, the corresponding optimal predictor for that scan is employed). The incentive for this work was the fact that under loss functions other than the logarithmic loss it is a priori unclear which scan achieves optimum performance.

The remainder of the paper is organized as follows. In Section 2 we present the notation, formulate the general setting, and introduce the notion of a "scandictor" and the concept of "scandictability". Sections 3, 4, and 5 are as elaborated on above. In Section 6 we briefly consider the problem where scandiction performance is evaluated under more than one loss function. Finally, in Section 7, we mention some directions for related future work. For simplicity of the presentation we treat the case of a 2-dimensional data array throughout the paper. All the results carry over to higher dimensions in a straightforward way.

2 The Model, Notation, and Definitions

Let A (the alphabet) be a Polish space and let $\mathcal{A}$ be its Borel σ-algebra. A *configuration* over $\mathbb{Z}^2$ is an element $x = \{x_i\}_{i\in\mathbb{Z}^2}$ of the measurable space of configurations $(\Omega, \mathcal{F}) = (A, \mathcal{A})^{\mathbb{Z}^2}$. Let $\mathcal{M}(\Omega)$ denote the space of probability measures on Ω, and denote by $\mathcal{M}_s(\Omega)$ the subspace consisting of all elements of $\mathcal{M}(\Omega)$ that are invariant under all shifts $\theta_i : \Omega \to \Omega$, $i \in \mathbb{Z}^2$, where $(\theta_i x)_j = x_{j+i}$. For $B \subseteq \mathbb{Z}^2$ let $\mathcal{M}(B)$ denote the space of probability measures on $(A, \mathcal{A})^B$. An element of $\mathcal{M}(\Omega)$, $\mathcal{M}_s(\Omega)$, $\mathcal{M}(B)$ will be referred to as a *random field*, a *stationary random field*, a *random field on* B, respectively.

For $n \geq 1$, we will use the notation y_1^n to denote $(y_1, \ldots, y_n)$. For any positive integer n, let V_n be the $n \times n$ square of all $t = (t_1, t_2) \in \mathbb{Z}^2$ with both coordinates nonnegative and strictly less than n. For $V \subseteq \mathbb{Z}^2$, we let $x(V) = \{x(i)\}_{i\in V}$ denote the restriction of $x \in \Omega$ to V. Let $\mathcal{V}$ denote the class of finite subsets of $\mathbb{Z}^2$. For a source $Q \in \mathcal{M}(\Omega)$ we denote by E_Q expectation w.r.t. Q (though we omit the subscript when it is clear from the context). For any $V \subset \mathbb{Z}^2$, define $D(V)$ as the interior diameter of V:

$$D(V) \stackrel{\text{def}}{=} \sup\{r : \exists c \text{ s.t. } B(c, r) \subseteq V\}, \tag{1}$$

where we let $B(c, r)$ denote the closed ball of radius r centered at c under the l_1-norm on $\mathbb{Z}^2$. Following [14] we further let $\mathcal{R}_\square$ denote the system of all rectangles of the form $V = \mathbb{Z}^2 \cap ([m_1, n_1] \times [m_2, n_2])$ with $m_k, n_k \in \mathbb{Z}$, $m_k \leq n_k$. We let $\mathcal{R}_\square^0$ denote the subset of $\mathcal{R}_\square$ consisting of all boxes of the form $V = \mathbb{Z}^2 \cap ([0, n_1] \times [0, n_2])$ with $n_k \in \mathbb{Z}_+$. For any $V \subseteq \mathbb{Z}^2$ and $i \in \mathbb{Z}^2$ we shall let the standard notation $V + i$ signify the set $\{j \in \mathbb{Z}^2 : j - i \in V\}$ and $-V$ stand for $\{j \in \mathbb{Z}^2 : -j \in V\}$. We shall let $\mathbf{1}$ denote the point $(1, 1) \in \mathbb{Z}^2$. The cardinality of a set will be denoted by $|\cdot|$. For $i, j \in \mathbb{Z}^2$ we let $< i, j >$ denote their inner product, i.e., $i_1 j_1 + i_2 j_2$. If $\{A_n\}$ is a sequence of sets then $A_n \nearrow A$ ($A_n \searrow A$) is synonymous to "$A_n \subseteq A_{n+1}$ ($A_n \supseteq A_{n+1}$) and $\bigcup A_n = A$ ($\bigcap A_n = A$)". If $\{A_n\}$ is a sequence of reals then $A_n \nearrow A$ ($A_n \searrow A$) is synonymous to "$\{A_n\}$ is non-decreasing (non-increasing) and $\lim A_n = A$".

For a finite set of random variables $\mathbf{N}$, jointly distributed according to the probability distribution q_N, we shall let $H(\mathbf{N})$ as well as $H(q_N)$ denote the entropy. More precisely, the components of $\mathbf{N}$ will be either all discrete-valued or all continuous-valued, so that in the latter case $H(\mathbf{N})$ and $H(q_N)$ will stand for the differential entropy. Throughout this work we take all logarithms to the natural base and entropy is measured in nats. For a discrete- (continuous-) valued random variable Z we shall let $E_q f(Z)$ denote the expectation of $f(Z)$ when Z is distributed according to the Probability Mass Function (PMF) (Probability Density Function (PDF)[2]) $q(\cdot)$. For $B \in \mathcal{V}$ and a random field $Y(B)$ taking values in $\mathbb{R}^B$ with continuous-valued components, we shall consider its PDF $q(\cdot)$ as a function $q : \mathbb{R}^B \to [0, \infty)$ integrating to unity, with the obvious interpretation. For any finite set Σ,

[2]Here and throughout the sequel by a "continuous-valued random variable" we mean one with a distribution which is absolutely continuous w.r.t. Lebesgue measure, i.e., one with a PDF.

we let $M_1(\Sigma)$ denote the set of all probability measures on Σ.

For $B \subseteq \mathbb{Z}^d$ and data arrays on B with real-valued components $x(B), y(B) \in \mathbb{R}^B$ we denote $\|x(B) - Y(B)\|_\infty \stackrel{\mathrm{def}}{=} \max_{i \in B} |x_i - y_i|$ and we let $x(B) + y(B) \in \mathbb{R}^B$ denote the data array formed by component-wise addition.

A "Scandiction" Defined

Given data that are indexed by the set B, a *scandictor* is a scheme for the sequential scanning and prediction of this data. We formalize this as follows.

Definition 1 *A* scandictor *for the finite set of sites $B \in \mathcal{V}$ is given by a pair (Ψ, F) as follows:*

- *Ψ, the "scan", is a sequence of measurable mappings $\{\Psi_t\}_{t=1}^{|B|}$, where $\Psi_t : A^{t-1} \to B$, with the property that*

$$\{\Psi_1, \Psi_2(y_1), \Psi_3(y_1, y_2), \ldots, \Psi_{|B|}(y_1^{|B|-1})\} = B \quad \forall y_1^{|B|-1} \in A^{|B|-1}. \tag{2}$$

- *F, the predictor, is a sequence of measurable mappings $\{F_t\}_{t=1}^{|B|}$, where $F_t : A^{t-1} \to P$ (P is the prediction or "action" space assumed Polish and equipped with its Borel σ-algebra).*

We shall let $\mathcal{S}(B)$ denote the class of all scandictors for the set of sites B.

A scandictor $(\Psi, F) \in \mathcal{S}(B)$ operates as follows: The scandictor gives its first prediction, F_1, for the value at site Ψ_1. It then moves to site Ψ_1 and incurs a loss $l(F_1, x(\Psi_1))$. The scandictor now gives its prediction $F_2 = F_2(x(\Psi_1))$ (based on the value $x(\Psi_1)$ observed at site Ψ_1) for the value at site $\Psi_2 = \Psi_2(x(\Psi_1))$, it then moves to site Ψ_2 and incurs a loss $l(F_2, x(\Psi_2))$. Similarly, the scandictor gives its t-th prediction $(1 \leq t \leq |B|)$ $F_t = F_t(x(\Psi_1), x(\Psi_2), \ldots, x(\Psi_{t-1}))$ (based on the values $x(\Psi_1), x(\Psi_2), \ldots, x(\Psi_{t-1})$ observed at the previously visited sites) for the value at site $\Psi_t = \Psi_t(x(\Psi_1), x(\Psi_2), \ldots, x(\Psi_{t-1}))$, it then moves to site Ψ_t and incurs a loss $l(F_t, x(\Psi_t))$, where $l : P \times A \to [0, \infty)$ is a given loss function. Note that property (2) implies that no site is visited more than once so that all the sites of B have been covered after precisely $|B|$ steps. We let

$$L^l_{(\Psi,F)}(x(B)) = \frac{1}{|B|} \sum_{t=1}^{|B|} l(F_t, x(\Psi_t)) \tag{3}$$

denote the normalized cumulative loss, w.r.t. the loss function l, of the scandictor $(\Psi, F) \in \mathcal{S}(B)$ when operating on the restriction of x to B. The superscript l will be omitted when there is no room for ambiguity. Note that a scandictor, according to Definition 1, is not allowed to randomize its prediction or choice of the next site. I.e., its strategy at each point is deterministic (given the available information). Similarly as in the case of standard prediction, however, it is easy to show that there is no loss of optimality in this restriction insofar as expected performance is concerned.

Definition 2 *Given a loss function l, we define the* scandictability *of any source $Q \in \mathcal{M}(\Omega)$ on $B \in \mathcal{V}$ by*

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

where E_{Q_B} denotes expectation when $X(B)$ has been generated by Q_B. We further define the scandictability *of $Q \in \mathcal{M}(\Omega)$ by*

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

whenever the limit exists.

Note the special case where $l = l_{log}$ is the logarithmic loss function: When the alphabet A is finite, the prediction space is $P = M_1(A)$, and

$$l_{log}(F, a) = -\log F(a) \quad a \in A. \tag{6}$$

In this case, for any $B \in \mathcal{V}$ and $Q_B \in \mathcal{M}(B)$ the scandictability coincides with the (normalized) entropy, i.e.

$$U(l_{log}, Q_B) = \frac{1}{|B|} H(Q_B). \tag{7}$$

The proof of this simple fact extends verbatim from the case of regular predictability (cf., e.g., [15, 9]) by showing that to every scandictor $(\Psi, F) \in \mathcal{S}(B)$ there corresponds a probability measure $Q_{(\Psi,F)}$ on A^B such that $E_{Q_B} L_{(\Psi,F)}(X(B)) = E_{Q_B}\left[-\log Q_{(\Psi,F)}(X(B))\right]$. Using the fact that $D(Q_B \| \tilde{Q}_B) \geq 0$ for all $\tilde{Q}_B \in \mathcal{M}(B)$ it is then easy to show that $\frac{1}{|B|}H(Q_B)$ is an attainable lower bound on the scandictability $U(l_{log}, Q_B)$. Another way of seeing why (7) should hold is to note that the expected loss of the optimal predictor (under log-loss) associated with any scan is given by a summation of conditional entropies, which always sum up to the joint entropy, regardless of the scan. Hence, not only does equality (7) hold, but the scandictability is attained by *any* scan. In this context, the scandictability notion of Definition 2 can be regarded an extension of entropy for the case of a general loss function.

Analogously as with the notation for entropy, we shall sometimes write $U(l, X(B))$ (resp. $U(l, X)$) instead of $U(l, Q_B)$ (resp. $U(l, Q)$) when it is clear from the context that $X(B)$ (resp. X) is distributed according to Q_B (resp. Q). The definition of $U(l, Q)$ in (5) through a limit over the squares V_n may seem, at first glance, somewhat arbitrary. The justification for such a definition lies in the following.

Theorem 1 *For any stationary source $Q \in \mathcal{M}_s(\Omega)$*

(a) The limit in (5) exists.

(b) For any sequence $\{B_n\}$ of elements of $\mathcal{R}_\Box$ satisfying $D(B_n) \to \infty$ we have

$$U(l, Q) = \lim_{n\to\infty} U(l, Q_{B_n}) = \inf_{\Delta\in\mathcal{R}_\Box} U(l, Q_\Delta). \tag{8}$$

Theorem 1 justifies the notion of scandictability as introduced in Definition 2 and substantiates the significance of this entity as the fundamental limit on prediction performance for stationary data arranged in a rectangular array, when any scheme for sequentially scanning the data is allowed. It tells us that the scandictability of a stationary data array is independent of the ratios between the edges of the rectangle confining the array when these become large. Furthermore, Theorem 1 assures us of the fact that if one's goal is to achieve the predictability (of any stationary source) to within some $\varepsilon > 0$ of a large rectangular box, it suffices to partition the data into rectangular non-overlapping blocks congruent to any Δ satisfying $U(l, Q_\Delta) \leq U(l, Q) + \varepsilon$. Finally, we note that by letting $l = l_{log}$ Theorem 1 and equation (7) recover what is known about the entropy of random fields, cf. [14, Theorem (15.12)], [16, Theorem 5.2.1].

Basically, the only property we rely upon for establishing Theorem 1 is the sub-additivity of $U(l, Q_V)$. Specifically, we have

Lemma 2 *For any $Q \in \mathcal{M}(\Omega)$ and $V, V' \in \mathcal{V}$*

$$|V \cup V'| U(l, Q_{V \cup V'}) \leq |V| U(l, Q_V) + |V'| U(l, Q_{V'}). \tag{9}$$

Proof: Note first that if $V \cap V' \neq \emptyset$ we can take, instead of V', $\tilde{V} = V' \backslash V \stackrel{\text{def}}{=} V' \cap V^c$. The validity of the lemma for disjoint subsets, together with the obvious fact that $|\tilde{V}| U(l, Q_{\tilde{V}}) \leq |V'| U(l, Q_{V'})$, would imply the lemma for V, V'. We can therefore assume that $V \cap V' = \emptyset$. By the definition of $U(l, Q_{V \cup V'})$ it will clearly suffice to establish the existence of $(\Psi, F) \in \mathcal{S}(V \cup V')$ for which

$$E_Q L_{(\Psi,F)}(X(B)) \leq |V| U(l, Q_V) + |V'| U(l, Q_{V'}). \tag{10}$$

But this is easy: take $(\Psi, F) \in \mathcal{S}(V \cup V')$ to be the scandictor obtained by concatenating $(\Psi, F)^*_V$ and $(\Psi, F)^*_{V'}$ (i.e., the scheme which scandicts the set of sites V according to $(\Psi, F)^*_V$ and then the set V' according to $(\Psi, F)^*_{V'}$), where we let $(\Psi, F)^*_V \in \mathcal{S}(V)$ denote the scandictor achieving the infimum in (4)[3]. □

The relevance of sub-additivity to establishing the existence of a limit is manifested in the following lemma.

Lemma 3 *Let $f : (\mathbb{Z}_+)^m \to [0, \infty)$ be sub-additive separately in each of its arguments, i.e., for all $1 \leq i \leq m$, $a_1, \cdots, a_m, b_i, c_i \in \mathbb{Z}_+$*

$$f(a_1, \ldots, b_i + c_i, \ldots, a_m) \leq f(a_1, \ldots, b_i, \ldots, a_m) + f(a_1, \ldots, c_i, \ldots, a_m). \tag{11}$$

Then for every $\{a^{(n)} = (a_1^{(n)}, \ldots, a_m^{(n)})\}_{n \geq 1}$ with $\left(\min_{1 \leq i \leq m} a_i^{(n)}\right) \to \infty$

$$\lim_{n \to \infty} \frac{f(a^{(n)})}{\prod_{i=1}^m a_i^{(n)}} = \inf_{a \in (\mathbb{Z}_+)^m} \frac{f(a)}{\prod_{i=1}^m a_i}. \tag{12}$$

[3]If the infimum is not achieved take any ε-achiever and the proof carries through.

The proof (cf., e.g., [16, Lemma 5.2.1]) is a straightforward generalization of that from the univariate case (cf., e.g., [17, Lemma 6.1.11]).

Proof of Theorem 1: Fix $Q \in \mathcal{M}_s(\Omega)$. Since the first item follows from the second item by taking $B_n = V_n$, it will suffice to establish the fact that for any sequence $\{B_n\}$ of elements of $\mathcal{R}_\square$ satisfying $D(B_n) \to \infty$

$$\lim_{n\to\infty} U(l, Q_{B_n}) = \inf_{\Delta \in \mathcal{R}_\square} U(l, Q_\Delta). \tag{13}$$

By stationarity of Q it will suffice to restrict attention to $\mathcal{R}^0_\square$, namely, to prove that for any $\{B_n\} \subseteq \mathcal{R}^0_\square$ with $D(B_n) \to \infty$

$$\lim_{n\to\infty} U(l, Q_{B_n}) = \inf_{\Delta \in \mathcal{R}^0_\square} U(l, Q_\Delta). \tag{14}$$

To this end define $f : (\mathbb{Z}_+)^2 \to [0, \infty)$ by

$$f(a_1, a_2) \stackrel{\text{def}}{=} a_1 a_2 U(l, Q_{V(a_1,a_2)}), \tag{15}$$

where $V(a_1, a_2) \stackrel{\text{def}}{=} \mathbb{Z}^2 \cap ([0, a_1) \times [0, a_2))$. The sub-additivity of f is a direct consequence of Lemma 2 and the stationarity of Q. The proof is completed by an appeal to Lemma 3. $\square$

3 The Case $A = P = \mathbb{R}$

We dedicate this section to the case where the source alphabet and the predictions are real-valued. Furthermore, we shall focus on the case where the loss function is of the form $l(F, y) = \rho(y - F)$, where the function $\rho(z)$ is monotonically increasing for $z > 0$, monotonically decreasing for $z < 0$, and $\rho(0) = 0$. With a slight abuse of notation we shall write $U(\rho, Q_B)$ and $U(\rho, Q)$ for $U(l, Q_B)$ and $U(l, Q)$ respectively when $l(F, y) = \rho(y - F)$. We assume that $\rho(\cdot)$ is sufficiently "steep" in the sense that $\int e^{-s\rho(z)}dz < \infty$ for every $s > 0$ and, following [9], we define the *log-moment generating function* associated with the loss function ρ by

$$\lambda_\rho(s) = -\log\left[\int e^{-s\rho(z)}dz\right], \quad s > 0, \tag{16}$$

and the *one-sided Fenchel-Legendre transform* of $\lambda_\rho(\cdot)$ by

$$\gamma_\rho(d) = \inf_{s>0}[sd - \lambda_\rho(s)], \quad d > 0. \tag{17}$$

As remarked in [9], the function $\gamma_\rho(d)$ can be interpreted as the differential entropy associated with the probability density function

$$q_{\rho,s}(z) = e^{-s\rho(z)+\lambda_\rho(s)}, \tag{18}$$

where s is tuned so that $E_{\rho,s}\rho(Z) = d$, $E_{\rho,s}$ being the expectation operation w.r.t. $q_{\rho,s}$. For a reason that will be clear from the proof of the first item of Proposition 4 below (cf., in particular, equation (A.3) of Appendix A), we refer to $q_{\rho,s}(\cdot)$ as a *maximum-entropy distribution w.r.t. ρ.* It can be seen that $\gamma_\rho(d)$ is strictly monotonically increasing and concave for $d \geq 0$ and therefore the inverse function $\gamma_\rho^{-1}(\cdot)$ exists and is continuous. Two additional important aspects of $\gamma_\rho(d)$, which will be of later use, are encapsulated in the following proposition, whose proof is deferred to the appendix.

Proposition 4 *1. For any probability density function $q(\cdot)$*

$$\gamma_\rho(E_q\rho(Z)) \geq H(q), \tag{19}$$

with equality if and only if $q(\cdot) = q_{\rho,s}(\cdot)$ for some $s > 0$.

2. For all $n \geq 1$ and all $d \geq 0$

$$\frac{1}{n}\log \mathrm{Vol}\left(\left\{e_1^n : \sum_{i=1}^n \rho(e_i) \leq nd\right\}\right) \leq \frac{\log 2}{n} + \gamma_\rho(d). \tag{20}$$

3.

$$\lim_{n\to\infty}\frac{1}{n}\log \mathrm{Vol}\left(\left\{e_1^n : \sum_{i=1}^n \rho(e_i) \leq nd\right\}\right) = \gamma_\rho(d). \tag{21}$$

A Volume-Preserving Injections

In this subsection we make the observation that the map from the data array to the sequence of prediction errors associated with any (sufficiently smooth) scandictor is one to one and volume-preserving. As will be seen, this fact is key to the derivation of lower bounds on scandiction performance based on volume considerations.

For any $B \in \mathcal{V}$ let $\mathcal{S}_D(B)$ denote the subset of $\mathcal{S}(B)$ consisting of those scandictors which have a predictor $F = \{F_t : \mathbb{R}^{t-1} \to \mathbb{R}\}_{t=1}^{|B|}$ consisting of functions that are continuous and have continuous first derivatives. We shall let $U_D(\rho, Q_B)$ and $U_D(\rho, Q)$ be defined analogously to $U(\rho, Q_B)$ and $U(\rho, Q)$ of Definition 2, with the only difference that the infimum for defining $U_D(\rho, Q_B)$ is taken over $\mathcal{S}_D(B)$, instead of over $\mathcal{S}(B)$ as in the right side of equation (4). Theorem 1 is easily verified to hold for $U_D(\rho, Q_B)$ and $U_D(\rho, Q)$ as well. Similarly as with the quantities of Definition 2, we shall sometimes write $U_D(\rho, X(B))$ and $U_D(\rho, X)$ when the underlying distributions are clear. Note that in non-pathological cases, when the components of $X(B)$ are continuous-valued and the conditional distribution of X_i, $i \in B$, given the values of $X(B)$ at other sites is a continuous functional of these values, we have $U_D(\rho, X(B)) = U(\rho, X(B))$ and $U_D(\rho, X) = U(\rho, X)$. Note further that even when the scandictor achieving $U(\rho, X(B))$ is not a member of $\mathcal{S}_D(B)$, it is enough that it can be approximated arbitrarily well by members of $\mathcal{S}_D(B)$ in order for $U_D(\rho, X(B)) = U(\rho, X(B))$. One important example of a random field for which $U_D(\rho, X(B)) = U(\rho, X(B))$ for all $B \in \mathcal{V}$ and $U_D(\rho, X) = U(\rho, X)$ is the Gaussian field of Section 4.

Let now, for any $B \in \mathcal{V}$ and any scandictor $(\Psi, F) \in \mathcal{S}(B)$, the transformation $T_{(\Psi,F)} : \mathbb{R}^B \to \mathbb{R}^{|B|}$ be defined by

$$T_{(\Psi,F)}(x(B)) = \left(x(\Psi_1) - F_1, x(\Psi_2) - F_2, \ldots, x(\Psi_{|B|}) - F_{|B|}\right), \tag{22}$$

where Ψ_t and F_t on the right side of (22) are, respectively, the t-th site and t-th prediction associated with the scandictor (Ψ, F) when operating on $x(B)$. In words, $T_{(\Psi,F)}$ maps $x(B)$ into the sequence of prediction errors incurred when the scandictor (Ψ, F) operates on $x(B)$. For any $B \in \mathcal{V}$ we extend the notion of volume to $\mathbb{R}^B$ in the trivial way: order the sites of B arbitrarily and identify any $x(B) \in \mathbb{R}^B$ with the corresponding point in $\mathbb{R}^{|B|}$. A measurable map $T : \mathbb{R}^B \to \mathbb{R}^{|B|}$ will be said to be *volume-preserving* if $\mathrm{Vol}(G) = \mathrm{Vol}(T(G))$ for all measurable G.

Theorem 5 *For any $B \in \mathcal{V}$ and any scandictor $(\Psi, F) \in \mathcal{S}_D(B)$, the transformation $T_{(\Psi,F)} : \mathbb{R}^B \to \mathbb{R}^{|B|}$ defined by equation (22) is one-to-one and volume-preserving.*

Proof of Theorem 5: We assume a fixed $B \in \mathcal{V}$ and $(\Psi, F) \in \mathcal{S}_D(B)$ throughout the proof. The mapping $T_{(\Psi,F)}$ can be decomposed as follows. Let $T_\Psi : \mathbb{R}^B \to \mathbb{R}^{|B|}$ be defined by

$$T_\Psi(x(B)) = \big(x(\Psi_1), x(\Psi_2), \ldots, x(\Psi_{|B|})\big) \tag{23}$$

and let $T_F : \mathbb{R}^{|B|} \to \mathbb{R}^{|B|}$ be defined by

$$T_F(y_1^{|B|}) = \Big(y_1 - F_1, y_2 - F_2(y_1), \ldots, y_{|B|} - F_{|B|}(y_1^{|B|-1})\Big) . \tag{24}$$

Clearly

$$T_{(\Psi,F)} = T_F \circ T_\Psi, \tag{25}$$

so it suffices to show that both T_F and T_Ψ are one-to-one and volume-preserving. To this end note first that T_F is clearly one-to-one as, given the sequence of prediction errors (the right side of (24)), assuming the predictor F is known, the source sequence $y_1^{|B|}$ is uniquely determined. As for the volume-preservation property of this transformation, it is easy to see that the associated Jacobian (which exists as, by the hypothesis, $(\Psi, F) \in \mathcal{S}_D(B)$) is a lower-triangular matrix with diagonal entries which are all equal to 1 (for all values of $y_1^{|B|}$). Hence, the determinant of the Jacobian of this mapping equals unity everywhere, which implies that T_F is volume-preserving.

Moving on to consider T_Ψ, note first that T_Ψ is obviously measurable (by the measurability of the mappings defining Ψ) and one-to-one as knowledge of the values observed along any scan of the sites in B uniquely determines $x(B)$. To establish the volume-preservation property of T_Ψ we define a *permutation* of B as any map ϕ from $\{1, 2, \ldots, |B|\}$ onto B. Letting Φ_B denote the class of all $|B|!$ permutations of B, the following two simple observations can be made:

1. For any $\phi \in \Phi_B$, the mapping $T_\phi : \mathbb{R}^B \to \mathbb{R}^{|B|}$, defined by

$$T_\phi(x(B)) = (x(\phi(1)), x(\phi(2)), \ldots, x(\phi(|B|))) , \tag{26}$$

is volume preserving (as it corresponds to a relabelling of the axes).

2. For each $x(B) \in \mathbb{R}^B$ there exists a unique $\phi \in \Phi_B$ such that

$$T_\Psi(x(B)) = T_\phi(x(B)). \tag{27}$$

Let now $G_B \subseteq \mathbb{R}^B$ be an arbitrary Borel set. For each $\phi \in \Phi_B$ define $G_B^\phi = \{x(B) \in G_B : T_\Psi(x(B)) = T_\phi(x(B))\}$ (note that G_B^ϕ is Borel by the measurability of the mappings defining Ψ). By the above second simple observation, $\{G_B^\phi\}_{\phi\in\Phi_B}$ is a disjoint partition of G_B, i.e.

$$\bigcup_{\phi\in\Phi_B} G_B^\phi = G_B \quad \text{and} \quad G_B^\phi \cap G_B^{\phi'} = \emptyset \ \ \forall \phi \neq \phi'. \tag{28}$$

Consequently, we have

$$\begin{aligned}
\mathrm{Vol}\left(T_\Psi(G_B)\right) &= \mathrm{Vol}\left(T_\Psi\left(\bigcup_{\phi\in\Phi_B} G_B^\phi\right)\right) && (29)\\
&= \mathrm{Vol}\left(\bigcup_{\phi\in\Phi_B} T_\Psi\left(G_B^\phi\right)\right) && (30)\\
&= \sum_{\phi\in\Phi_B} \mathrm{Vol}\left(T_\Psi\left(G_B^\phi\right)\right) && (31)\\
&= \sum_{\phi\in\Phi_B} \mathrm{Vol}\left(T_\phi\left(G_B^\phi\right)\right) && (32)\\
&= \sum_{\phi\in\Phi_B} \mathrm{Vol}\left(G_B^\phi\right) && (33)\\
&= \mathrm{Vol}\left(G_B\right), && (34)
\end{aligned}$$

where the measurability of T_Ψ and the fact that G_B as well as the G_B^ϕ-s are Borel guarantee that all quantities in equations (29) - (31) are well-defined. Equation (30) follows from the facts that the sets in $\{G_B^\phi\}_{\phi\in\Phi_B}$ are disjoint and that T_Ψ is one-to-one. Equation (31) follows from the fact that T_Ψ is one-to-one and, hence, the sets in $\left\{T_\Psi\left(G_B^\phi\right)\right\}_{\phi\in\Phi_B}$ are disjoint. Equation (32) follows from the definition of the sets G_B^ϕ and equation (33) follows from the first simple observation above. □

Remark: As is clear from the above proof, the one-to-one property holds for *any* scandictor $(\Psi, F) \in \mathcal{S}(B)$. As for volume preservation, the condition $(\Psi, F) \in \mathcal{S}_D(B)$ allowed for the simple argument based on evaluation of the Jacobian of the map T_F. With a somewhat more elaborate argument it can be shown that it is enough, for example, that the functions defining the predictor associated with (Ψ, F) be piecewise differentiable.

Note that we can, conversely, look at $T^{-1}_{(\Psi,F)}$, the inverse transformation of $T_{(\Psi,F)}$, i.e., the transformation taking the prediction error sequence associated with the scandictor (Ψ, F) into the original data array $x(B)$. More specifically, $T^{-1}_{(\Psi,F)} : \mathbb{R}^{|B|} \to \mathbb{R}^B$ is given as follows: For any $W = (W_1, W_2, \ldots, W_{|B|}) \in \mathbb{R}^{|B|}$, if $x(B) = T^{-1}_{(\Psi,F)}(W)$ then $x(B)$ can be auto-regressively constructed using W as the innovations process according to: $x(\Psi_1) = F_1 + W_1$, $x\left(\Psi_2\left(x(\Psi_1)\right)\right) = F_2\left(x(\Psi_1)\right) + W_2$, and so forth. Note that Theorem 5 implies that for any scandictor $(\Psi, F) \in \mathcal{S}_D(B)$, the mapping $T^{-1}_{(\Psi,F)}$ is one-to-one and volume preserving. We thus have the following corollary to Theorem 5.

Corollary 6 *For any $B \in \mathcal{V}$ and any scandictor $(\Psi, F) \in \mathcal{S}_D(B)$ we have the following:*

1. *Let $X(B)$ be a discrete- or continuous-valued random field on B and let $N = (N_1, N_2, \dots, N_{|B|}) \in \mathbb{R}^{|B|}$ be the error sequence associated with the scandictor (Ψ, F) when operating on $X(B)$, i.e., $N = T_{(\Psi,F)}(X(B))$. Then*
$$H(N) = H(X(B)). \tag{35}$$

2. *Let $W = (W_1, W_2, \dots, W_{|B|}) \in \mathbb{R}^{|B|}$ be a discrete- or continuous-valued random vector and let $X(B)$ be a random field on B auto-regressively defined by $X(B) = T^{-1}_{(\Psi,F)}(W)$. Then*
$$H(X(B)) = H(W). \tag{36}$$

To derive another corollary note that Theorem 5 implies, in particular, that for all $B \in \mathcal{V}, (\Psi, F) \in \mathcal{S}_D(B)$

$$\mathrm{Vol}\left(\left\{e_1^{|B|} \in A^{|B|} : \sum_{i=1}^{|B|} \rho(e_i) \leq nd\right\}\right) = \mathrm{Vol}\left(\left\{x(B) \in A^B : L_{(\Psi,F)}(x(B)) \leq d\right\}\right). \tag{37}$$

Combined with the third item of Proposition 4, this implies that for large $B \in \mathcal{V}$ and any scandictor $(\Psi, F) \in \mathcal{S}_D(B)$, $\mathrm{Vol}\left(\left\{x(B) \in A^B : L_{(\Psi,F)}(x(B)) \leq d\right\}\right) \approx e^{|B|\gamma_\rho(d)}$. Thus, if $G_B \subseteq A^B$ is a set of a volume which is exponentially larger than $e^{|B|\gamma_\rho(d)}$, then $L_{(\Psi,F)}(x(B)) > d$ for all $x(B) \in G_B \setminus J_B$, where the volume of J_B is an exponentially negligible fraction of the volume of G_B. More formally, equations (21) and (37) lead to the following.

Corollary 7 *For any $a > \gamma_\rho(d)$ and $\varepsilon < a - \gamma_\rho(d)$, there exists n_0 such that: For all $B \in \mathcal{V}$ with $|B| \geq n_0$, for all $G_B \subseteq \mathbb{R}^B$ with $\mathrm{Vol}(G_B) \geq e^{|B|a}$, and any scandictor $(\Psi, F) \in \mathcal{S}_D(B)$*

$$L_{(\Psi,F)}(x(B)) > d \quad \forall x(B) \in G_B \setminus J_B, \tag{38}$$

where
$$\frac{\mathrm{Vol}(J_B)}{\mathrm{Vol}(G_B)} \leq e^{-|B|\varepsilon}. \tag{39}$$

Corollary (7) is an "individual-sequence" type of result which gives a lower bound on scandiction loss for "most" sequences in G_B. We now progress to derive a result for the probabilistic setting. For future reference, we first state the following, which is a direct consequence of equation (37) and the second item of Proposition 4.

Corollary 8 *For any $B \in \mathcal{V}$ and any scandictor $(\Psi, F) \in \mathcal{S}_D(B)$*

$$\frac{1}{|B|} \log \mathrm{Vol}\left(\left\{x(B) \in \mathbb{R}^B : L_{(\Psi,F)}(x(B)) \leq d\right\}\right) \leq \frac{\log 2}{|B|} + \gamma_\rho(d). \tag{40}$$

We can now state the following result, whose main significance is in the introduction of single-letter upper and lower bounds on scandiction performance in the probabilistic setting.

Theorem 9 *Let* $\{B_n\}_{n\geq 1}$*,* $B_n \in \mathcal{V}$*, be an arbitrary sequence satisfying* $|B_n| \to \infty$*. Let* $\mathbf{W} = \{W_i\}_{i\geq 1}$ *be a sequence of independent continuous random variables, where the density function of* W_i *is* $f_{W_i}(\cdot)$*,* $\mathrm{Var}(\log f_{W_i}(W_i)) \leq C < \infty$*, and for which there exists a continuous random variable* W_* *such that*

$$\frac{1}{|B_n|}\sum_{i=1}^{|B_n|} H(W_i) \to H(W_*) \tag{41}$$

and

$$\frac{1}{|B_n|}\sum_{i=1}^{|B_n|} E\rho(W_i) \to E\rho(W_*). \tag{42}$$

Let further $\{(\Psi^{(n)}, F^{(n)})\}_{n\geq 1}$ *be an arbitrary sequence such that* $(\Psi^{(n)}, F^{(n)}) \in \mathcal{S}_D(B_n)$*. Finally let, for each* n*,* $X(B_n)$ *be the random field on* B_n *which is auto-regressively generated by the scandictor* $(\Psi^{(n)}, F^{(n)})$ *with the innovation process* $W^{(n)} = (W_1, W_2, \ldots, W_{|B_n|})$*, i.e.*

$$X(B_n) = T^{-1}_{(\Psi^{(n)}, F^{(n)})}(W^{(n)}). \tag{43}$$

Then

$$\gamma_\rho^{-1}(H(W_*)) \leq \liminf_{n\to\infty} U_D(\rho, X(B_n)) \leq \limsup_{n\to\infty} U_D(\rho, X(B_n)) \leq E\rho(W_*). \tag{44}$$

The upper bound in (44) is easily seen to be attainable by employing the scandictor $(\Psi^{(n)}, F^{(n)})$ from which $X(B_n)$ was generated. To see why the lower bound in (44) should hold note that if d is such that $\gamma_\rho(d) < H(W_*)$ then, by Corollary 8, there exists $\varepsilon > 0$ such that for all n sufficiently large and any scandictor $(\Psi, F) \in \mathcal{S}_D(B_n)$, $\mathrm{Vol}\left(\left\{x(B_n) \in \mathbb{R}^{B_n} : L_{(\Psi,F)}(x(B_n)) \leq d\right\}\right) \leq e^{|B_n|(H(W_*)-\varepsilon)}$. Since $T^{-1}_{(\Psi^{(n)},F^{(n)})}(\cdot)$ is volume-preserving, this implies that in order for $L_{(\Psi,F)}(X(B_n)) \leq d$, the innovation vector $W^{(n)}$ through which $X(B_n)$ was defined must lie in a set whose volume is $\leq e^{|B_n|(H(W_*)-\varepsilon)}$. But the fact that (41) holds implies (by an AEP-type argument) that the probability of this being the case is arbitrarily small for sufficiently large n. This line of argumentation leads to $\liminf_{n\to\infty} U_D(\rho, X(B_n)) \geq d$ whenever $d < \gamma_\rho^{-1}(H(W_*))$, which implies the left inequality in (44). This is the essential idea behind the formal proof below. Prior to the proof of Theorem 9, we note the following, which is a direct consequence of Theorem 9 and the first item of proposition 4.

Corollary 10 *Let the setting of Theorem 9 hold and suppose further that* $f_{W_*}(\cdot) = q_{\rho,s}(\cdot)$ *for some* $s > 0$*. Then*

$$\lim_{n\to\infty} U_D(\rho, X(B_n)) = E\rho(W_*). \tag{45}$$

Note that, in particular, Corollary 10 tells us that for large $B \in \mathcal{V}$, if $X(B)$ is auto-regressively generated via any scandictor $(\Psi, F) \in \mathcal{S}_D(B)$ and the innovation process has independent components with a maximum-entropy distribution w.r.t. ρ, then the optimal scandictor for $X(B)$ is (Ψ, F) itself. It should be noted that, though seemingly plausible and even intuitive, this is a non-trivial fact. Indeed, the question of optimally scanning the data for prediction is not trivial even in the one-dimensional case. As one simple example, consider the problem of scandicting the binary-valued first-order symmetric Markov source under Hamming loss. It is a priori unclear whether chronological ordering is better than, for example, scanning first the even indices and then the odd ones. As will be shown in Section 5, the finite-alphabet version of Corollary 10 (Corollary 19) implies that it is the chronological ordering which is optimal in this case. The chronological ordering, however, is not always optimal, as the following example shows: Let $R_1, R_2, \dots$ be i.i.d. Bernoulli(1/2) and construct the process $X = X_1, X_2, \dots$, where $X_i \in A = \{0,1\}^2$ via $X_i = (R_i, R_{i+1})$. X is clearly a first-order Markov process. Consider now the problem of scandicting X when the prediction space $P = \{0,1\}$ and the loss function l is given by

$$l(P, a) = P \oplus a_2, \qquad a = (a_1, a_2) \in A, \tag{46}$$

where $\oplus$ denotes addition modulo-2. For this case it is easy to see that chronological scandiction gives an average loss of 1/2 while scanning first the even indices and then the odd ones gives an average loss of 1/4. The loss function in (46) does not satisfy the conditions required for our results, so there is no contradiction. What this "counterexample" does demonstrate is that the question of finding the optimal scandiction scheme is a non-trivial one even for relatively simple one-dimensional processes and, a fortiori, for multi-dimensional random fields.

Two concrete examples for the significance of Corollary 10 are:

Gaussian innovation and squared-error loss: If W_* in Theorem 9 is Gaussian, with variance σ^2, and $\rho(z) = z^2$, Corollary 10 gives

$$\lim_{n\to\infty} U_D(\rho, X(B_n)) = \sigma^2. \tag{47}$$

This fact will play a key role in the proof of the main result of Section 4.

Laplacian innovation and absolute-error loss: If W_* in Theorem 9 is Laplacian, i.e. $f_{W_*}(z) = \frac{1}{2\beta} e^{-|z|/\beta}$ for some $\beta > 0$ so that $E|W_*| = \beta$, and $\rho(z) = |z|$, Corollary 10 gives

$$\lim_{n\to\infty} U_D(\rho, X(B_n)) = \beta. \tag{48}$$

Proof of Theorem 9: To establish the upper bound on the limit supremum in (44) note that for all n the normalized cumulative loss of the scandictor $(\Psi^{(n)}, F^{(n)})$ when applied to $X(B_n)$ is given by

$$L_{(\Psi^{(n)}, F^{(n)})}(X(B_n)) = \frac{1}{|B_n|} \sum_{i=1}^{|B_n|} \rho(W_i). \tag{49}$$

Thus

$$\begin{aligned} \limsup_{n\to\infty} U_D(\rho, X(B_n)) &\leq \limsup_{n\to\infty} E L_{(\Psi^{(n)}, F^{(n)})}(X(B_n)) \\ &= \limsup_{n\to\infty} E \frac{1}{|B_n|} \sum_{i=1}^{|B_n|} \rho(W_i) \\ &= E\rho(W_*), \end{aligned} \tag{50}$$

where the last equality follows from (42).

We now progress to establish the lower bound on the limit infimum in (44). To this end, fix an arbitrary $d < \gamma_\rho^{-1}(H(W_*))$ and an arbitrary $\delta > 0$. Let further $\varepsilon = \min\left\{\frac{H(W_*)-\gamma_\rho(d)}{2}, \delta/2\right\}$ and let $A_\varepsilon^{(m)}$ denote the ε-typical set with respect to $f_{\mathbf{W}}(\cdot)$ defined as follows:

$$A_\varepsilon^{(m)} = \left\{(w_1, w_2, \ldots, w_m) \in \mathbb{R}^m : \left|-\frac{1}{m}\log f_{\mathbf{W}}(w_1, w_2, \ldots, w_m) - H(W_*)\right| \le \varepsilon\right\}, \tag{51}$$

where $f_{\mathbf{W}}(w_1, w_2, \ldots, w_m) = \prod_{i=1}^m f_{W_i}(w_i)$. As is easily shown in Appendix B to follow from the hypotheses that $\mathrm{Var}\left(\log f_{W_i}(W_i)\right) \le C < \infty$, that (41) holds and that $|B_n| \to \infty$, for all $\varepsilon > 0$

$$\lim_{n\to\infty} \Pr\{W^{(n)} \in A_\varepsilon^{(|B_n|)}\} = 1. \tag{52}$$

This is the analogue for our setting of the standard ε-typicality result of the i.i.d. case (cf., e.g., item 1 of [18, Theorem 9.2.2]). Thus there exists $n_0(d, \delta)$ such that for all $n \ge n_0(d, \delta)$ and all $(\Psi, F) \in \mathcal{S}_D(B_n)$ we have

$$\begin{aligned}
& \Pr\{L_{(\Psi,F)}(X(B_n)) \le d\} \\
\le\ & \Pr\left(\{L_{(\Psi,F)}(X(B_n)) \le d\} \cap \{W^{(n)} \in A_\varepsilon^{(|B_n|)}\}\right) + \Pr\{W^{(n)} \notin A_\varepsilon^{(|B_n|)}\} \\
\le\ & \int_{\left\{\{w^{(n)}: L_{(\Psi,F)}(x(B_n)) \le d\} \cap A_\varepsilon^{(|B_n|)}\right\}} f_{\mathbf{W}}(w^{(n)}) dw^{(n)} + \delta/2 && (53) \\
\le\ & \int_{\left\{\{w^{(n)}: L_{(\Psi,F)}(x(B_n)) \le d\} \cap A_\varepsilon^{(|B_n|)}\right\}} e^{-|B_n|(H(W_*)-\varepsilon)} dw^{(n)} + \delta/2 \\
=\ & \mathrm{Vol}\left(\{w^{(n)} : L_{(\Psi,F)}(x(B_n)) \le d\} \cap A_\varepsilon^{(|B_n|)}\right) e^{-|B_n|(H(W_*)-\varepsilon)} + \delta/2 \\
\le\ & \mathrm{Vol}\left(\{w^{(n)} : L_{(\Psi,F)}(x(B_n)) \le d\}\right) e^{-|B_n|(H(W_*)-\varepsilon)} + \delta/2 \\
=\ & \mathrm{Vol}\left(\{x(B_n) : L_{(\Psi,F)}(x(B_n)) \le d\}\right) e^{-|B_n|(H(W_*)-\varepsilon)} + \delta/2 && (54) \\
\le\ & \exp\left\{|B_n|\left(\frac{\log 2}{|B_n|} + \gamma_\rho(d) - H(W_*) + \varepsilon\right)\right\} + \delta/2 && (55) \\
\le\ & \exp\left\{-|B_n|\left(\frac{H(W_*) - \gamma_\rho(d)}{2} - \frac{\log 2}{|B_n|}\right)\right\} + \delta/2 && (56) \\
\le\ & \delta, && (57)
\end{aligned}$$

where the inequality in (53) follows by taking $n_0(d, \delta)$ sufficiently large so that

$$\Pr\{W^{(n)} \notin A_\varepsilon^{(|B_n|)}\} \le \delta/2 \quad \forall n \ge n_0(d, \delta), \tag{58}$$

which is possible by (52). Equality in (54) follows from the fact that the transformation taking $w^{(n)}$ into $x(B_n)$, namely $T^{-1}_{(\Psi^{(n)}, F^{(n)})}$, is volume-preserving (Theorem 5). Inequality (55) follows from Corollary 8. Inequality (56) follows since $\varepsilon \le \frac{H(W_*)-\gamma_\rho(d)}{2}$. Inequality (57) follows by taking a sufficiently large $n_0(d, \delta)$ such that, in addition to satisfying (58), the first term in (56) is upper bounded by $\delta/2$ for all $n \ge n_0(d, \delta)$. This is possible since $|B_n| \to \infty$. Consequently, for all $n \ge n_0(d, \delta)$ and all $(\Psi, F) \in \mathcal{S}_D(B_n)$

$$EL_{(\Psi,F)}(X(B_n)) \quad = \quad E[L_{(\Psi,F)}(X(B_n)) | L_{(\Psi,F)}(X(B_n)) \le d] \cdot \Pr\{L_{(\Psi,F)}(X(B_n)) \le d\}$$

$$
\begin{aligned}
&+E[L_{(\Psi,F)}(X(B_n))|L_{(\Psi,F)}(X(B_n)) > d] \cdot \big(1 - \Pr\{L_{(\Psi,F)}(X(B_n)) \leq d\}\big) \\
\geq \;& 0 + d\big(1 - \Pr\{L_{(\Psi,F)}(X(B_n)) \leq d\}\big) \\
\geq \;& d(1-\delta). \qquad (59)
\end{aligned}
$$

The fact that the right side does not depend on $(\Psi, F) \in \mathcal{S}_D(B_n)$ implies that for all $n \geq n_0(d, \delta)$

$$U_D(\rho, X(B_n)) \geq d(1-\delta) \tag{60}$$

which, in turn, implies

$$\liminf_{n\to\infty} U_D(\rho, X(B_n)) \geq d(1-\delta). \tag{61}$$

The arbitrariness of $d < \gamma_\rho^{-1}(H(W_1))$ and $\delta > 0$ on the right side of (61) completes the proof. □

In the course of the preceding proof (cf., in particular, the inequalities leading to (57)) we have, in fact, established the following result, from which equation (44) was easily derived.

Corollary 11 *Let the setting of Theorem 9 hold. For any* $d < \gamma_\rho^{-1}(H(W_*))$

$$\lim_{n\to\infty} \sup_{(\Psi,F)\in\mathcal{S}_D(B_n)} \Pr\{L_{(\Psi,F)}(X(B_n)) \leq d\} = 0 \tag{62}$$

and for any $d > E\rho(W_*)$

$$\lim_{n\to\infty} \sup_{(\Psi,F)\in\mathcal{S}_D(B_n)} \Pr\{L_{(\Psi,F)}(X(B_n)) \leq d\} = 1. \tag{63}$$

In fact, the convergence in equations (62) and (63) is exponentially (in $|B_n|$) rapid (because the convergence of $\Pr\{W^{(n)} \notin A_\varepsilon^{(|B_n|)}\}$ to 0 is).

B An Alternative Route to a Converse on Scandictability Performance

The observation that $T_{(\Psi,F)}$ is measure-preserving for any $(\Psi, F) \in \mathcal{S}_D(B)$ was the key to the results of the previous subsection. When the scandictor, in $\mathcal{S}(B)$, is not a member of $\mathcal{S}_D(B)$ however, the volume-preservation property may no longer hold. In this subsection, we take a somewhat different route for the derivation of lower bounds, utilizing MDL-type lower bounds [6, 7, 8]. We shall use an approach which was applied in [9, subsection III.A] in the context of prediction of time series. This will lead, in particular, to lower bounds on scandiction performance for scandictors which are not necessarily members of $\mathcal{S}_D(B)$.

Let $\{f_\theta, \theta \in \Lambda\}$ be a general class of information sources emitting continuous-valued random variables. Suppose that the source alphabet $\mathcal{I}$ is some bounded interval. With a customary abuse of notation, we shall let $f_\theta(y^n) = f_\theta(y_1, \ldots, y_n)$ denote the probability density function (PDF) of $Y^n = Y_1, \ldots Y_n$ when emitted by f_θ. Let ω be an arbitrary probability measure on Λ (which is equipped with a σ-algebra) and assume that $\{f_\theta, \theta \in \Lambda\}$ is such that $f_\theta(y^n)$ is a measurable function of θ for every $y^n \in \mathcal{I}^n$. Following [7], we shall refer to this measurability assumption as *Assumption A*. Let now $\Theta_1, \ldots, \Theta_M$ denote M independent random points selected from Λ under ω. Suppose, without loss of generality, that Θ_1 has generated Y^n. Let $\overline{P}_e(M,n,\omega)$ denote the average probability of error in the random coding sense; namely, the probability that $\Theta_1,\ldots,\Theta_M$ and Y^n are such that for some $2 \le i \le M$, $f_{\Theta_i}(Y^n) \ge f_{\Theta_1}(Y^n)$. Mathematically,

$$\overline{P}_e(M,n,\omega) = 1 - \int_\Lambda \omega(d\theta) \int_{y^n \in \mathcal{I}^n} dy^n f_\theta(y^n) \left[1 - \omega\left\{\theta' : f_{\theta'}(y^n) \ge f_\theta(y^n)\right\}\right]^{M-1}. \tag{64}$$

Now let $M(n,\delta,\omega)$ be the largest integer M such that

$$\overline{P}_e(M,n,\omega) \le \delta \tag{65}$$

and finally, define the *random coding δ-capacity with respect to ω* as

$$C_R(n,\delta,\omega) \stackrel{\text{def}}{=} \log M(n,\delta,\omega). \tag{66}$$

Note that $\overline{P}_e(M,n,\omega)$ can be upper bounded by the union bound

$$\overline{P}_e(M,n,\omega) \le (M-1) \cdot \int_\Lambda \omega(d\theta) \int_{y^n \in \mathcal{I}^n} dy^n f_\theta(y^n) \omega\left\{\theta' : f_{\theta'}(y^n) \ge f_\theta(y^n)\right\} \tag{67}$$

so clearly $M(n,\delta,\omega)$ is lower bounded by the largest integer M for which the right side of (67) is less than δ, namely,

$$M(n,\delta,\omega) \ge \frac{\delta}{\int_\Lambda \omega(d\theta) \int_{y^n \in \mathcal{I}^n} dy^n f_\theta(y^n) \omega\left\{\theta' : f_{\theta'}(y^n) \ge f_\theta(y^n)\right\}}. \tag{68}$$

Though the precise expression for $C_R(n,\delta,\omega)$ is hard to obtain, the lower bound in (68) is easier to work with in many cases (and will be made use of in the sequel). Let further $E_\theta\{\cdot\}$ denote the mathematical expectation with respect to f_θ and let $H_\theta(Y^n)$ denote the differential entropy of Y^n under f_θ. The following is one of the main results of [7].

Theorem 12 (Theorem 3 of [7]) *Let $\{f_\theta, \theta \in \Lambda\}$ satisfy Assumption A and let ω be any probability measure on Λ. Then for every $\varepsilon > 0$, $0 < \delta < 1$, every PDF $q(\cdot)$, and every $n \ge 1$*

$$E_\theta[-\log q(Y^n)] \ge H_\theta(Y^n) + (1-\varepsilon) C_R(n,\delta,\omega) \tag{69}$$

for every $\theta \in \Lambda$ except for a subset of points $\Lambda_s \subseteq \Lambda$ such that

$$\omega(\Lambda_s) \le \frac{\delta C_R(n,\delta,\omega) + 2}{\varepsilon C_R(n,\delta,\omega)}. \tag{70}$$

Theorem 3 of [7], in fact, is formulated for the discrete case, where $\{f_\theta, \theta \in \Lambda\}$ are finite-alphabet sources and q is, correspondingly, a PMF. The proof of the continuous version presented above is easily seen to carry over (under our Assumption A and the assumption that the source alphabet is a bounded interval) from the finite-alphabet case.

For $B \in \mathcal{V}$ Theorem 12 can now be applied to derive a lower bound on the attainable scandiction performance for "most" data arrays in a given subset of $\mathbb{R}^B$, of the type obtained in Corollary 7.

Specifically, let $x(B) \in \mathbb{R}^B$ be a deterministic ("individual") data-array indexed by the elements of B. Suppose that we observe a noisy version $Y(B) = x(B)+W(B)$, where $W(B)$ is a stochastic noise field with continuous-valued components. We will assume first that the components of $x(B)$ and $W(B)$ (and hence also of $Y(B)$) are bounded. We shall be interested in the attainable performance of an arbitrary scandictor $(\Psi, F) \in \mathcal{S}(B)$ when the underlying data-array $x(B)$ belongs to a certain subset G_B of $\mathbb{R}^B$. Let $E_{x(B)}L_{(\Psi,F)}(Y(B))$ denote the expected scandiction performance of $(\Psi, F) \in \mathcal{S}(B)$ on $Y(B)$ when the underlying data-array is $x(B)$. We further let $C_R\,(B, \delta, \omega)$ denote the random coding δ-capacity with respect to ω of the additive channel $Y(B) = x(B) + W(B)$ when the input is constrained to G_B. An application of Theorem 12, letting the clean data-array $x(B)$ play the role of θ, G_B the role of Λ, $Y(B)$ the role of Y^n, and $\mathrm{Vol}(\cdot)/\mathrm{Vol}(G_B)$ the role of $\omega(\cdot)$, gives: For every PDF $q(\cdot)$ that is independent of $x(B)$, we have

$$E_{x(B)}[-\log q(Y(B))] \geq H(W(B)) + (1-\varepsilon)C_R\left(B, \delta, \frac{\mathrm{Vol}(\cdot)}{\mathrm{Vol}(G_B)}\right) \tag{71}$$

for every $x(B) \in G_B$ except for a subset of points $J_B \subseteq G_B$ such that

$$\frac{\mathrm{Vol}(J_B)}{\mathrm{Vol}(G_B)} \leq \frac{\delta C_R\left(B, \delta, \frac{\mathrm{Vol}(\cdot)}{\mathrm{Vol}(G_B)}\right) + 2}{\varepsilon C_R\left(B, \delta, \frac{\mathrm{Vol}(\cdot)}{\mathrm{Vol}(G_B)}\right)}. \tag{72}$$

Taking a route similar to that taken in Section III.A of [9] (cf. derivation in equations (29)-(32) therein), for the given $(\Psi, F) \in \mathcal{S}(B)$, we now define a PDF on $\mathbb{R}^B$ as follows:

$$q(y(B)) = \int_0^\infty ds \nu(s) \left[\prod_{t=1}^{|B|} q_{\rho,s}\left([y(\Psi_t) - F_t(y(\Psi_1), \cdots, y(\Psi_{t-1}))]\right)\right], \tag{73}$$

where the Ψ_t-s and F_t-s on the right side are those associated with the scandictor $(\Psi, F) \in \mathcal{S}(B)$, $\nu(\cdot)$ is a locally bounded away from zero "prior" on s, and $q_{\rho,s}(\cdot)$ is the maximum-entropy distribution defined in (18). Note that for each s, the bracketed expression in the right side of (73) is a bona fide PDF and, consequently, so is $q(\cdot)$. Furthermore, according to the main result of [19] (cf. also equation (30) in [9]), $-\log q(y(B))$ can be approximated as follows:

$$\begin{aligned} -\log q(y(B)) &= |B| \cdot \inf_{s>0}\left[s \cdot L_{(\Psi,F)}(y(B)) - \lambda_\rho(s)\right] + \frac{1}{2}\log|B| + R(y(B)) \\ &= |B| \cdot \gamma_\rho\left(L_{(\Psi,F)}(y(B))\right) + \frac{1}{2}\log|B| + R(y(B)), \end{aligned} \tag{74}$$

where the remainder $R(y(B))$ is an increasing function of $L_{(\Psi,F)}(y(B))$. Since the components of $y(B)$ are assumed bounded, $R(y(B))$ is bounded as well by some constant R. Substituting into (71) implies that

$$|B| \cdot E_{x(B)}\left[\gamma_\rho\left(L_{(\Psi,F)}(Y(B))\right)\right] \geq H(W(B)) + (1-\varepsilon)C_R\left(B, \delta, \frac{\mathrm{Vol}(\cdot)}{\mathrm{Vol}(G_B)}\right) - \frac{1}{2}\log|B| - R \tag{75}$$

for all $x(B) \in G_B \setminus J_B$. The concavity of $\gamma_\rho(\cdot)$ allows to insert the expectation into the argument of $\gamma_\rho(\cdot)$ on the left side of (75) which gives

$$|B| \cdot \gamma_\rho\left(E_{x(B)}L_{(\Psi,F)}(Y(B))\right) \geq H(W(B)) + (1-\varepsilon)C_R\left(B, \delta, \frac{\mathrm{Vol}(\cdot)}{\mathrm{Vol}(G_B)}\right) - \frac{1}{2}\log|B| - R \tag{76}$$

for all $x(B) \in G_B \setminus J_B$. Narrowing down even further, assume henceforth that the components of $W(B)$ are i.i.d. and uniformly distributed on $[-\Delta/2, \Delta/2]$. To make the dependence explicit, we add the superscript Δ in the notation for expectation, thus writing $E^{\Delta}_{x(B)}\{\cdot\}$. For this case, we clearly have

$$H(W(B)) = |B| \log \Delta. \tag{77}$$

To get a more explicit handle on the right side of (76) for this case, we now lower bound $C_R\left(B, \delta, \frac{\mathrm{Vol}(\cdot)}{\mathrm{Vol}(G_B)}\right)$ as follows. Letting $f^{\Delta}_{x(B)}(\cdot)$ denote the PDF of $Y(B)$ when the underlying data-array is $x(B)$, it is clear that for any $x(B) \in G_B$ and any $y(B)$

$$\mathrm{Vol}\left(x'(B) : f^{\Delta}_{x'(B)}(y(B)) > f^{\Delta}_{x(B)}(y(B))\right) \leq \Delta^{|B|}. \tag{78}$$

Combining this with the lower bound (68) we obtain

$$C_R\left(B, \delta, \frac{\mathrm{Vol}(\cdot)}{\mathrm{Vol}(G_B)}\right) \geq \log\left[\frac{\delta \mathrm{Vol}(G_B)}{\Delta^{|B|}}\right]. \tag{79}$$

Substituting (77) and (79) into the right side of (76) gives

$$E^{\Delta}_{x(B)} L_{(\Psi,F)}(Y(B)) \geq \gamma_\rho^{-1}\left((1-\varepsilon)\frac{\log \mathrm{Vol}(G_B)}{|B|} + \varepsilon \log \Delta + (1-\varepsilon)\frac{\log \delta}{|B|} - \frac{\log |B|}{2|B|} - \frac{R}{|B|}\right) \tag{80}$$

for all $x(B) \in G_B \setminus J_B$. By maintaining a regime where $\varepsilon \to 0$, $\delta \to 0$, $|B| \to \infty$, $\Delta \to 0$, $\varepsilon \log \Delta \to 0$, $0 < a \leq \frac{1}{|B|}\log \mathrm{Vol}(G_B) \leq b < \infty$, $\delta \log\left[\frac{\delta e^{|B|a}}{\Delta^{|B|}}\right] \to \infty$, $\varepsilon \log\left[\frac{\delta e^{|B|a}}{\Delta^{|B|}}\right] \to \infty$ and $\delta/\varepsilon \to 0$, we have by equations (71) and (72) and the continuity of $\gamma_\rho^{-1}(\cdot)$ established the following.

Theorem 13 *$\forall \eta > 0, 0 < a < b < \infty$, $\exists n_0 = n_0(\eta, a, b)$ such that: $\forall B \in \mathcal{V}$ with $|B| \geq n_0$, $\forall G_B \in \mathbb{R}^B$ with $a \leq \frac{1}{|B|}\log \mathrm{Vol}(G_B) \leq b$, and $\forall (\Psi, F) \in \mathcal{S}(B)$*

$$E^{\eta}_{x(B)} L_{(\Psi,F)}(Y(B)) \geq \gamma_\rho^{-1}\left(\frac{\log \mathrm{Vol}(G_B)}{|B|}\right) - \eta \quad \forall x(B) \in G_B \setminus J_B, \tag{81}$$

where

$$\frac{\mathrm{Vol}(J_B)}{\mathrm{Vol}(G_B)} \leq \eta. \tag{82}$$

As opposed to the previous subsection, where the converse statements were valid for scandictors with a continuously differentiable predictor, Theorem 13 holds for an arbitrary scandictor. Note also that when $\eta > 0$ is small, $Y(B)$ on the left side of (81) is close, under sup-norm, to $x(B)$. One example of a way of exploiting this it to let $\mathcal{S}^K(B)$ denote the subset of $\mathcal{S}(B)$ consisting of all scandictors which are K-Lipschitz in the sense that

$$|L_{(\Psi,F)}(x(B)) - L_{(\Psi,F)}(x'(B))| \leq K\|x(B) - x'(B)\|_\infty \tag{83}$$

for all $x(B), x'(B) \in \mathbb{R}^B$. Note, for example, that any scandictor $(\Psi, F) \in \mathcal{S}(B)$ with a deterministic (non data-dependent) scan, such that the functions comprising F are K-Lipschitz, is a member of

$\mathcal{S}^K(B)$. Note also that when the underlying data-array is $x(B)$ and the components of $W(B)$ are $U[-\eta/2, \eta/2]$ then, with probability 1, $\|x(B) - Y(B)\|_\infty \leq \eta/2$ and, hence,

$$E^{\eta}_{x(B)} L_{(\Psi,F)}(Y(B)) \leq L_{(\Psi,F)}(x(B)) + K\eta/2 \quad \forall (\Psi, F) \in \mathcal{S}^K(B), x(B) \in \mathbb{R}^B. \tag{84}$$

We thus have the following corollary to Theorem 13.

Corollary 14 *$\forall \eta > 0, 0 < a < b < \infty$, $\exists n_0 = n_0(\eta, a, b)$ such that: $\forall B \in \mathcal{V}$ with $|B| \geq n_0$, $\forall G_B \in \mathbb{R}^B$ with $a \leq \frac{1}{|B|} \log \mathrm{Vol}(G_B) \leq b$, and $\forall K > 0, (\Psi, F) \in \mathcal{S}^K(B)$*

$$L_{(\Psi,F)}(x(B)) \geq \gamma_\rho^{-1}\left(\frac{\log \mathrm{Vol}(G_B)}{|B|}\right) - (K+2)\eta/2 \quad \forall x(B) \in G_B \setminus J_B, \tag{85}$$

where

$$\frac{\mathrm{Vol}(J_B)}{\mathrm{Vol}(G_B)} \leq \eta. \tag{86}$$

Note that similarly to Corollary 7, Corollary 14 is a purely "individual - sequence" statement. Where the former was valid for scandictors in $\mathcal{S}_D(B)$, the latter holds for those in $\mathcal{S}^K(B)$. Note also that Corollary (14) can be further specialized as follows (the details, which are similar as in the proof of Corollary 7, can be made precise and are only sketched here for brevity): For $B \in \mathcal{V}$ such that $|B|$ is large, for any scandictor $(\Psi_0, F_0) \in \mathcal{S}_D(B)$ we know, from the previous subsection, that $\mathrm{Vol}\left(\left\{x(B) \in \mathbb{R}^B : L_{(\Psi_0,F_0)}(x(B)) \leq d\right\}\right)$ is exponentially equivalent to $e^{|B|\gamma_\rho(d)}$. Hence, for large $|B|$, taking $G_B = \left\{x(B) \in \mathbb{R}^B : L_{(\Psi_0,F_0)}(x(B)) \leq d\right\}$ in Corollary 14 implies that (Ψ_0, F_0) is the optimal scandictor for the set G_B in the sense that there is no Lipschitz scandictor that can perform better than (Ψ_0, F_0) for most data-arrays in G_B. This is true because, while (Ψ_0, F_0) attains a scandiction error no larger than d for every $x(B) \in G_B$ (by definition), any alternative scandictor will have scandiction error essentially lower bounded by d (by inequality (85)) for all but a set of data-arrays whose volume is a negligible fraction of the volume of G_B.

To end this subsection we remark that the "individual-sequence" results developed above can be used to obtain results for the probabilistic setting, similarly as was done in the previous subsection. It should also be pointed out that the derivation of Theorem 13 and Corollary 14 was based on an application of Theorem 12 with the assignment $\omega(\cdot) = \mathrm{Vol}(\cdot)/\mathrm{Vol}(G_B)$. This gave an upper bound on the ratio between the volumes of the sets J_B and G_B. Other choices of $\omega(\cdot)$ can similarly give analogues of the above results with upper bounds on the ratio between the ω-measures of the sets J_B and G_B.

4 Scandictability of the Stationary Gaussian Field on $\mathbb{Z}^2$

We dedicate this section to the scandictability of the spatially stationary Gaussian field on $\mathbb{Z}^2$ with respect to the squared-error loss function. The main result and the analysis carries over to $\mathbb{Z}^d$, for any $d \geq 1$.

To fix notation, we recall here the basics regarding spectral representations of wide-sense (second-order) stationary processes. There are no fundamental differences between time-series and the multi-dimensional case. Let $X = \{X_t\}_{t\in\mathbb{Z}^d}$ be a Wide-Sense Stationary (w.s.s.) and centered process taking (in general) complex values: $X_t \in L^2$, $EX_t = 0$, $E(X_t\overline{X_s}) = R(t-s)$. For any $V \subseteq \mathbb{Z}^2$ let $\mathcal{H}(X(V))$ denote the closed span of $\{X_t\}_{t\in V}$, i.e., the smallest closed subspace which contains each X_t, $t \in V$ (under the scalar covariance product). For any $t \in \mathbb{Z}^2$ and $V \subseteq \mathbb{Z}^2$ we will let $\hat{X}_t(V)$ denote the projection of X_t onto $\mathcal{H}(V)$ (in other words, $\hat{X}_t(V)$ is the best linear predictor of X_t in terms of $\{X_{t'}\}_{t'\in V}$).

The extension of Herglotz's theorem [20, Section 4.3] to the multi-dimensional case dates at least as far back as [21], asserting the following representation of the covariance:

$$R(n) = \frac{1}{(2\pi)^d}\int_{[0,2\pi)^d} e^{i<n,\lambda>} G(d\lambda), \quad n \in \mathbb{Z}^d, \tag{87}$$

where G, the *spectral measure*, is a nonnegative and bounded measure over $[0, 2\pi)^d$.

A subset $S \subseteq \mathbb{Z}^2$ is called a *half plane* if:

$$S \text{ is closed to addition}, \quad S \cup (-S) = \mathbb{Z}^2, \quad S \cap (-S) = \{0\}. \tag{88}$$

A half plane S defines a total order relationship on $\mathbb{Z}^2$ via

$$(i, j) \leq (k, l) \Longleftrightarrow (k-i, l-j) \in S. \tag{89}$$

Examples for half planes include

$$S_{lex} = \{(m, n) \in \mathbb{Z}^2 : [m > 0] \text{ or } [m = 0, n \geq 0]\}, \tag{90}$$

where the corresponding total order is known as the *lexicographic order*. If α is irrational, the subset

$$S_\alpha = \{(m, n) \in \mathbb{Z}^2 : m\alpha + n \geq 0\} \tag{91}$$

is easily verified to be a half space as well. The following result is due to Helson and Lowdenslager [22] (cf. also [23, Section 1.2.3]). It is a nontrivial generalization of the well-known Szegö's Theorem (also known in the literature as Kolmogorov's Formula [20, Section 5.8]).

Theorem 15 ([22]) *Let $X = \{X_t\}_{t\in\mathbb{Z}^2}$ be a wide-sense stationary process and let g denote the density function associated with the absolutely continuous component in the Lebesgue decomposition of its spectral measure. Then for any half plane S*

$$E\left|X_0 - \hat{X}_0((-S)\setminus\{0\})\right|^2 = \exp\left\{\frac{1}{4\pi^2}\int_{[0,2\pi)^2} \log g(\lambda) d\lambda\right\}. \tag{92}$$

Note that $(-S)\setminus\{0\} = \{s \in \mathbb{Z}^2 : s < 0\}$ where, in the right side, we use the total order relationship defined by S. Under this convention $\hat{X}_0((-S)\setminus\{0\}) = \hat{X}_0(\{s \in \mathbb{Z}^2 : s < 0\})$ is the best linear predictor of X_0 based on its infinite "past". In the sequel, we shall write $\{s < 0\}$ as short-hand notation for $\{s \in \mathbb{Z}^2 : s < 0\}$, where the total order relationship should be clear from the context.

The main result of this section is the following.

Theorem 16 *Let Q be any stationary Gaussian field on $\mathbb{Z}^2$. Let $\rho(\cdot)$ be the squared-error loss function. Then*

$$U(\rho, Q) = \exp\left\{\frac{1}{4\pi^2}\int_{[0,2\pi)^2} \log f_Q(\lambda)d\lambda\right\}, \tag{93}$$

where f_Q is the density function associated with the absolutely continuous component in the Lebesgue decomposition of the spectral measure of Q.

For notational convenience in what follows, we let σ_Q^2 denote the right side of (93). To discuss the implication of Theorem 16 and for future reference, we make an explicit note of the following elementary fact, which is easily established using the properties of Hilbert spaces (cf., e.g., [20, Problem 2.18]).

Fact 1 *Let $\{B_n\}_{n\geq 1}$, $B_n \subseteq \mathbb{Z}^2$, satisfy $B_n \nearrow B$ for some $B \subseteq \mathbb{Z}^2$. Then*

$$E\left|X_0 - \hat{X}_0(B_n)\right|^2 \searrow E\left|X_0 - \hat{X}_0(B)\right|^2. \tag{94}$$

Note, in particular, that Theorem 15 combined with Fact 1 imply that if X is distributed according to (the wide-sense stationary) Q, if S is any half plane, and if we let

$$\varepsilon_n \stackrel{\text{def}}{=} E\left|X_0 - \hat{X}_0\left(\{s<0\} \cap (V_n - \lceil n/2\rceil \cdot \mathbf{1})\right)\right|^2 - \sigma_Q^2, \tag{95}$$

where $\{s < 0\}$ is with respect to the total order defined by S, then

$$\varepsilon_n \searrow 0. \tag{96}$$

One notable consequence of the combination of Theorem 16 with equations (95)-(96) is that for large rectangles of a stationary Gaussian field: *The scandictability is (essentially) attained by any scandictor which scans the data according to the total order defined by any half plane S (and, of-course, employs the corresponding optimal linear predictor).*

Another consequence of Theorem 16 and equations (95)-(96) is that of all w.s.s. fields with a given spectrum *the Gaussian field is hardest to scandict.* To see this note that the performance (i.e., the normalized cumulative MSE) of the scandictor which (asymptotically) achieves optimum performance in the Gaussian case depends only on the second-order statistics of the field. In the non-Gaussian case, however, it may not be the optimal scheme.

The main idea behind the proof of Theorem 16 is the following: Fix a half-space S. The fact that $\{N_i = X_i - \hat{X}_i(\{s<0\}+i)\}_{i\in\mathbb{Z}^2}$ is a two-dimensional white noise process (due to the orthogonality principle) and is Gaussian (because of the Gaussianity of $X = \{X_t\}_{i\in\mathbb{Z}^2}$ and the linearity of $\hat{X}_i(\cdot)$)

implies that it is a Gaussian i.i.d. process and, in particular, has components with a maximum-entropy distribution w.r.t. the squared loss function. Since $\{X_i\}_{i\in\mathbb{Z}^2}$ is generated auto-regressively by $\{N_i\}$ (i.e., $X_i = N_i + \hat{X}_i(\{s<0\}+i)$), then the conditions of Corollary 10 are satisfied, e.g., by $\{X(V_n)\}_{n\geq 1}$ (recall that $V_n \in \mathcal{V}$ is the $n\times n$ rectangle whose lower left corner is at the origin). By predicting $\{X_i\}$ on finite, growing rectangles, we are approximating better and better the optimal predictor, based on the infinite past (associated with S). This idea is made precise in the formal proof which follows.

Proof of Theorem 16: Let $X=\{X_t\}_{t\in\mathbb{Z}^2}$ be distributed according to Q. Let $\{m_n\}_{n\geq 1}$ be an arbitrary increasing sequence of positive integers satisfying

$$m_n/m_{n+1} \to 0. \tag{97}$$

By item (b) of Theorem 1 it will suffice to show that

$$\lim_{n\to\infty} U(\rho, X(V_{m_n})) = \sigma_Q^2. \tag{98}$$

Furthermore, since $X(V_{m_n})$ is a Gaussian field on V_{m_n}, for any scan Ψ the corresponding optimal (under the MSE criterion) predictor F is a linear combination of the values of the field at the previously observed sites. A fortiori, such a predictor consists of smoothly differentiable functions so that $U(\rho, X(V_{m_n})) = U_D(\rho, X(V_{m_n}))$ for each n and, consequently, we will be done upon showing that

$$\lim_{n\to\infty} U_D(\rho, X(V_{m_n})) = \sigma_Q^2. \tag{99}$$

To this end, we fix a half-space, say, for concreteness, S_{lex} of equation (90) so that, in the remainder of the proof, inequalities between members of $\mathbb{Z}^2$ should be understood in the sense of the lexicographic order. Note that this total order also induces a deterministic (data-independent) scan on any $V\in\mathcal{V}$ according to which site $i\in V$ is reached before site $j\in V$ if and only if $i<j$. We construct now the sequence $\mathbf{W}=\{W_t\}_{t\geq 1}$ inductively through the following steps:

- At the first step, $W_1,\ldots,W_{(m_1)^2}$ are defined to be the prediction errors when scanning V_{m_1} lexicographically and employing the optimal linear predictor. I.e., if $i\in V_{m_1}$ is the t-th site reached when scanning V_{m_1} lexicographically, then

$$W_t = X_i - \hat{X}_i(\{s<i\}\cap V_{m_1}). \tag{100}$$

- At the $n+1$-th step, the components $W_{(m_n)^2+1},\ldots,W_{(m_{n+1})^2}$ are defined to be the prediction errors when scanning $V_{m_{n+1}}\setminus V_{m_n}$ lexicographically and employing the optimal linear predictor which bases its prediction for site $i\in V_{m_{n+1}}\setminus V_{m_n}$ on the values observed at the previously scanned sites of $V_{m_{n+1}}\setminus V_{m_n}$ as well as on $X(V_{m_n})$ (which is known from the n-th step). I.e., if $i\in V_{m_{n+1}}\setminus V_{m_n}$ is the t-th site reached in the lexicographic scan of $V_{m_{n+1}}\setminus V_{m_n}$ then

$$W_{(m_n)^2+t} = X_i - \hat{X}_i\left(\left\{\{s<i\}\cap V_{m_{n+1}}\right\}\cup V_{m_n}\right). \tag{101}$$

Clearly, the components of $\mathbf{W}$ are zero-mean (the optimal linear predictor is always unbiased), Gaussian (each is a finite linear mixture of components of a Gaussian field), and independent (by the orthogonality principle). Furthermore, by the construction of $\mathbf{W}$, Theorem 15, Fact 1, and the stationarity of X we have: On the one hand

$$\mathrm{Var}(W_t) \geq \sigma_Q^2 \quad \forall t. \tag{102}$$

On the other hand, for each $n \geq 1$ there are clearly more than $(m_{n+1} - 2m_n)^2$ sites $i \in V_{m_{n+1}}$ for which $\left\{\{s < i\} \cap V_{m_{n+1}}\right\} \supseteq \{\{s < i\} \cap \{V_{m_n} - \lceil m_n/2 \rceil \cdot \mathbf{1} + i\}\}$. By stationarity, this means that the mean squared error associated with each such i, namely the variance of W_t for t-s corresponding to such i-s, is upper bounded by $E\left(X_0 - \hat{X}_0\left(\{s < 0\} \cap (V_{m_n} - \lceil m_n/2 \rceil \cdot \mathbf{1})\right)\right)^2$. Consequently, for each such i we have, for the corresponding W_t

$$\mathrm{Var}(W_t) \leq \sigma_Q^2 + \varepsilon_{m_n}, \tag{103}$$

where $\{\varepsilon_n\}$ was defined in equation (95). At the remaining sites of $V_{m_{n+1}}$, the corresponding W_t clearly satisfies

$$\mathrm{Var}(W_t) \leq R(0). \tag{104}$$

Hence, we have both

$$\sigma_Q^2 \leq \frac{1}{|V_{m_n}|} \sum_{i=t}^{|V_{m_n}|} \rho(W_t) \leq \left(\frac{m_{n+1} - 2m_n}{m_{n+1}}\right)^2 \left(\sigma_Q^2 + \varepsilon_{m_n}\right) + \left[1 - \left(\frac{m_{n+1} - 2m_n}{m_{n+1}}\right)^2\right] \cdot R(0) \tag{105}$$

and (by the Gaussianity of each W_t)

$$\begin{aligned}\frac{1}{2}\log\left[2\pi e \sigma_Q^2\right] \leq \frac{1}{|V_{m_n}|} \sum_{i=t}^{|V_{m_n}|} H(W_t) \quad &\leq \quad \left(\frac{m_{n+1} - 2m_n}{m_{n+1}}\right)^2 \frac{1}{2}\log\left[2\pi e\left(\sigma_Q^2 + \varepsilon_{m_n}\right)\right] \\ &+ \left[1 - \left(\frac{m_{n+1} - 2m_n}{m_{n+1}}\right)^2\right] \cdot \frac{1}{2}\log\left[2\pi e R(0)\right].\end{aligned} \tag{106}$$

Equations (105) and (106), combined with (97), imply that $\mathbf{W} = \{W_t\}_{t \geq 1}$ satisfies equations (41) and (42) for $B_n = V_{m_n}$ and $W_* \sim N(0, \sigma_Q^2)$. Furthermore, letting $\Psi^{(n)}$ stand for the scan corresponding to that by which $W^{(n)} = (W_1, \ldots, W_{|V_{m_n}|})$ was constructed and $F^{(n)}$ correspond to the associated optimal linear predictor, clearly $(\Psi^{(n)}, F^{(n)}) \in \mathcal{S}_D(V_{m_n})$ and $X(V_{m_n}) = T^{-1}_{(\Psi^{(n)}, F^{(n)})}(W^{(n)})$. Thus, the setting of Theorem 9 holds and Corollary 10 (recall, in particular, the Gaussian example following it) implies that (99) holds, thereby completing the proof. □

We point out that the proof idea extends to the case of any stationary field $Q \in \mathcal{M}_s(\Omega)$ that can be auto-regressively represented as:

$$X_t = f(X(\{s < 0\} + t)) + W_t, \tag{107}$$

where $W = \{W_t\}_{t \in \mathbb{Z}^2}$ is an i.i.d. field (the "innovations" process with continuous-valued components), $f : \mathbb{R}^{\{s<0\}} \to \mathbb{R}$ is a measurable map, and $\{s < 0\}$ is w.r.t. any half-plane. Slightly more formally, $Q \in \mathcal{M}_s(\Omega)$ must be such that for any Borel $I \subseteq \mathbb{R}$

$$Q(X_0 \in I|\mathcal{F}_{\{s<0\}}) = \int_I f_{W_*}(a - f(X(\{s<0\})))da \quad \text{a.s.}, \tag{108}$$

where $f_{W_*}(\cdot)$ is the PDF of the W_t-s. For such $Q \in \mathcal{M}_s(\Omega)$, the above proof idea easily extends to show that

$$\gamma_\rho^{-1}(H(W_*)) \leq U(\rho, Q) \leq E\rho(W_*), \tag{109}$$

with equality when $f_{W_*}(\cdot)$ is a maximum-entropy distribution w.r.t. ρ.

5 The Case $A = P < \infty$

We dedicate this section to the case where the components of the data-array, as well as the predictions of the scandictors, take values in the same finite alphabet A. We shall further assume throughout this section that the subtraction operation is well defined and that, as in Section 3, we have a difference loss function. This will allow us to follow a line of reasoning analogous to that from the case of real-valued observations and predictions treated in previous sections. In particular, the volume-preservation arguments of Section 3 are replaced here by (somewhat simpler) "cardinality-preservation" arguments, to obtain lower bounds on the attainable scandiction performance.

More concretely, assume throughout this section that $(A, +)$ is a group. I.e., the operation $+$ is associative and there exists $0 \in A$ such that

$$\forall a \in A \;\; \exists(-a) \in A: \;\; a + (-a) = (-a) + a = 0. \tag{110}$$

Following the usual convention, for $a, b \in A$ we write $a - b$ for $a + (-b)$. We assume that the loss function $l(\cdot, \cdot)$ is of the form

$$l(F, a) = \rho(a - F) \quad \forall F, a \in A, \tag{111}$$

for a given $\rho : A \to [0, \infty)$ satisfying $\rho(a) = 0$ if and only if $a = 0$. Let now, analogously as in Section 3, for any $B \in \mathcal{V}$ and any scandictor $(\Psi, F) \in \mathcal{S}(B)$, the transformation $T_{(\Psi,F)} : A^B \to A^{|B|}$ be defined by

$$T_{(\Psi,F)}(x(B)) = \left(x(\Psi_1) - F_1, x(\Psi_2) - F_2, \ldots, x(\Psi_{|B|}) - F_{|B|}\right), \tag{112}$$

where Ψ_t and F_t on the right side of (22) are, respectively, the t-th site and t-th prediction associated with the scandictor (Ψ, F) when operating on $x(B)$ and the subtractions on the right side of (112) are in the group sense of (110). In words, $T_{(\Psi,F)}$ maps $x(B) \in A^B$ into the sequence of prediction errors incurred when the scandictor (Ψ, F) operates on $x(B)$. For any scandictor (Ψ, F), given the sequence of prediction errors, the data-array $x(B)$ is uniquely (auto-regressively) determined (recall analogous discussion following the proof of Theorem 5). Hence:

Fact 2 *For any scandictor $(\Psi, F) \in \mathcal{S}(B)$, the transformation $T_{(\Psi,F)} : A^B \to A^{|B|}$ defined in (112) is one-to-one.*

An immediate consequence of Fact (2), which is key to the results of this section, is the following discrete analogue of equation (37):

$$\left|\left\{e_1^{|B|} \in A^{|B|} : \sum_{i=1}^{|B|} \rho(e_i) \leq nd\right\}\right| = \left|\left\{x(B) \in A^B : L_{(\Psi,F)}(x(B)) \leq d\right\}\right| \quad \forall B \in \mathcal{V}, (\Psi, F) \in \mathcal{S}(B). \tag{113}$$

We now define quantities analogous to those in Section 3 as follows[4]: The *log-moment generating function* associated with the loss function ρ is defined by

$$\lambda_\rho(s) = -\log\left[\sum_{z\in A} e^{-s\rho(z)}\right], \quad s > 0, \tag{114}$$

and its *one-sided Fenchel-Legendre transform* is, as before, defined by

$$\gamma_\rho(d) = \inf_{s>0}[sd - \lambda_\rho(s)], \quad d > 0. \tag{115}$$

Analogously as for the continuous case, the function $\gamma_\rho(d)$ can be interpreted as the entropy associated with the PMF

$$q_{\rho,s}(z) = e^{-s\rho(z)+\lambda_\rho(s)} \quad z \in A, \tag{116}$$

where $s \geq 0$ is tuned so that $E_{\rho,s}\rho(Z) = d$ (for d sufficiently small so that such an s exists), $E_{\rho,s}$ (as before) being the expectation operation w.r.t. $q_{\rho,s}$. It is also easy to verify[5] that $q_{\rho,s}(\cdot)$ is a maximum entropy distribution for the discrete case as well, i.e., that

$$E_q\rho(Z) \leq E_{\rho,s}\rho(Z) \Rightarrow H(q) \leq H(q_{\rho,s}), \tag{117}$$

with equality on the right side if and only if $q = q_{\rho,s}$. Hence, the first item of Proposition 4 holds verbatim for this case as well. Furthermore, (117) implies that the right side of (115) is the explicit expression for the more qualitative form

$$\gamma_\rho(d) = \max_{p:E_p\rho(Z)\leq d} H(p). \tag{118}$$

Since $\gamma_\rho(\cdot)$ is continuous (as is seen from its definition (115)) it follows, e.g. by combining the relation (118) with a typical-sequences analysis [24], that

$$\lim_{n\to\infty} \frac{1}{n} \log\left|\left\{e_1^n \in A^n : \sum_{i=1}^{n} \rho(e_i) \leq nd\right\}\right| = \gamma_\rho(d), \tag{119}$$

which is the discrete-alphabet analogue of equation (21). Equation (119), combined with (113), implies that for large $B \in \mathcal{V}$ and any scandictor $(\Psi, F) \in \mathcal{S}(B)$, $\left|\left\{x(B) \in A^B : L_{(\Psi,F)}(x(B)) \leq d\right\}\right| \approx e^{|B|\gamma_\rho(d)}$. Thus, if $G_B \subseteq A^B$ is a set of size which is exponentially larger than $e^{|B|\gamma_\rho(d)}$, then $L_{(\Psi,F)}(x(B)) > d$ for all $x(B) \in G_B \setminus J_B$, where the size of J_B is an exponentially negligible fraction of the size of G_B. More formally, equations (119) and (113) lead to the following.

[4]We maintain the notation from the previous sections to emphasize the analogy.

[5]The proof follows that from the continuous case (cf. proof of the first item of Proposition 4) verbatim up to the replacement of integrals by sums.

Theorem 17 *For any $a > \gamma_\rho(d)$ and $\varepsilon < a - \gamma_\rho(d)$, there exists n_0 such that: For all $B \in \mathcal{V}$ with $|B| \geq n_0$, for all $G_B \subseteq A^B$ with $|G_B| \geq e^{|B|a}$, and any scandictor $(\Psi, F) \in \mathcal{S}(B)$*

$$L_{(\Psi,F)}(x(B)) > d \quad \forall x(B) \in G_B \setminus J_B, \tag{120}$$

where

$$\frac{|J_B|}{|G_B|} \leq e^{-|B|\varepsilon}. \tag{121}$$

Theorem 17 is an "individual-sequence" type of result. For the probabilistic setting, we have the following analogue of Theorem 9.

Theorem 18 *Let $\{B_n\}_{n\geq 1}$, $B_n \in \mathcal{V}$, be an arbitrary sequence satisfying $|B_n| \to \infty$. Let $\mathbf{W} = \{W_i\}_{i\geq 1}$ be a sequence of independent A-valued random variables converging in distribution to some W_*. Let further $\{(\Psi^{(n)}, F^{(n)})\}_{n\geq 1}$ be an arbitrary sequence of scandictors, where $(\Psi^{(n)}, F^{(n)}) \in \mathcal{S}(B_n)$. Finally let, for each n, $x(B_n)$ be the random field on B_n which is auto-regressively generated by the scandictor $(\Psi^{(n)}, F^{(n)})$ with the innovation process $W^{(n)} = (W_1, W_2, \ldots, W_{|B_n|})$, i.e.*

$$x(B_n) = T^{-1}_{(\Psi^{(n)},F^{(n)})}(W^{(n)}). \tag{122}$$

Then

$$\gamma_\rho^{-1}(H(W_*)) \leq \liminf_{n\to\infty} U(\rho, x(B_n)) \leq \limsup_{n\to\infty} U(\rho, x(B_n)) \leq E\rho(W_*). \tag{123}$$

The proof of Theorem 18 is analogous to (though simpler than) that of Theorem 9.
Proof sketch: The upper bound in (123) is established by considering the expected performance of $(\Psi^{(n)}, F^{(n)})$ on $x(B_n)$ which, by construction of $x(B_n)$, is precisely $\frac{1}{|B_n|}\sum_{i=1}^{|B_n|} E\rho(W_i)$, which converges to $E\rho(W_*)$. For the lower bound we observe that, by the AEP[6] and Fact 2, for any $\varepsilon > 0$ and sufficiently large n, if $J_{B_n} \subset A^{B_n}$ with $|J_{B_n}| \leq e^{|B_n|(H(W_*)-\varepsilon)}$ then $x(B_n) \in J_{B_n}$ with probability $\leq \varepsilon$. In particular, for large n and any scandictor $(\Psi, F) \in \mathcal{S}(B_n)$, we can take $J_{B_n} = \left\{x(B_n) \in A^{B_n} : L_{(\Psi,F)}(x(B_n)) \leq d\right\}$. Since, as discussed above, $J_{B_n} \approx e^{|B_n|\gamma_\rho(d)}$, if $H(W_*) > \gamma_\rho(d)$ then we will have $L_{(\Psi,F)}(x(B_n)) \leq d$ with probability $\leq \varepsilon$. Using this line of reasoning, one can show that $\liminf_{n\to\infty} U(\rho, x(B_n)) \geq d$ whenever $d < \gamma_\rho^{-1}(H(W_*))$, which implies the lower bound in (123). $\square$
For simplicity, in the hypotheses of Theorem 18 we have required the convergence in distribution of $\{W_i\}_{i\geq 1}$ to W_*, which implies in the present finite-alphabet setting that equations (41) and (42) hold[7].

Since, as discussed above, the first item of Proposition 4 holds for the current setting, Theorem 18 implies, similarly as Corollary 10 from the continuous case, the following.

[6] In particular, Theorem 22 of Appendix B can be harnessed for this setting to show that for any $\varepsilon > 0$, large n, and set of size $\leq e^{n(H(W_*)-\varepsilon)}$, the probability of $(W_1, W_2, \ldots, W_n)$ belonging to that set is $\leq \varepsilon$.

[7] This is in contrast to the continuous setting of Section 3, where convergence in distribution does not imply that equations (41) and (42) hold.

Corollary 19 *Let the setting of Theorem 18 hold and suppose further that W_* has a maximum-entropy distribution (i.e., of the form (116)) w.r.t. ρ. Then*

$$\lim_{n\to\infty} U(\rho, x(B_n)) = E\rho(W_*). \tag{124}$$

The following subsections are dedicated to the application of Theorem 18 and Corollary 19 to a few concrete cases of special interest.

A Auto-Regressive Markov Random Fields

Consider $\mathbb{Z}^2$ and let S be any half-plane (so that inequalities among elements of $\mathbb{Z}^2$ in what follows are w.r.t. the total order defined by S). Let $X = \{X_t\}_{t\in\mathbb{Z}^2}$, $X_t \in A$, be a stationary random field, governed by $Q \in \mathcal{M}_s(\Omega)$, which can be auto-regressively represented as:

$$X_t = f(X(\{s < 0\} + t)) + W_t, \tag{125}$$

where $W = \{W_t\}_{t\in\mathbb{Z}^2}$, $W_t \in A$, is an i.i.d. field (the "innovations" process), $f : A^{\{s<0\}} \to A$ is a given mapping, and addition in the right side of (125) is in the group sense of this section. In other words, the conditional distribution of X_0 based on its past is given by

$$Q(X_0 = a|X(\{s < 0\})) = p_W\left(a - f(X(\{s < 0\}))\right) \quad \text{a.s.}, \tag{126}$$

where p_W is the PMF of W_1. For this case, one can use Theorem 18, very similarly to (yet even more simply than) the way that Theorem 9 was used to establish Theorem 16, to show that

$$\gamma_\rho^{-1}(H(W_1)) \le U(\rho, Q) \le E\rho(W_1), \tag{127}$$

with equality when W_1 has a maximum-entropy distribution w.r.t. ρ. Furthermore, the upper bound on $U(\rho, Q)$ in (127) is achieved via the deterministic scan induced by the half-plane S. In particular, when the distribution of W_1 is maximum-entropy, such a scan achieves the optimum scandiction performance.

For a concrete example, let ρ_H stand for Hamming loss:

$$\rho_H(a) = \begin{cases} 0 & \text{if } a = 0 \\ 1 & \text{otherwise,} \end{cases} \tag{128}$$

so that the associated maximum-entropy distributions are easily seen to be of the form

$$p_H^\varepsilon(a) = \begin{cases} 1 - (|A| - 1)\varepsilon & \text{if } a = 0 \\ \varepsilon & \text{otherwise,} \end{cases} \tag{129}$$

for $0 < \varepsilon < \frac{1}{|A|}$. For an MRF characterized by (125) or, equivalently, by (126), where W_1 is distributed according to p_H^ε, we thus have

$$U(\rho_H, X) = (|A| - 1)\varepsilon. \tag{130}$$

Specializing this observation even further, consider now the binary case where $A = 2$ and $+$ denotes modulo-2 addition. For this case, if $\Pr(W_1 = 1) = \varepsilon < 1/2$, then W_1 has a maximum-entropy distribution. Furthermore, here it is easy to see that equation (126) holds for some f if and only if

$$Q(X_0 = 1 | X(\{s < 0\})) \in \{\varepsilon, 1 - \varepsilon\} \quad \text{a.s.} \tag{131}$$

We thus have

Corollary 20 *Let $Q \in \mathcal{M}_s(\Omega)$ be a binary field satisfying (131) (w.r.t. any half-plane S). Then*

$$U(\rho_H, Q) = \varepsilon, \tag{132}$$

where the (asymptotically) optimal performance is achieved by scanning the data according to the order corresponding to S.

Examples are:

Symmetric First-Order Markov Binary Source in One Dimension: This case was mentioned in the discussion following Corollary 10. If X is a first-order Markov process (on $\mathbb{Z}$) with a symmetric transition matrix $\begin{pmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{pmatrix}$, then Corollary 20 implies that $U(\rho_H, X) = \varepsilon$. In particular, for this process the chronological scan is optimal. A more interesting example is:

Eight Nearest Neighbors Binary MRFs: Take, for concreteness, $S = S_{lex}$ and suppose that X is a binary MRF on $\mathbb{Z}^2$ governed by $Q \in \mathcal{M}_s(\Omega)$ such that

$$Q(X_0 = 1 | X(\{s < 0\})) = Q(X_0 = 1 | X_{(-1,0)}, X_{(0,-1)}) \quad \text{a.s.} \tag{133}$$

Suppose further that

$$Q(X_0 = 1 | X_{(-1,0)} = a, X_{(0,-1)} = b) = p_{ab}, \tag{134}$$

where $p_{ab} \in \{\varepsilon, 1 - \varepsilon\}$. The presentation (134) has an equivalent eight- (in effect six-) nearest-neighbors presentation (cf. [23, Section 2.2.5]). Corollary 20 implies that $U(\rho_H, Q) = \varepsilon$ for this case, which can be achieved via the lexicographic scan.

6 Multiple Loss Criteria

Suppose that rather than a single loss function we are presented with a list of loss functions w.r.t. which scandiction performance is to be evaluated. For concreteness, assume the setting of Section 3 where the source alphabet and the predictions are real-valued and the loss functions are of the form $l_i(F, y) = \rho_i(y - F)$, each $\rho_i(\cdot)$ satisfying the regularity assumptions of Section 3. In this context, given a list of k loss functions $(\rho_1, \ldots, \rho_k)$, it is natural to try and characterize the achievable region of the vector of corresponding losses $(d_1, \ldots, d_k)$. We formalize this as follows.

Definition 3 *For a sequence of random fields $\{X(B_n)\}_{n\geq 1}$, $B_n \in \mathcal{V}$, the tuple $(d_1, \ldots, d_k)$ is* achievable *w.r.t. the loss functions $(\rho_1, \ldots, \rho_k)$ if for every $\varepsilon > 0$*

$$\lim_{n\to\infty} \sup_{(\Psi,F)\in\mathcal{S}(B_n)} \Pr\left(L^{\rho_i}_{(\Psi,F)}(X(B_n)) \leq d_i + \varepsilon,\ 1 \leq i \leq k\right) = 1. \tag{135}$$

Note that Definition 3 is a "with-high-probability" definition, as opposed to the scandictability of Definition 2, which was given in terms of expectation. We recall, however, that the probabilistic results of the previous sections (cf., in particular, proof of Theorem 9 and Corollary 11) are fundamentally high-probability results, from which expected performance was easily derived. Hence it should be clear that the setting of the previous sections is attained from the setting of this section as a special case by taking $k = 1$.

Analogues of lower and upper bounds on scandiction performance in previous sections for the case of multiple loss criteria would be in terms of inner and outer bounds on the achievable region. Such bounds can be obtained by generalizing the techniques of Section 3, as we outline in what follows. For notational simplicity we address the case of $k = 2$. The idea carries over to any k in a straightforward way. We define the log-moment generating function associated with the loss functions ρ_1, ρ_2 by

$$\lambda_{\rho_1,\rho_2}(s_1, s_2) = -\log\left[\int e^{-s_1\rho_1(z)-s_2\rho_2(z)}dz\right], \quad s_1, s_2 > 0, \tag{136}$$

and its one-sided Fenchel-Legendre transform by

$$\gamma_{\rho_1,\rho_2}(d_1, d_2) = \inf_{s_1,s_2>0}[s_1 d_1 + s_2 d_2 - \lambda_{\rho_1,\rho_2}(s_1, s_2)], \quad d_1, d_2 > 0. \tag{137}$$

Letting

$$q_{s_1,s_2}(z) = e^{-s_1\rho_1(z)-s_2\rho_2(z)+\lambda_{\rho_1,\rho_2}(s_1,s_2)}, \tag{138}$$

it is easy to verify that:

$$\frac{\partial}{\partial s_i}\lambda_{\rho_1,\rho_2}(s_1, s_2) = E\rho_i(Z) \quad i = 1, 2, \tag{139}$$

$$\frac{\partial^2}{\partial s_i^2}\lambda_{\rho_1,\rho_2}(s_1, s_2) = -\mathrm{Var}\left(\rho_i(Z)\right) \quad i = 1, 2, \tag{140}$$

and

$$\frac{\partial^2}{\partial s_1 \partial s_2}\lambda_{\rho_1,\rho_2}(s_1, s_2) = -\mathrm{Cov}\left(\rho_1(Z), \rho_2(Z)\right), \tag{141}$$

where Z on the right sides of equations (139)-(141) is distributed according to the PDF $q_{s_1,s_2}(z)$ of equation (138). Hence, similarly as in the case of a single loss function, $\gamma_{\rho_1,\rho_2}(d_1, d_2)$ can be thought of as the differential entropy of a random variable Z with the PDF of equation (138), where s_1 and s_2 are tuned so that $E\rho_i(Z) = d_i$ (whenever a unique such pair of $s_1, s_2 > 0$ exists). Note that equations (140) and (141) imply that $\left(\frac{\partial^2}{\partial s_1^2}\lambda_{\rho_1,\rho_2}(s_1, s_2)\right)\left(\frac{\partial^2}{\partial s_2^2}\lambda_{\rho_1,\rho_2}(s_1, s_2)\right) > \left[\frac{\partial^2}{\partial s_1 \partial s_2}\lambda_{\rho_1,\rho_2}(s_1, s_2)\right]^2$ for all $s_1, s_2 > 0$ so that it is indeed the minimum in the right side of (137) that such a distribution achieves. It is also easy to extend the proof in Appendix A for this case and to show that $q_{s_1,s_2}(\cdot)$ is a maximum-entropy distribution in the sense that for all $s_1, s_2 > 0$

$$E_q \rho_i(Z) \leq E_{s_1,s_2} \rho_i(Z),\ i = 1,2 \Rightarrow H(q) \leq H(q_{s_1,s_2}), \tag{142}$$

where E_{s_1,s_2} denotes expectation w.r.t. q_{s_1,s_2} and equality on the right side of (142) holds if and only if $q(\cdot) = q_{s_1,s_2}(\cdot)$. Using this, Proposition 4 can be extended to this case and, in particular, it can be shown that

$$\lim_{n\to\infty} \frac{1}{n} \log \mathrm{Vol}\left(\left\{e_1^n : \sum_{i=1}^{n} \rho_j(e_i) \leq n d_j, \quad j = 1,2\right\}\right) = \gamma_{\rho_1,\rho_2}(d_1, d_2). \tag{143}$$

We can now state the analogue of Theorem 9 for this setting.

Theorem 21 *Let $\{B_n\}_{n\geq 1}$, $B_n \in \mathcal{V}$, be an arbitrary sequence satisfying $|B_n| \to \infty$. Let $\mathbf{W} = \{W_i\}_{i\geq 1}$ be a sequence of independent continuous random variables, where the density function of W_i is $f_{W_i}(\cdot)$, $\mathrm{Var}(\log f_{W_i}(W_i)) \leq C < \infty$, and for which there exists a continuous random variable W_* such that*

$$\frac{1}{|B_n|} \sum_{i=1}^{|B_n|} H(W_i) \to H(W_*) \tag{144}$$

and

$$\frac{1}{|B_n|} \sum_{i=1}^{|B_n|} E\rho_j(W_i) \to E\rho_j(W_*) \quad j = 1,2. \tag{145}$$

Let further $\{(\Psi^{(n)}, F^{(n)})\}_{n\geq 1}$ be an arbitrary sequence such that $(\Psi^{(n)}, F^{(n)}) \in \mathcal{S}_D(B_n)$. Finally let, for each n, $X(B_n)$ be the random field on B_n which is auto-regressively generated by the scandictor $(\Psi^{(n)}, F^{(n)})$ with the innovation process $W^{(n)} = (W_1, W_2, \ldots, W_{|B_n|})$, i.e.

$$X(B_n) = T^{-1}_{(\Psi^{(n)},F^{(n)})}(W^{(n)}). \tag{146}$$

Then

1. *The pair (d_1, d_2) is achievable for $\{X(B_n)\}_{n\geq 1}$ if $d_j \geq E\rho_j(W_*)$, $j = 1,2$.*

2. *The pair (d_1, d_2) is not achievable for $\{X(B_n)\}_{n\geq 1}$ if $\gamma_{\rho_1,\rho_2}(d_1, d_2) < H(W_*)$.*

We omit the proof of Theorem 21 as it is completely analogous to that of Theorem 9. In particular, the proof of the first item follows by employing the scandictor $(\Psi^{(n)}, F^{(n)})$ from which $X(B_n)$ was generated. The proof of the second item follows by combining the volume-preservation property, the volume estimate (143), and an AEP-type argument, as in the proof of Theorem 9. Note that analogously to Corollary 10, Theorem 21 and equation (142) imply that when W_* has a maximum-entropy distribution, i.e., is distributed according to q_{s_1,s_2}, then the point $(d_1, d_2) = (E\rho_1(W_*), E\rho_2(W_*))$ is a boundary point of the achievable region in the sense that it is the only achievable point in the rectangle $[0, d_1] \times [0, d_2]$.

Outer bounds on the achievable region can also be attained in terms of the structure of the loss functions and the relationships between them. Thus, for example, if $\rho_1(z) = |z|$ and $\rho_2(z) = z^2$ then, by the Cauchy-Schwartz inequality, it is clear that any pair of the form $(\sqrt{d}, d)$ is achievable when d is achievable for ρ_2 and is not achievable when $\sqrt{d}$ is not achievable for ρ_1. Unfortunately, it is hard to evaluate $\gamma_{\rho_1,\rho_2}(d_1, d_2)$ explicitly even for such simple loss functions.

7 Future Directions

Future research directions related to this work include:

1. Assessing the tightness of the upper and lower bounds in Theorem 9 (equation (44)) for the case where the distribution of W_* is not maximum-entropy. Suppose, for example, that the field $X(B_n)$ is auto-regressively generated by some $(\Psi^{(n)}, F^{(n)}) \in \mathcal{S}_D(B_n)$, where the driving noise is zero-mean Gaussian, yet performance is evaluated relative to the absolute loss function $\rho(\cdot) = |\cdot|$. Or, conversely, that the driving noise is Laplacian and performance is evaluated under squared-error loss. Is it still true in these cases that the optimal scandictor for $X(B_n)$ is $(\Psi^{(n)}, F^{(n)})$? An affirmative answer would imply that the "blame" for the gap between the upper and lower bounds in equation (44) lies in the lower bound and that, in fact, equation (45) holds in cases other than when the distribution of W_* is maximum-entropy.

2. As remarked in Section 4, the stationary Gaussian field is hardest to scandict, under squared-error loss, among all the w.s.s. fields of a given spectrum. It would be interesting to find out whether this is strictly so. The key to this may lie in the first item.

3. Universal scandiction. It is not hard to extend the idea underlying universal predictors and construct universal schemes for the scandiction setting. The scandictors resulting from such an approach, however, are far too complex to have any practical value. Thus, it is of interest to find universal scandictors of moderate complexity.

4. Scandiction under the large deviations performance criterion. Is there no loss of optimality in restricting attention to deterministic scandictors for this case? Is it still true that an auto-regressively generated field is best scandicted the way it was generated?

5. Noisy scandiction. Suppose that a scandictor is to operate on a noise-corrupted image (e.g., a Gaussian image corrupted by additive Gaussian white noise), yet its performance is evaluated relative to the clean image (cf., e.g., [25, 26] for the time-series origin of this problem and for its motivation). Do the main results of this work carry over to the noisy setting? In particular, does the main result of Section 4 carry over to the case of a Gaussian image corrupted by additive Gaussian white noise ?

6. Extension of the analysis to non-difference loss functions and to more general probabilistic models.

Some of these directions are under current research.

Acknowledgement

Interesting discussions with Shie Mannor are gratefully acknowledged.

Appendix

A Proof of Proposition 4

Proof of item 1): We start by verifying that $q_{\rho,s}(\cdot)$ $(s>0)$ is a maximum-entropy distribution w.r.t. the loss function ρ. To see this note that if $q(\cdot)$ is any PDF satisfying

$$E_q\rho(Z) \le E_{\rho,s}\rho(Z) \tag{A.1}$$

then

$$\begin{aligned} 0 \le D(q\|q_{\rho,s}) &= -H(q) + \int q(z)\log\frac{1}{q_{\rho,s}(z)}dz \\ &= -H(q) + sE_q\rho(Z) - \lambda_\rho(s) \\ &\le -H(q) + sE_{\rho,s}\rho(Z) - \lambda_\rho(s) = -H(q) + H(q_{\rho,s}). \end{aligned} \tag{A.2}$$

We have thus shown that

$$E_q\rho(Z) \le E_{\rho,s}\rho(Z) \Rightarrow H(q) \le H(q_{\rho,s}), \tag{A.3}$$

where the right side of (A.3) holds with equality if and only if $q(\cdot) = q_{\rho,s}(\cdot)$. To see why this implies item 1) let $E_q\rho(Z) = d$ and recall that $\gamma_\rho(d)$ is the differential entropy of $q_{\rho,s}$, where $E_{\rho,s}\rho(Z) = d$. $\square$

Proof of item 2): Fix an $a>0$ and let $\{X_i\}$ be an i.i.d. sequence $X_i \sim U[-a,a]$. On the one hand clearly

$$\Pr\left\{\sum_{i=1}^n \rho(X_i) \le nd\right\} = \frac{\mathrm{Vol}\left(\{\sum_{i=1}^n \rho(X_i) \le nd\} \cap [-a,a]^n\right)}{(2a)^n} \tag{A.4}$$

so that

$$\frac{1}{n}\log\Pr\left\{\sum_{i=1}^n \rho(X_i) \le nd\right\} = \frac{1}{n}\log\mathrm{Vol}\left(\left\{\sum_{i=1}^n \rho(X_i) \le nd\right\} \cap [-a,a]^n\right) - \log(2a). \tag{A.5}$$

On the other hand, by the non-asymptotic upper bound in Cramér's Theorem (cf. [17, Theorem 2.2.3] and, in particular, remark (c) therein), we have

$$\Pr\left\{\sum_{i=1}^n \rho(X_i) \le nd\right\} \le 2e^{-n\inf_{x\le d}\Lambda_a^*(x)}, \tag{A.6}$$

where

$$\Lambda_a(\lambda) = \log E e^{\lambda\rho(X_1)} = \log\left[\frac{1}{2a}\int_{-a}^{a} e^{\lambda\rho(z)}dz\right] = -\log(2a) + \log\left[\int_{-a}^{a} e^{\lambda\rho(z)}dz\right] \tag{A.7}$$

and

$$\begin{aligned}
\Lambda_a^*(x) &= \sup_{\lambda\in\mathbb{R}}[\lambda x - \Lambda_a(\lambda)] \\
&= \log(2a) + \sup_{\lambda\in\mathbb{R}}\left[\lambda x - \log\left[\int_{-a}^{a} e^{\lambda\rho(z)}dz\right]\right] \\
&\geq \log(2a) + \sup_{\lambda\in\mathbb{R}}\left[\lambda x - \log\left[\int_{-\infty}^{\infty} e^{\lambda\rho(z)}dz\right]\right] \\
&= \log(2a) - \inf_{\lambda\in\mathbb{R}}\left[-\lambda x - \left(-\log\left[\int_{-\infty}^{\infty} e^{\lambda\rho(z)}dz\right]\right)\right] \\
&= \log(2a) - \inf_{s\in\mathbb{R}}\left[sx - \left(-\log\left[\int_{-\infty}^{\infty} e^{-s\rho(z)}dz\right]\right)\right] \\
&\geq \log(2a) - \inf_{s>0}\left[sx - \left(-\log\left[\int_{-\infty}^{\infty} e^{-s\rho(z)}dz\right]\right)\right] \\
&= \log(2a) - \gamma_\rho(x).
\end{aligned} \tag{A.8}$$

Combining inequality (A.6) with (A.8) we obtain

$$\begin{aligned}
\frac{1}{n}\log\Pr\left\{\sum_{i=1}^{n}\rho(X_i) \leq nd\right\} &\leq \frac{\log 2}{n} - \inf_{x\leq d}\Lambda_a^*(x) \\
&\leq \frac{\log 2}{n} - \inf_{x\leq d}[\log(2a) - \gamma_\rho(x)] \\
&= \frac{\log 2}{n} - \log(2a) + \sup_{x\leq d}\gamma_\rho(x) \\
&= \frac{\log 2}{n} - \log(2a) + \gamma_\rho(d),
\end{aligned} \tag{A.9}$$

where the last equality follows by the fact that $\gamma_\rho(x)$ is monotonically increasing in x. Combining equality (A.5) with (A.9) gives

$$\frac{1}{n}\log\mathrm{Vol}\left(\left\{\sum_{i=1}^{n}\rho(X_i) \leq nd\right\} \cap [-a,a]^n\right) \leq \frac{\log 2}{n} + \gamma_\rho(d). \tag{A.10}$$

Finally, taking the limit of the left side of (A.10) as $a \to \infty$ gives the desired result. □

Proof of item 3): Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sequence drawn according to the probability density function $q_{\rho,s}(\cdot)$ (recall (18) for notation), where s is tuned so that $E_{\rho,s}\rho(X_1) = d$. It is then easy to verify that the differential entropy of X_1 is $H(X_1) = \gamma_\rho(d)$. Furthermore, letting $A_n(d) = \{x_1^n : \sum_{i=1}^n \rho(x_i) \leq nd\}$, the weak law of large numbers implies

$$\lim_{n\to\infty}\Pr\{X^n \in A_n(d+\varepsilon)\} = 1 \quad \forall\varepsilon > 0. \tag{A.11}$$

Evidently, $A_n(d+\varepsilon)$ carries most of the probability mass and, therefore, must have volume which is exponentially no less than $e^{H(X_1)}$ (cf., e.g., [18, Theorem 9.2.3]). More precisely:

$$\liminf_{n\to\infty}\frac{1}{n}\log\mathrm{Vol}\left(A_n(d+\varepsilon)\right) \geq H(X_1) = \gamma_\rho(d) \quad \forall\varepsilon > 0. \tag{A.12}$$

Combining inequality (A.12) with the continuity of $\gamma_\rho(d)$ and item 2) of the proposition completes the proof. □

B Proof of equation (52)

By the hypotheses of Theorem 9 we have $\mathrm{Var}\left(\log f_{W_i}(W_i)\right) \leq C < \infty$, $\frac{1}{|B_n|}\sum_{i=1}^{|B_n|} H(W_i) \to H(W_*)$ and $|B_n| \to \infty$. It would therefore be more than enough to prove the following weak law of large numbers.

Theorem 22 *Let $X_1, X_2, \ldots$ be uncorrelated random variables with $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} EX_i = \mu$ and $Var(X_i) \leq C < \infty$. If $S_n = \sum_{i=1}^{n} X_i$ then $S_n/n \to \mu$ in L^2.*

Proof:

$$\begin{aligned} E\left(S_n/n - \mu\right)^2 &= E\left(\frac{S_n}{n} - \frac{1}{n}\sum_{i=1}^{n} EX_i + \frac{1}{n}\sum_{i=1}^{n} EX_i - \mu\right)^2 \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left(X_i\right) + \left[\frac{1}{n}\sum_{i=1}^{n} EX_i - \mu\right]^2 \\ &\leq \frac{C}{n} + \left[\frac{1}{n}\sum_{i=1}^{n} EX_i - \mu\right]^2 \to 0.\Box \end{aligned} \tag{B.1}$$

References

[1] L. Rabiner and R. Schafer. *Digital Processing of Speech Signals.* Prentice-Hall, 1978.

[2] M. Weinberger, G. Seroussi, and G. Sapiro. The LOCO-I lossless image compression algorithm: Principles and standardization into JPEG-LS. *IEEE Trans. Image Proc.*, 1998. Submitted. Available as Hewlett-Packard Laboratories Technical Report No. HPL-98-193, 1998.

[3] Hua, Harry, and Li. *Video Data Compression for Multimedia Computing.* Kluwer academic publishers, January 1997.

[4] Seul, O'Gorman, and Sammon. *Practical Algorithms for Image Analysis.* Cambridge university press, 1999.

[5] T. Sikora. MPEG digital video-coding standards. *IEEE Sig. Proc. Magazine*, pages 82–100, September 1997.

[6] J. Rissanen. Universal coding, information, prediction, and estimation. *IEEE Trans. Inform. Theory*, IT-30(4):629–636, July 1984.

[7] N. Merhav and M. Feder. A strong version of the redundancy-capacity theorem of universal coding. *IEEE Trans. Inform. Theory*, 41(3):714–722, May 1995.

[8] J. Rissanen. Fisher information and stochastic complexity. *IEEE Trans. Inform. Theory*, 42(1):40–47, January 1996.

[9] N. Merhav and M. Feder. Universal prediction. *IEEE Trans. Inform. Theory*, IT-44(6):2124–2147, October 1998.

[10] M. Gardner. Mathematical games. *Sci. Amer.*, pages 124–133, 1976.

[11] B. B. Mandelbrot. *Fractals: Form, Chance, and Dimension.* Freeman, San Francisco, 1977.

[12] A. Lempel and J. Ziv. Compression of two-dimensional data. *IEEE Trans. Inform. Theory*, IT-32(1):2–8, January 1986.

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

[14] H. O. Georgii. *Gibbs Measures and Phase Transitions.* Walter de Gruyter, Berlin - New York, 1988.

[15] N. Merhav and M. Feder. Universal schemes for sequential decision from individual data sequences. *IEEE Trans. Inform. Theory*, 39(4):1280–1292, July 1993.

[16] Z. Ye and T. Berger. *Information Measures for Discrete Random Fields.* Science Press, Beijing, New York, 1998.

[17] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications.* Springer-Verlag, New York, 2nd edition, 1998.

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

[19] G. Schwartz. Estimating the dimension of a model. *Ann. Stat.*, 6(2):461–464, 1978.

[20] P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods.* Springer, New York, 2nd edition, 1991.

[21] W. Rudin. *Fourier Analysis on Group.* Interscience, 1962.

[22] H. Helson and D. Lowdenslager. Prediction theory and fourier series in several variables. *Acta Mathematica*, (99):165–202, 1958.

[23] X. Guyon. *Random Fields on a Network.* Springer-Verlag, New York, 1995.

[24] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems.* Academic Press, New York, 1981.

[25] T. Weissman and N. Merhav. Universal prediction of individual binary sequences in the presence of noise. *To appear in IEEE Trans. Inform. Theory*, September 2001.

[26] T. Weissman, N. Merhav, and A. Baruch. Twofold universal pediction schemes for achieving the finite-state predictability of a noisy individual binary sequence. *To appear in IEEE Trans. Inform. Theory*, July 2001.