Scanning and Prediction in Multidimensional Data Arrays

Neri Merhav, *Fellow, IEEE*, and Tsachy Weissman, *Member, IEEE*

***Abstract*—The problem of sequentially scanning and predicting data arranged in a multidimensional array is considered. We introduce the notion of a *scandictor*, which is any scheme for the sequential scanning and prediction of such multidimensional 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. The limit is shown to exist for any stationary field, and essentially be independent of the ratios between the box dimensions. Fundamental limitations on scandiction performance in both the probabilistic and the deterministic settings are characterized for the family of difference loss functions. We find that any stochastic process or random field that can be generated autoregressively with a maximum-entropy innovation process is optimally "scandicted" the way it was generated. 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 and is shown to be attained by the raster scan. For a family of binary Markov random fields (MRFs), the scandictability under the Hamming distortion measure is fully characterized.**

***Index Terms*—Autoregressive representations, Gaussian fields, Kolmogorov's formula, Markov random fields (MRFs), prediction, random fields, scandiction, scanning.**

I. 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 prediction error. Examples for predictive coding include linear prediction coding (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 that the encoder that acts on the prediction error is fixed, the degrees of freedom left to be optimized are the predictor itself and the scanning strategy, i.e., the choice of the order at which the data are 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 multidimensional array.

In typical prediction problems the data are 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, [2]–[5]) there is no natural direction 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 arranged in a multidimensional 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 tth 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 $\sum_{t=1}^{n_1 \cdot n_2} l(F_t, x_t)$.

Arising naturally in the multidimensional setting, the question of optimally scanning the data for prediction turns out to be an intricate one already for the one-dimensional case. To see this, consider the simple symmetric first-order Markov process defined autoregressively by

$$X_{t+1} = X_t + W_{t+1} \tag{1}$$

where W_t, $t \geq 1$, are independent and identically distributed (i.i.d.), taking values in $\{0, 1, \ldots, M-1\}$ $(M \geq 3)$, with distribution

$$\Pr(W_t = i) = \begin{cases} 1-p, & \text{if } i = 0 \\ p/(M-1), & \text{otherwise} \end{cases}$$

$0 \leq p \leq 1$, and addition in (1) is modulo-M. Assume further, for concreteness, that X_0 is uniformly distributed over $\{0, 1, \ldots, M-1\}$, so that the process is stationary. Suppose now that, for some large n, we are interested in "scandicting" the data $X_1, \ldots, X_n$ in a way that will minimize the expected number of prediction errors. At first glance, the autoregressive representation of the process may seem to suggest that the trivial scan (left to right) is optimal. This indeed turns out to be

Manuscript received August 5, 2001; revised June 13, 2002. The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Laussane, Switzerland, July 2002.

N. Merhav is with the Department of Electrical Engineering, Technion– Israel Institute of Technology, Haifa 32000, Israel (e-mail: merhav@ee.technion.ac.il).

T. Weissman was with the Department of Electrical Engineering, Technion–Israel Institute of Technology, Haifa 32000, Israel. He is now with the Statistics Department, Stanford University, Stanford, CA 94395 USA (e-mail: tsachy@stat.stanford.edu).

Communicated by M. Weinberger, Associate Editor for Source Coding.

Digital Object Identifier 10.1109/TIT.2002.806134

the case, as our results will show, if $p < (M-1)/M$. However, when $p > (M-1)/M$ (i.e., when staying in the previous location is less probable than a transition into each of the other states), it can be shown by direct calculation that scanning first the data indexed by the odd points (say, from left to right) and then "filling in" the even points, attains better performance than the trivial scan. For a concrete numerical example, it is easily verified that for simple random walk $(p = 1)$ on the ternary alphabet $(M = 3)$, the expected error rate of the trivial scandictor is $1/2$, while that of the "odds-then-evens" scandictor is $3/8$. We shall elaborate on this example in Section V.

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 subadditivity 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. In particular, one can take the infinite limit in one dimension first, and only then the limit in the other dimension.

After introducing the notions of a scandictor and scandictability in a general setting, we shall focus in Section III on the case where the data, as well as the predictions, are real valued, and loss is measured with respect to (w.r.t.) a difference loss function.[1] Two approaches for assessing fundamental limitations on scandiction performance will be developed. The first, in Section III-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 elaborated on later, this observation leads to several general lower bounds on scandiction performance in a probabilistic as well as an "individual-data-array" setting. The second approach, in Section III-B, is based on minimum description length (MDL)-type lower bounds [6]–[8]. More specifically, we extend an idea, which was applied in [9, Subsec. 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 Section III-A.

In Section IV, we pay special attention to the stationary Gaussian field on $\mathbb{Z}^2$. The main probabilistic result of Section III 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 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.

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

In Section V, 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 III 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 III, 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 autoregressively represented, the scandictability is fully characterized.

The bottom line of this work is in attaining upper and lower bounds on the achievable scandiction performance for the case of a difference loss function. In particular, we characterize a family of stochastic processes for which the bounds coincide. This family includes all processes (or multidimensional fields) which can be autoregressively represented with an innovation process which has a maximum-entropy distribution w.r.t. the relevant loss function. Any stationary Gaussian field, for example, belongs to this family, under the squared-error loss. We find that an optimal scandictor for such processes is one corresponding to the autoregressive way in which they can be represented.

The essence of our approach for obtaining a lower bound on scandiction performance is 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. This implies that for any such scandictor, the volume of the set of all data arrays for which the scandiction loss is beneath a certain value is the same as the volume of the ρ-"ball" of a radius which equals this value. Therefore, the least expected scandiction error cannot be less than the radius of a ρ-sphere whose volume is equal to the volume of the set of typical sequences of the given source. In other words, since objects cannot "shrink" under the mapping from source sequences onto scandiction error sequences, the best scenario that one can hope for is the one where the typical set of source sequences, which possesses most of the probability, is mapped onto a ρ-sphere in the domain of the error sequences. In particular, this happens to be the case with autoregressively generated processes having a maximum-entropy innovation process, and, therefore, this lower bound is indeed tight for this class of processes. Thus, for example, if the n components of the innovation process are i.i.d. with entropy H then (by the converse to the asymptotic equipartition property (AEP)) the latter probability is small when the radius is taken such that the volume of the ρ-"ball" is (exponentially) less than e^{nH} (cf. Fig. 1 for a schematic illustration of this point).

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 online 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 consid-

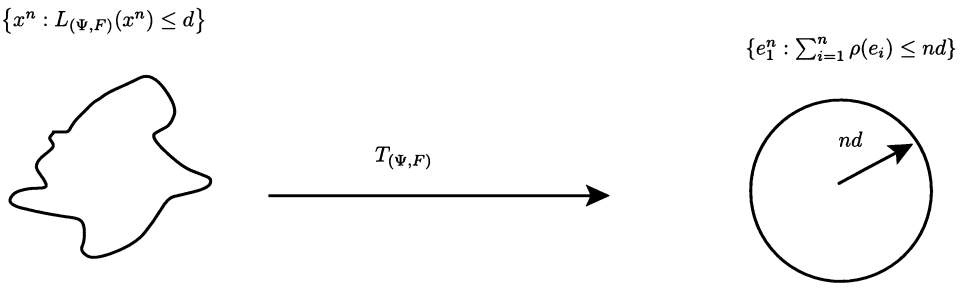

Fig. 1. The volume-preserving mapping $T_{(\Psi,\,F)}$, taking the source sequence into the error sequence associated with the scandictor $(\Psi,\,F)$.

ered 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 multidimensional 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 II, however, in the context of (nonuniversal) compression of probabilistic multidimensional 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, that 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. Finally, we remark that while most of the results of this work are asymptotic in nature, they can often lead to conclusions of nonasymptotic value. For example, the asymptotically optimal normalized expected scandiction (Hamming) loss for the process discussed above (see (1)) will be shown to equal p when $p < (M-1)/M$. It is easy to see, however, that for a block of length n, trivial scandiction gives expected loss of $p+\frac{1}{n}\,[\frac{M-1}{M}-p]$. Thus, given any $\varepsilon>0$, it is clear how large the block length must be to attain optimum scandiction to within ε.

The following summarizes a few of the central themes and conclusions of this work, as previously discussed and as will be elaborated on and established later.

1) Volume preservation considerations as a basis for lower bounds.
2) Suboptimality of natural scandiction even for simple processes.
3) Optimal scandiction performance for all stationary Gaussian fields is attained by the lexicographic (raster) scan. In particular, in one dimension, optimal scandiction is attained by the trivial scan for all stationary Gaussian processes.
4) If a process or field is autoregressively generated with innovations having a maximum-entropy distribution, then it is optimally scandicted the way it was generated.

The remainder of the paper is organized as follows. In Section II, we present the notation, formulate the general setting, and formally introduce the notion of a "scandictor" and the concept of "scandictability." Sections III–V are as elaborated on above. Section VI contains some concluding remarks along with some directions for related future work. For simplicity of the presentation we treat the case of a two-dimensional data array throughout the paper. All the results carry over to higher dimensions in a straightforward way.

II. The Model, Notation, and Definitions

We shall assume the alphabet, denoted generically by A, to be either the real line, or a finite set. We let $\Omega = A^{\mathbb{Z}^2}$ denote the set of all possible realizations. Let further $\mathcal{M}(\Omega)$ denote the space of probability measures on Ω (equipped with the cylinder σ-algebra), and denote by $\mathcal{M}_s(\Omega)$ the subspace consisting of all (spatially) stationary measures, i.e., measures that are invariant under all shifts θ_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 (Borel) probability measures on 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*, and a *random field on* B, respectively.

For $n \ge 1$, 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

$$D(V) \stackrel{\text{def}}{=} \sup\{r\colon \exists\, c \text{ s.t. } B(c,\, r) \subseteq V\} \tag{2}$$

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 $m_k,\, n_k \in \mathbb{Z}$, $m_k \le n_k$. 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 $\{j \in \mathbb{Z}^2\colon j - i \in V\}$ and $-V$ stand for $\{j \in \mathbb{Z}^2\colon -j \in V\}$. 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 $\langle i,\, j\rangle$ 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 $\cup A_n = A$ $(\cap A_n = A)$." If $\{A_n\}$ is a sequence of reals then $A_n \nearrow A$ $(A_n \searrow A)$ is synonymous to "$\{A_n\}$ is nondecreasing (nonincreasing) and $\lim A_n = A$."

For a finite set of random variables $\boldsymbol{N}$, jointly distributed according to the probability distribution q_N, we shall let $H(\boldsymbol{N})$ as well as $H(q_N)$ denote the entropy. More precisely, the components of $\boldsymbol{N}$ will be either all discrete valued or all continuous valued, so that in the latter case $H(\boldsymbol{N})$ 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, we let $M_1(\Sigma)$ denote the set of all probability measures on Σ.

For $B \subseteq \mathbb{Z}^2$ and data arrays on B with real-valued components $x(B)$, $y(B) \in \mathbb{R}^B$ we denote

$$\|x(B) - Y(B)\|_\infty \stackrel{\text{def}}{=} \max_{i \in B} |x_i - y_i|$$

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 Ψ_t: $A^{t-1} \to B$, with the property that
$$\left\{\Psi_1, \Psi_2(y_1), \Psi_3(y_1, y_2), \ldots, \Psi_{|B|}\left(y_1^{|B|-1}\right)\right\} = B, \quad \forall\, y_1^{|B|-1} \in A^{|B|-1}. \tag{3}$$
- F, the predictor, is a sequence of measurable mappings $\{F_t\}_{t=1}^{|B|}$, where F_t: $A^{t-1} \to A$.

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 tth prediction $(1 \le t \le |B|)$

$$F_t = F_t(x(\Psi_1), x(\Psi_2), \ldots, x(\Psi_{t-1}))$$

(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: $A \times A \to [0, \infty)$ is a given loss function. Note that property (3) 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

$$L_{(\Psi, F)}(x(B)) = \frac{1}{|B|} \sum_{t=1}^{|B|} l(F_t, x(\Psi_t)) \tag{4}$$

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. Note that a scandictor, according to Definition 1, is not allowed to randomize its prediction or choice of the next site. That is, 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

$$U(l, Q_B) = \inf_{(\Psi, F) \in \mathcal{S}(B)} E_{Q_B} L_{(\Psi, F)}(X(B)) \tag{5}$$

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

$$U(l, Q) = \lim_{n \to \infty} U(l, Q_{V_n}) \tag{6}$$

whenever the limit exists.

Note that the scandictability of $Q \in \mathcal{M}(\Omega)$ is defined as the limit of the scandictability of the finitely indexed fields Q_{V_n}. Thus, henceforth, the term "scandictor for Q" will be short-hand terminology for the more precise phrasing "sequence of scandictors for the respective fields Q_{V_n}." We also remark that while most of our results are asymptotic in nature, they can lead to nonasymptotic conclusions.

Notice 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

$$l_{\log}(F, a) = -\log F(a), \qquad a \in A. \tag{7}$$

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.,

$$U(l_{\log}, Q_B) = \frac{1}{|B|} H(Q_B). \tag{8}$$

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

$$E_{Q_B} L_{(\Psi, F)}(X(B)) = E_{Q_B}[-\log Q_{(\Psi, F)}(X(B))].$$

Using the fact that $D(Q_B \| \tilde{Q}_B) \ge 0$ 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 (8) 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 (8) hold, but the scandictability is attained by *any* scan. In this context, the scandictability notion of Definition 2 can be

[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.

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 (6) 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 (6) exists; and
b) for any sequence $\{B_n\}$ of elements of $\mathcal{R}_\square$ satisfying $D(B_n) \to \infty$, we have

$$U(l, Q) = \lim_{n\to\infty} U(l, Q_{B_n}) = \inf_{\Delta\in\mathcal{R}_\square} U(l, Q_\Delta). \quad (9)$$

Theorem 1, proof of which is given later, 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 nonoverlapping blocks congruent to any Δ satisfying $U(l, Q_\Delta) \le U(l, Q) + \varepsilon$. Finally, we note that by letting $l = l_{\log}$ Theorem 1 and (8) 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 subadditivity of $U(l, Q_V)$. Specifically, we have the following.

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

$$|V\cup V'|U(l, Q_{V\cup V'}) \le |V|U(l, Q_V)+|V'|U(l, Q_{V'}). \quad (10)$$

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 $|\tilde{V}|U(l, Q_{\tilde{V}}) \le |V'|U(l, Q_{V'})$ 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

$$E_Q L_{(\Psi,F)}(X(B)) \le |V|U(l, Q_V) + |V'|U(l, Q_{V'}). \quad (11)$$

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 (5).[3] □

The relevance of subadditivity to establishing the existence of a limit is manifested in the following lemma.

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

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

$$f(a_1, \ldots, b_i + c_i, \ldots, a_m) \le f(a_1, \ldots, b_i, \ldots, a_m) + f(a_1, \ldots, c_i, \ldots, a_m). \quad (12)$$

Then for every

$$\{a^{(n)} = (a_1^{(n)}, \ldots, a_m^{(n)})\}_{n\ge 1}$$

with $(\min_{1\le i\le m} a_i^{(n)}) \to \infty$

$$\lim_{n\to\infty} \frac{f\left(a^{(n)}\right)}{\prod_{i=1}^m a_i^{(n)}} = \inf_{a\in(\mathbb{Z}_+)^m} \frac{f(a)}{\prod_{i=1}^m a_i}. \quad (13)$$

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 $D(B_n) \to \infty$

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

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 $D(B_n) \to \infty$

$$\lim_{n\to\infty} U(l, Q_{B_n}) = \inf_{\Delta\in\mathcal{R}^0_\square} U(l, Q_\Delta). \quad (15)$$

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

$$f(a_1, a_2) \stackrel{\text{def}}{=} a_1 a_2 U\left(l, Q_{V(a_1,a_2)}\right) \quad (16)$$

where

$$V(a_1, a_2) \stackrel{\text{def}}{=} \mathbb{Z}^2 \cap ([0, a_1) \times [0, a_2)).$$

The subadditivity of f is a direct consequence of Lemma 2 and the stationarity of Q. The proof is completed by an appeal to Lemma 3. □

Note that it also follows from the above derivations that the scandictability can be reached by taking the limits "one dimension at a time" (note that this does not follow directly as a special case of Theorem 1 b) because the diameter does not tend to infinity). To see this, let $\{\varepsilon_n\}$ be any sequence of positive reals satisfying $\varepsilon_n \to 0$. Let now U_n be the scandictability when the first dimension is sent to infinity and the other one is fixed at n (note that it necessarily exists by subadditivity in that first dimension). Construct now the increasing sequence $\{m_n\}$ by letting m_n be the smallest integer which is larger than m_{n-1} and which is also sufficiently large so that $|U_{m_n,n} - U_n| \le \varepsilon_n$, $U_{m,n}$ denoting the scandictability of the $m \times n$ rectangle. By Theorem 1, we know that the $\lim_{n\to\infty} U_{m_n,n}$ exists and equals the scandictability. On the other hand, by construction of the sequence $\{m_n\}$, the limit of $\{U_{m_n,n}\}$ must coincide with the quantity obtained by taking the limits "one dimension at a time."

III. The Case $A = \mathbb{R}$

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

$$\lambda(s) = -\log\left[\int e^{-s\rho(z)}\,dz\right], \qquad s > 0 \tag{17}$$

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

$$\gamma(d) = \inf_{s>0}[sd - \lambda(s)], \qquad d > 0. \tag{18}$$

As remarked in [9], the function $\gamma(d)$ can be interpreted as the differential entropy associated with the PDF

$$q_s(z) = e^{-s\rho(z)+\lambda(s)} \tag{19}$$

where s is tuned so that $E_s\rho(Z) = d$, E_s being the expectation operation w.r.t. q_s. For a reason that will be clear from the proof of the first item of Proposition 4 later (cf., in particular, (A1) of Appendix A), we refer to $q_s(\cdot)$ as a *maximum-entropy distribution w.r.t.* ρ. It can be seen that $\gamma(d)$ is strictly monotonically increasing and concave for $d \geq 0$ and, therefore, the inverse function $\gamma^{-1}(\cdot)$ exists and is continuous. Two additional important aspects of $\gamma(d)$, which will be of later use, are encapsulated in the following proposition, whose proof is deferred to the Appendix.

Proposition 4:

1) For any PDF $q(\cdot)$
$$\gamma(E_q\rho(Z)) \geq H(q) \tag{20}$$
with equality if and only if $q(\cdot) = q_s(\cdot)$ for some $s > 0$.

2) For all $n \geq 1$ and all $d \geq 0$
$$\frac{1}{n}\log \mathrm{Vol}\left(\left\{e_1^n : \sum_{i=1}^n \rho(e_i) \leq nd\right\}\right) \leq \frac{\log 2}{n} + \gamma(d). \tag{21}$$

3)
$$\lim_{n\to\infty}\frac{1}{n}\log \mathrm{Vol}\left(\left\{e_1^n : \sum_{i=1}^n \rho(e_i) \leq nd\right\}\right) = \gamma(d). \tag{22}$$

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-hand side of (5). 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 nonpathological 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 $U_D(\rho, X) = U(\rho, X)$. We omit the proof of this fact (which can be more rigorously formulated), as it is cumbersome in detail but straightforward. The key is to note 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 trivially satisfying $U_D(\rho, X(B)) = U(\rho, X(B))$ for all $B \in \mathcal{V}$ and $U_D(\rho, X) = U(\rho, X)$ is the Gaussian field of Section IV (as the optimal predictor is always linear and, *a fortiori*, continuously differentiable).

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

$$T_{(\Psi,F)}(x(B)) = \big(x(\Psi_1) - F_1, x(\Psi_2) - F_2, \ldots, x\left(\Psi_{|B|}\right) - F_{|B|}\big) \tag{23}$$

where Ψ_t and F_t on the right-hand side of (23) are, respectively, the tth site and tth 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 (23) 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

$$T_\Psi(x(B)) = \big(x(\Psi_1), x(\Psi_2), \ldots, x\left(\Psi_{|B|}\right)\big) \tag{24}$$

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

$$T_F\left(y_1^{|B|}\right) = \Big(y_1 - F_1, y_2 - F_2(y_1), \ldots, y_{|B|} - F_{|B|}\left(y_1^{|B|-1}\right)\Big). \tag{25}$$

Clearly

$$T_{(\Psi,F)} = T_F \circ T_\Psi \tag{26}$$

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-hand side of (25)), 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 $T_\phi\colon \mathbb{R}^B \to \mathbb{R}^{|B|}$, defined by
$$T_\phi(x(B)) = (x(\phi(1)),\, x(\phi(2)),\, \ldots,\, x(\phi(|B|))) \quad (27)$$
is volume preserving (as it corresponds to a relabeling of the axes).
2) For each $x(B) \in \mathbb{R}^B$ there exists a unique $\phi \in \Phi_B$ such that
$$T_\Psi(x(B)) = T_\phi(x(B)). \quad (28)$$

Let now $G_B \subseteq \mathbb{R}^B$ be an arbitrary Borel set. For each $\phi \in \Phi_B$ define
$$G_B^\phi = \{x(B) \in G_B\colon T_\Psi(x(B)) = T_\phi(x(B))\}$$
(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.,
$$\bigcup_{\phi\in\Phi_B} G_B^\phi = G_B \quad \text{and} \quad G_B^\phi \cap G_B^{\phi'} = \emptyset \qquad \forall\, \phi \neq \phi'. \quad (29)$$
Consequently, we have
$$\begin{aligned}
\mathrm{Vol}(T_\Psi(G_B)) &= \mathrm{Vol}\left(T_\Psi\left(\bigcup_{\phi\in\Phi_B} G_B^\phi\right)\right) && (30)\\
&= \mathrm{Vol}\left(\bigcup_{\phi\in\Phi_B} T_\Psi(G_B^\phi)\right) && (31)\\
&= \sum_{\phi\in\Phi_B} \mathrm{Vol}(T_\Psi(G_B^\phi)) && (32)\\
&= \sum_{\phi\in\Phi_B} \mathrm{Vol}(T_\phi(G_B^\phi)) && (33)\\
&= \sum_{\phi\in\Phi_B} \mathrm{Vol}(G_B^\phi) && (34)\\
&= \mathrm{Vol}(G_B), && (35)
\end{aligned}$$
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 (30)–(32) are well-defined. Equation (31) 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 (32) follows from the fact that T_Ψ is one-to-one and, hence, the sets in $\{T_\Psi(G_B^\phi)\}_{\phi\in\Phi_B}$ are disjoint. Equation (33) follows from the definition of the sets G_B^ϕ and (34) 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 $(\Psi,\, F)$ 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 $(\Psi,\, F)$ into the original data array $x(B)$. More specifically, $T^{-1}_{(\Psi,\, F)}\colon \mathbb{R}^{|B|} \to \mathbb{R}^B$ 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 autoregressively constructed using W as the innovation process according to
$$x(\Psi_1) = F_1 + W_1, \quad x(\Psi_2(x(\Psi_1))) = F_2(x(\Psi_1)) + W_2$$
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 $N = (N_1,\, N_2,\, \ldots,\, N_{|B|}) \in \mathbb{R}^{|B|}$ be the error sequence associated with the scandictor $(\Psi,\, F)$ when operating on $X(B)$, i.e., $N = T_{(\Psi,\, F)}(X(B))$. Then
$$H(N) = H(X(B)). \quad (36)$$
2) Let $W = (W_1,\, W_2,\, \ldots,\, W_{|B|}) \in \mathbb{R}^{|B|}$ be a discrete- or continuous-valued random vector and let $X(B)$ be a random field on B autoregressively defined by $X(B) = T^{-1}_{(\Psi,\, F)}(W)$. Then
$$H(X(B)) = H(W). \quad (37)$$

To derive another corollary note that Theorem 5 implies, in particular, that for all $B \in \mathcal{V}$, $(\Psi,\, F) \in \mathcal{S}_D(B)$
$$\begin{aligned}
&\mathrm{Vol}\left(\left\{e_1^{|B|} \in A^{|B|}\colon \sum_{i=1}^{|B|} \rho(e_i) \le nd\right\}\right)\\
&\qquad = \mathrm{Vol}\left(\left\{x(B) \in A^B\colon L_{(\Psi,\, F)}(x(B)) \le d\right\}\right). \quad (38)
\end{aligned}$$
Combined with the third item of Proposition 4, this implies that for large $B \in \mathcal{V}$ and any scandictor $(\Psi,\, F) \in \mathcal{S}_D(B)$
$$\mathrm{Vol}(\{x(B) \in A^B\colon L_{(\Psi,\, F)}(x(B)) \le d\}) \approx e^{|B|\gamma(d)}.$$
Thus, if $G_B \subseteq A^B$ is a set of a volume which is exponentially larger than $e^{|B|\gamma(d)}$, 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, (22) and (38) lead to the following.

Corollary 7: For any $a > \gamma(d)$ and $\varepsilon < a - \gamma(d)$, 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 $(\Psi, F) \in \mathcal{S}_D(B)$

$$L_{(\Psi, F)}(x(B)) > d \qquad \forall\, x(B) \in G_B \setminus J_B \tag{39}$$

where

$$\frac{\mathrm{Vol}(J_B)}{\mathrm{Vol}(G_B)} \leq e^{-|B|\varepsilon}. \tag{40}$$

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 (38) and the second item of Proposition 4.

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

$$\frac{1}{|B|} \log \mathrm{Vol}\left(\{x(B) \in \mathbb{R}^B \colon L_{(\Psi, F)}(x(B)) \leq d\}\right) \leq \frac{\log 2}{|B|} + \gamma(d). \tag{41}$$

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 $\boldsymbol{W} = \{W_i\}_{i \geq 1}$ 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 exist values H_* and ρ_* such that

$$\frac{1}{|B_n|} \sum_{i=1}^{|B_n|} H(W_i) \to H_* \tag{42}$$

and

$$\frac{1}{|B_n|} \sum_{i=1}^{|B_n|} E\rho(W_i) \to \rho_*. \tag{43}$$

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 autoregressively generated by the scandictor $(\Psi^{(n)}, F^{(n)})$ with the innovation process $W^{(n)} = (W_1, W_2, \ldots, W_{|B_n|})$, i.e.,

$$X(B_n) = T^{-1}_{(\Psi^{(n)}, F^{(n)})}(W^{(n)}). \tag{44}$$

Then

$$\gamma^{-1}(H_*) \leq \liminf_{n \to \infty} U_D(\rho, X(B_n)) \leq \limsup_{n \to \infty} U_D(\rho, X(B_n)) \leq \rho_*. \tag{45}$$

The upper bound in (45) 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 (45) should hold note that if d is such that $\gamma(d) < H_*$ then, by Corollary 8, there exists $\varepsilon > 0$ such that for all n sufficiently large and any scandictor $(\Psi, F) \in \mathcal{S}_D(B_n)$

$$\mathrm{Vol}(\{x(B_n) \in \mathbb{R}^{B_n} \colon L_{(\Psi, F)}(x(B_n)) \leq d\}) \leq e^{|B_n|(H_* - \varepsilon)}.$$

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 $\leq e^{|B_n|(H_* - \varepsilon)}$. But the fact that (42) 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 $d < \gamma^{-1}(H_*)$, which implies the left inequality in (45). This is the essential idea behind the formal proof that follows. Prior to the proof of Theorem 9, we note the following two corollaries regarding the tightness of the upper and lower bounds in (45), which are direct consequences of Theorem 9 and the first item of Proposition 4.

Corollary 10: Let $X = (X_1, X_2, \ldots)$ be a stochastic process autoregressively generated by

$$X_t = f_t(X_1^{t-1}) + W_t, \qquad t \geq 1 \tag{46}$$

where $\{f_t\}_{t \geq 1}$ is a sequence of continuously differentiable functions $f_t \colon \mathbb{R}^{t-1} \to \mathbb{R}$ and $\{W_t\}$ are i.i.d. with a ρ-maximum-entropy distribution. Then $U(\rho, X) = E\rho(W_1)$.

Corollary 10 implies that the scandictor achieving (asymptotically) optimal performance for a stochastic process representable in the form (46) is that which scans the data from left to right and predicts $f_t(X_1^{t-1})$ for the value at t. Somewhat more generally we have the following.

Corollary 11: Let the setting of Theorem 9 hold and suppose further that there exists a continuous random variable W_* with a max-entropy distribution $f_{W_*}(\cdot) = q_s(\cdot)$ for some $s > 0$ such that $H(W_*) = H_*$ and $E\rho(W_*) = \rho_*$. Then

$$\lim_{n \to \infty} U_D(\rho, X(B_n)) = \rho_*. \tag{47}$$

Note that, in particular, Corollary 11 tells us that for large $B \in \mathcal{V}$, if $X(B)$ is autoregressively 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. When the innovations are not maximum entropy, characterizing optimal scandiction performance is currently an open problem. In general, when there is a gap between the left- and the right-hand side of (45), both the upper bound and the lower bound are to "blame." One demonstration of this is the process mentioned in Section I (see (1)) when $p > (M-1)/M$. For a concrete example, consider scandiction under Hamming loss of the simple random walk defined by $X_{t+1} = X_t + W_{t+1}$, where the process takes values in $\{0, 1, 2\}$, addition is modulo-2, and

$$W_t = \begin{cases} 1, & \text{w.p. } 1/2 \\ 2, & \text{w.p. } 1/2. \end{cases}$$

For this process, the right-hand side of (45) gives $1/2$ (attained by trivial scandiction), while the left-hand side is easily verified to be given by the root of the equation $h(d) + d = 1$ (≈ 0.227092). On the other hand, as one can show via "brute-force" calculations for this case [25], optimal scandiction for this process is attained by the odds-then-evens predictor, which

is easily verified to attain scandiction loss of $3/8$. Evidently, for this process there is a gap between the upper and the lower bound in (45), and neither are tight. We shall return to the example from Section I in Section V, where the finite-alphabet version of Corollary 11 (Corollary 20) will be shown to imply the optimality of the trivial scan for the range of p discussed in Section I.

Two concrete examples for the significance of Corollary 11 are as follows.

Gaussian Innovation and Squared-Error Loss: If there exists a Gaussian W_* with $H(W_*)=H_*$ and $E\rho(W_*)=\sigma^2$, $\rho(z)=z^2$, Corollary 11 gives

$$\lim_{n\to\infty} U_D(\rho,\, X(B_n)) = \sigma^2. \tag{48}$$

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

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 11 gives

$$\lim_{n\to\infty} U_D(\rho,\, X(B_n)) = \beta. \tag{49}$$

Proof of Theorem 9: To establish the upper bound on the limit supremum in (45) 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

$$L_{(\Psi^{(n)},\, F^{(n)})}(X(B_n)) = \frac{1}{|B_n|}\sum_{i=1}^{|B_n|}\rho(W_i). \tag{50}$$

Thus,

$$\begin{aligned}\limsup_{n\to\infty} U_D(\rho,\, X(B_n)) &\le \limsup_{n\to\infty} E L_{(\Psi^{(n)},\, F^{(n)})}(X(B_n)) \\ &= \limsup_{n\to\infty} E\frac{1}{|B_n|}\sum_{i=1}^{|B_n|}\rho(W_i) \\ &= \rho_* \end{aligned} \tag{51}$$

where the last equality follows from (43).

We now progress to establish the lower bound on the limit infimum in (45). To this end, fix an arbitrary $d < \gamma^{-1}(H_*)$ and an arbitrary $\delta > 0$. Let further

$$\varepsilon = \min\left\{\frac{H_* - \gamma(d)}{2},\, \delta/2\right\}$$

and let $A_\varepsilon^{(m)}$ denote the ε-typical set with respect to $f_{\boldsymbol{W}}(\cdot)$ defined as follows:

$$A_\varepsilon^{(m)} = \left\{(w_1,\, w_2,\, \ldots,\, w_m)\in\mathbb{R}^m:\ \left|-\frac{1}{m}\log f_{\boldsymbol{W}}(w_1,\, w_2,\, \ldots,\, w_m) - H_*\right| \le \varepsilon\right\} \tag{52}$$

where $f_{\boldsymbol{W}}(w_1,\, w_2,\, \ldots,\, w_m) = \prod_{i=1}^m f_{W_i}(w_i)$. As easily shown in Appendix B, to follow from the hypotheses that

$$\mathrm{Var}(\log f_{W_i}(W_i)) \le C < \infty$$

that (42) holds and that $|B_n| \to \infty$, for all $\varepsilon > 0$

$$\lim_{n\to\infty} \Pr\left\{W^{(n)} \in A_\varepsilon^{(|B_n|)}\right\} = 1. \tag{53}$$

This is the analog for our setting of the standard ε-typicality result of the i.i.d. case (cf., e.g., [18, Theorem 9.2.2, item 1]). 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

$$\begin{aligned}&\Pr\left\{L_{(\Psi,\, F)}(X(B_n)) \le d\right\} \\ &\quad\le \Pr\left(\left\{L_{(\Psi,\, F)}(X(B_n)) \le d\right\} \cap \left\{W^{(n)} \in A_\varepsilon^{(|B_n|)}\right\}\right) \\ &\qquad + \Pr\left\{W^{(n)} \notin A_\varepsilon^{(|B_n|)}\right\} \\ &\quad\le \int_{\{\{w^{(n)}:\, L_{(\Psi,\, F)}(x(B_n))\le d\}\cap A_\varepsilon^{(|B_n|)}\}} \\ &\qquad \cdot f_{\boldsymbol{W}}(w^{(n)})\, dw^{(n)} + \delta/2 \qquad (54) \\ &\quad\le \int_{\{\{w^{(n)}:\, L_{(\Psi,\, F)}(x(B_n))\le d\}\cap A_\varepsilon^{(|B_n|)}\}} \\ &\qquad \cdot e^{-|B_n|(H_*-\varepsilon)}\, dw^{(n)} + \delta/2 \\ &\quad= \mathrm{Vol}\left(\left\{w^{(n)}:\, L_{(\Psi,\, F)}(x(B_n)) \le d\right\} \cap A_\varepsilon^{(|B_n|)}\right) \\ &\qquad \cdot e^{-|B_n|(H_*-\varepsilon)} + \delta/2 \\ &\quad\le \mathrm{Vol}\left(\left\{w^{(n)}:\, L_{(\Psi,\, F)}(x(B_n)) \le d\right\}\right) \\ &\qquad \cdot e^{-|B_n|(H_*-\varepsilon)} + \delta/2 \\ &\quad= \mathrm{Vol}\left(\left\{x(B_n):\, L_{(\Psi,\, F)}(x(B_n)) \le d\right\}\right) \\ &\qquad \cdot e^{-|B_n|(H_*-\varepsilon)} + \delta/2 \qquad (55) \\ &\quad\le \exp\left\{|B_n|\left(\frac{\log 2}{|B_n|} + \gamma(d) - H_* + \varepsilon\right)\right\} + \delta/2 \qquad (56) \\ &\quad\le \exp\left\{-|B_n|\left(\frac{H_* - \gamma(d)}{2} - \frac{\log 2}{|B_n|}\right)\right\} + \delta/2 \qquad (57) \\ &\quad\le \delta \qquad (58)\end{aligned}$$

where the inequality in (54) follows by taking $n_0(d,\, \delta)$ sufficiently large so that

$$\Pr\left\{W^{(n)} \notin A_\varepsilon^{(|B_n|)}\right\} \le \delta/2, \qquad \forall\, n \ge n_0(d,\, \delta) \tag{59}$$

which is possible by (53). Equality in (55) 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 (56) follows from Corollary 8. Inequality (57) follows since $\varepsilon \le \frac{H_* - \gamma(d)}{2}$. Inequality (58) follows by taking a sufficiently large $n_0(d,\, \delta)$ such that, in addition to satisfying (59), the first term in (57) 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 $(\Psi,\, F) \in \mathcal{S}_D(B_n)$

$$\begin{aligned}&E L_{(\Psi,\, F)}(X(B_n)) \\ &\quad= E\left[L_{(\Psi,\, F)}(X(B_n)) \middle| L_{(\Psi,\, F)}(X(B_n)) \le d\right] \\ &\qquad \cdot \Pr\left\{L_{(\Psi,\, F)}(X(B_n)) \le d\right\} \\ &\qquad + E\left[L_{(\Psi,\, F)}(X(B_n)) \middle| L_{(\Psi,\, F)}(X(B_n)) > d\right] \\ &\qquad \cdot \left(1 - \Pr\left\{L_{(\Psi,\, F)}(X(B_n)) \le d\right\}\right) \\ &\quad\ge 0 + d\left(1 - \Pr\left\{L_{(\Psi,\, F)}(X(B_n)) \le d\right\}\right) \\ &\quad\ge d(1-\delta). \end{aligned} \tag{60}$$

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

$$U_D(\rho,\, X(B_n)) \ge d(1-\delta) \tag{61}$$

which, in turn, implies

$$\liminf_{n\to\infty} U_D(\rho,\, X(B_n)) \ge d(1-\delta). \tag{62}$$

The arbitrariness of $d < \gamma^{-1}(H(W_1))$ and $\delta > 0$ on the right-hand side of (62) completes the proof. □

In the course of the preceding proof (cf., in particular, the inequalities leading to (58)) we have, in fact, established the following result, from which (45) was easily derived.

Corollary 12: Let the setting of Theorem 9 hold. For any $d < \gamma^{-1}(H_*)$

$$\lim_{n\to\infty} \sup_{(\Psi, F)\in\mathcal{S}_D(B_n)} \Pr\left\{L_{(\Psi, F)}(X(B_n)) \le d\right\} = 0 \quad (63)$$

and for any $d > \rho_*$

$$\lim_{n\to\infty} \sup_{(\Psi, F)\in\mathcal{S}_D(B_n)} \Pr\left\{L_{(\Psi, F)}(X(B_n)) \le d\right\} = 1. \quad (64)$$

In fact, the convergence in (63) and (64) 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]–[8]. We shall use an approach which was applied in [9, Subsec. 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 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

$$\overline{P}_e(M, n, \omega) = 1 - \int_\Lambda \omega(d\theta) \int_{y^n\in\mathcal{I}^n} dy^n \cdot f_\theta(y^n)[1 - \omega\{\theta' \colon f_{\theta'}(y^n) \ge f_\theta(y^n)\}]^{M-1}. \quad (65)$$

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

$$\overline{P}_e(M, n, \omega) \le \delta \quad (66)$$

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

$$C_R(n, \delta, \omega) \stackrel{\text{def}}{=} \log M(n, \delta, \omega). \quad (67)$$

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

$$\overline{P}_e(M, n, \omega) \le (M-1) \cdot \int_\Lambda \omega(d\theta) \int_{y^n\in\mathcal{I}^n} dy^n \cdot f_\theta(y^n)\omega\{\theta' : f_{\theta'}(y^n) \ge f_\theta(y^n)\} \quad (68)$$

so clearly $M(n, \delta, \omega)$ is lower-bounded by the largest integer M for which the right-hand side of (68) is less than δ, namely,

$$M(n, \delta, \omega) \ge \frac{\delta}{\int_\Lambda \omega(d\theta) \int_{y^n\in\mathcal{I}^n} dy^n f_\theta(y^n)\omega\{\theta' \colon f_{\theta'}(y^n) \ge f_\theta(y^n)\}}. \quad (69)$$

Though the precise expression for $C_R(n, \delta, \omega)$ is hard to obtain, the lower bound in (69) 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 13 [7, Theorem 3]: 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 $n \ge 1$

$$E_\theta[-\log q(Y^n)] \ge H_\theta(Y^n) + (1-\varepsilon)C_R(n, \delta, \omega) \quad (70)$$

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

$$\omega(\Lambda_s) \le \frac{\delta C_R(n, \delta, \omega) + 2}{\varepsilon C_R(n, \delta, \omega)}. \quad (71)$$

The preceding theorem [7, Theorem 3] is, in fact, 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 13 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 13, 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 the following. For every PDF $q(\cdot)$ that is independent of $x(B)$, we have

$$E_{x(B)}[-\log q(Y(B))] \ge H(W(B)) + (1-\varepsilon)C_R\left(B, \delta, \frac{\mathrm{Vol}(\cdot)}{\mathrm{Vol}(G_B)}\right) \quad (72)$$

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

$$\frac{\mathrm{Vol}(J_B)}{\mathrm{Vol}(G_B)} \le \frac{\delta C_R\left(B,\, \delta,\, \frac{\mathrm{Vol}(\cdot)}{\mathrm{Vol}(G_B)}\right) + 2}{\varepsilon C_R\left(B,\, \delta,\, \frac{\mathrm{Vol}(\cdot)}{\mathrm{Vol}(G_B)}\right)}. \tag{73}$$

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

$$q(y(B)) = \int_0^\infty ds\nu(s) \cdot \left[\prod_{t=1}^{|B|} q_s([y(\Psi_t) - F_t(y(\Psi_1),\, \ldots,\, y(\Psi_{t-1}))])\right] \tag{74}$$

where the Ψ_t's and F_t's on the right-hand 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 $q_s(\cdot)$ is the maximum-entropy distribution defined in (19). Note that for each s, the bracketed expression in the right-hand side of (74) is a *bona fide* PDF and, consequently, so is $q(\cdot)$. Furthermore, according to the main result of [19] (cf. also [9, eq. (30)]), $-\log q(y(B))$ can be approximated as follows:

$$\begin{aligned} -\log q(y(B)) &= |B| \cdot \inf_{s>0}\left[s \cdot L_{(\Psi,\, F)}(y(B)) - \lambda(s)\right] \\ &\quad + \frac{1}{2}\log|B| + R(y(B)) \\ &= |B| \cdot \gamma\left(L_{(\Psi,\, F)}(y(B))\right) \\ &\quad + \frac{1}{2}\log|B| + R(y(B)) \end{aligned} \tag{75}$$

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 (72) implies that

$$\begin{aligned} |B| \cdot E_{x(B)}\left[\gamma\left(L_{(\Psi,\, F)}(Y(B))\right)\right] &\ge H(W(B)) \\ +(1-\varepsilon)C_R\left(B,\, \delta,\, \frac{\mathrm{Vol}(\cdot)}{\mathrm{Vol}(G_B)}\right) &- \frac{1}{2}\log|B| - R \end{aligned} \tag{76}$$

for all $x(B) \in G_B \setminus J_B$. The concavity of $\gamma(\cdot)$ allows to insert the expectation into the argument of $\gamma(\cdot)$ on the left-hand side of (76) which gives

$$\begin{aligned} |B| \cdot \gamma\left(E_{x(B)} L_{(\Psi,\, F)}(Y(B))\right) &\ge H(W(B)) \\ +(1-\varepsilon)C_R\left(B,\, \delta,\, \frac{\mathrm{Vol}(\cdot)}{\mathrm{Vol}(G_B)}\right) &- \frac{1}{2}\log|B| - R \end{aligned} \tag{77}$$

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

$$H(W(B)) = |B|\log\Delta. \tag{78}$$

To get a more explicit handle on the right-hand side of (77) for this case, we now lower-bound $C_R(B,\, \delta,\, \frac{\mathrm{Vol}(\cdot)}{\mathrm{Vol}(G_B)})$ 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 $y(B)$

$$\mathrm{Vol}\left(x'(B)\colon f^{\Delta}_{x'(B)}(y(B)) > f^{\Delta}_{x(B)}(y(B))\right) \le \Delta^{|B|}. \tag{79}$$

Combining this with the lower bound (69) we obtain

$$C_R\left(B,\, \delta,\, \frac{\mathrm{Vol}(\cdot)}{\mathrm{Vol}(G_B)}\right) \ge \log\left[\frac{\delta\mathrm{Vol}(G_B)}{\Delta^{|B|}}\right]. \tag{80}$$

Substituting (78) and (80) into the right-hand side of (77) gives

$$\begin{aligned} E^{\Delta}_{x(B)} L_{(\Psi,\, F)}(Y(B)) \ge \gamma^{-1}\Bigg((1-\varepsilon)\frac{\log\mathrm{Vol}(G_B)}{|B|} + \varepsilon\log\Delta \\ +(1-\varepsilon)\frac{\log\delta}{|B|} - \frac{\log|B|}{2|B|} - \frac{R}{|B|}\Bigg) \end{aligned} \tag{81}$$

for all $x(B) \in G_B \setminus J_B$. By maintaining a regime where $\varepsilon \to 0$, $\delta \to 0$, $|B| \to \infty$, $\Delta \to 0$, $\varepsilon\log\Delta \to 0$, $0 < a \le \frac{1}{|B|}\log\mathrm{Vol}(G_B) \le b < \infty$, $\delta\log[\frac{\delta e^{|B|a}}{\Delta^{|B|}}] \to \infty$, $\varepsilon\log[\frac{\delta e^{|B|a}}{\Delta^{|B|}}] \to \infty$, and $\delta/\varepsilon \to 0$, we have by (72) and (73) and the continuity of $\gamma^{-1}(\cdot)$ established the following.

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

$$E^{\eta}_{x(B)} L_{(\Psi,\, F)}(Y(B)) \ge \gamma^{-1}\left(\frac{\log\mathrm{Vol}(G_B)}{|B|}\right) - \eta, \quad \forall\, x(B) \in G_B \setminus J_B \tag{82}$$

where

$$\frac{\mathrm{Vol}(J_B)}{\mathrm{Vol}(G_B)} \le \eta. \tag{83}$$

As opposed to the previous subsection, where the converse statements were valid for scandictors with a continuously differentiable predictor, Theorem 14 holds for an arbitrary scandictor. Note also that when $\eta > 0$ is small, $Y(B)$ on the left-hand side of (82) 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

$$\left|L_{(\Psi,\, F)}(x(B)) - L_{(\Psi,\, F)}(x'(B))\right| \le K\|x(B) - x'(B)\|_\infty \tag{84}$$

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, $\|x(B) - Y(B)\|_\infty \le \eta/2$ and, hence,

$$E^{\eta}_{x(B)} L_{(\Psi,\, F)}(Y(B)) \le L_{(\Psi,\, F)}(x(B)) + K\eta/2, \quad \forall\, (\Psi,\, F) \in \mathcal{S}^K(B),\ x(B) \in \mathbb{R}^B. \tag{85}$$

We thus have the following corollary to Theorem 14.

Corollary 15: $\forall\, \eta > 0$, $0 < a < b < \infty$, $\exists\, n_0 = n_0(\eta,\, a,\, b)$ such that: $\forall\, B \in \mathcal{V}$ with $|B| \ge n_0$, $\forall\, G_B \in \mathbb{R}^B$ with $a \le \frac{1}{|B|}\log\mathrm{Vol}(G_B) \le b$, and $\forall\, K > 0$, $(\Psi,\, F) \in \mathcal{S}^K(B)$

$$L_{(\Psi,\, F)}(x(B)) \ge \gamma^{-1}\left(\frac{\log\mathrm{Vol}(G_B)}{|B|}\right) - (K+2)\eta/2, \quad \forall\, x(B) \in G_B \setminus J_B \tag{86}$$

where

$$\frac{\mathrm{Vol}(J_B)}{\mathrm{Vol}(G_B)} \le \eta. \tag{87}$$

Note that similarly to Corollary 7, Corollary 15 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 15 can be further specialized as follows (the details, which are similar to those 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

$$\mathrm{Vol}(\{x(B) \in \mathbb{R}^B \colon L_{(\Psi_0, F_0)}(x(B)) \le d\})$$

is exponentially equivalent to $e^{|B|\gamma(d)}$. Hence, for large $|B|$, taking

$$G_B = \{x(B) \in \mathbb{R}^B \colon L_{(\Psi_0, F_0)}(x(B)) \le d\}$$

in Corollary 15 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 (86)) for all but a set of data arrays whose volume is a negligible fraction of the volume of G_B.

To see the connection between Corollary 15 and the lower bound of Theorem 9, note that if $x(B)$ is assumed generated by a probabilistic source of entropy rate H, then by letting G_B above be the typical set (of exponential size $e^{|B|H}$) one gets a lower bound of $\gamma^{-1}(H)$ on the scandiction performance of any scandictor on most typical sequences, from which the same lower bound for expected scandiction performance is easily attained, essentially recovering the lower bound of Theorem 9.

To end this subsection we point out that the derivation of Theorem 14 and Corollary 15 was based on an application of Theorem 13 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 analogs of the above results with upper bounds on the ratio between the ω-measures of the sets J_B and G_B.

IV. 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 $d \ge 1$.

To fix notation, we recall here the basics regarding spectral representations of wide-sense (second-order) stationary processes. There are no fundamental differences between the time-series and the multidimensional 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, Sec. 4.3] to the multidimensional case dates at least as far back as [21], asserting the following representation of the covariance:

$$R(n) = \frac{1}{(2\pi)^d} \int_{[0, 2\pi)^d} e^{i\langle n, \lambda \rangle} G(d\lambda), \qquad n \in \mathbb{Z}^d \tag{88}$$

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

$$S \text{ is closed to addition}, \quad S \cup (-S) = \mathbb{Z}^2, \quad S \cap (-S) = \{0\}. \tag{89}$$

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

$$(i, j) \le (k, l) \Longleftrightarrow (k - i, l - j) \in S. \tag{90}$$

Examples for half-planes include

$$S_{\mathrm{lex}} = \{(m, n) \in \mathbb{Z}^2 \colon [m > 0] \text{ or } [m = 0, n \ge 0]\} \tag{91}$$

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

$$S_\alpha = \{(m, n) \in \mathbb{Z}^2 \colon m\alpha + n \ge 0\} \tag{92}$$

is easily verified to be a half-plane as well.

The following result is due to Helson and Lowdenslager [22] (cf. also [23, Sec. 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, Sec. 5.8]).

Theorem 16 [22]: Let $X = \{X_t\}_{t \in \mathbb{Z}^2}$ be a w.s.s. 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

$$E \left| X_0 - \hat{X}_0((-S) \setminus \{0\}) \right|^2 = \exp\left\{ \frac{1}{4\pi^2} \int_{[0, 2\pi)^2} \log g(\lambda)\, d\lambda \right\}. \tag{93}$$

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

The main result of this section is the following.

Theorem 17: Let Q be any stationary Gaussian field on $\mathbb{Z}^2$. Let $\rho(\cdot)$ be the squared-error loss function. Then

$$U(\rho, Q) = \exp\left\{ \frac{1}{4\pi^2} \int_{[0, 2\pi)^2} \log f_Q(\lambda)\, d\lambda \right\} \tag{94}$$

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-hand side of (94). To discuss the implication of Theorem 17 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

$$E\left|X_0 - \hat{X}_0(B_n)\right|^2 \searrow E\left|X_0 - \hat{X}_0(B)\right|^2. \tag{95}$$

Note, in particular, that Theorem 16 combined with Fact 1 imply that if X is distributed according to (the w.s.s.) Q, if S is any half-plane, and if we let

$$\varepsilon_n \stackrel{\text{def}}{=} E\left|X_0 - \hat{X}_0(\{s<0\} \cap (V_n - \lceil n/2 \rceil \cdot \mathbf{1}))\right|^2 - \sigma_Q^2 \tag{96}$$

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

$$\varepsilon_n \searrow 0. \tag{97}$$

One notable consequence of the combination of Theorem 17 with (96) and (97) 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 17 and (96) and (97) 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 mean-square error (MSE)) of the scandictor which 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 17 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 autoregressively by $\{N_i\}$ (i.e., $X_i = N_i + \hat{X}_i(\{s<0\}+i)$), then the conditions of Corollary 11 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 17: 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

$$m_n/m_{n+1} \to 0. \tag{98}$$

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

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

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

$$\lim_{n\to\infty} U_D(\rho, X(V_{m_n})) = \sigma_Q^2. \tag{100}$$

To this end, we fix a half-space, say, for concreteness, S_{lex} of (91) 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 $\boldsymbol{W} = \{W_t\}_{t\geq 1}$ 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. That is, if $i \in V_{m_1}$ is the tth site reached when scanning V_{m_1} lexicographically, then
$$W_t = X_i - \hat{X}_i(\{s<i\} \cap V_{m_1}). \tag{101}$$
- 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 nth step). That is, if $i \in V_{m_{n+1}} \setminus V_{m_n}$ is the tth site reached in the lexicographic scan of $V_{m_{n+1}} \setminus V_{m_n}$ then
$$W_{(m_n)^2+t} = X_i - \hat{X}_i\left(\left\{\{s<i\} \cap V_{m_{n+1}}\right\} \cup V_{m_n}\right). \tag{102}$$

Clearly, the components of $\boldsymbol{W}$ 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 $\boldsymbol{W}$, Theorem 16, Fact 1, and the stationarity of X we have

$$\mathrm{Var}(W_t) \geq \sigma_Q^2, \qquad \forall t. \tag{103}$$

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

$$\{\{s<i\} \cap V_{m_{n+1}}\} \supseteq \{\{s<i\} \cap \{V_{m_n} - \lceil m_n/2 \rceil \cdot \mathbf{1} + i\}\}.$$

By stationarity, this means that the MSE associated with each such i, namely, the variance of W_t for t's corresponding to such i's, is upper-bounded by

$$E(X_0 - \hat{X}_0(\{s<0\} \cap (V_{m_n} - \lceil m_n/2 \rceil \cdot \mathbf{1})))^2.$$

Consequently, for each such i we have, for the corresponding W_t

$$\mathrm{Var}(W_t) \leq \sigma_Q^2 + \varepsilon_{m_n} \tag{104}$$

where $\{\varepsilon_n\}$ was defined in (96). At the remaining sites of $V_{m_{n+1}}$, the corresponding W_t clearly satisfies

$$\mathrm{Var}(W_t) \le R(0). \tag{105}$$

Hence, we have both

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

and (by the Gaussianity of each W_t)

$$\begin{aligned}&\frac{1}{2}\log[2\pi e\sigma_Q^2]\\ &\le \frac{1}{|V_{m_n}|}\sum_{i=t}^{|V_{m_n}|}H(W_t)\\ &\le \left(\frac{m_{n+1}-2m_n}{m_{n+1}}\right)^2\frac{1}{2}\log\left[2\pi e\left(\sigma_Q^2+\varepsilon_{m_n}\right)\right]\\ &\quad+\left[1-\left(\frac{m_{n+1}-2m_n}{m_{n+1}}\right)^2\right]\cdot\frac{1}{2}\log[2\pi eR(0)].\end{aligned} \tag{107}$$

Equations (106) and (107), combined with (98), imply that $\boldsymbol{W}=\{W_t\}_{t\ge1}$ satisfies (42) and (43) 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 11 (recall, in particular, the Gaussian example following it) implies that (100) 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 autoregressively represented as

$$X_t=f(X(\{s<0\}+t))+W_t \tag{108}$$

where $W=\{W_t\}_{t\in\mathbb{Z}^2}$ is an i.i.d. field (the innovation process with continuous-valued components), $f\colon\mathbb{R}^{\{s<0\}}\to\mathbb{R}$ 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 $I\subseteq\mathbb{R}$

$$Q\left(X_0\in I\,\middle|\,\mathcal{F}_{\{s<0\}}\right)=\int_I f_{W_*}(a-f(X(\{s<0\})))\,da \text{ a.s.} \tag{109}$$

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

$$\gamma^{-1}(H_*)\le U(\rho,Q)\le\rho_* \tag{110}$$

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

V. The Case $A<\infty$

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 III, 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 III 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. That is, the operation $+$ is associative and there exists $0\in A$ such that

$$\forall\, a\in A,\quad \exists\,(-a)\in A\colon a+(-a)=(-a)+a=0. \tag{111}$$

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

$$l(F,a)=\rho(a-F),\qquad \forall\, F,\,a\in A \tag{112}$$

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

$$\begin{aligned}T_{(\Psi,F)}(x(B))=\big(&x(\Psi_1)-F_1,\,x(\Psi_2)-F_2,\,\ldots,\\ &x\left(\Psi_{|B|}\right)-F_{|B|}\big)\end{aligned} \tag{113}$$

where Ψ_t and F_t on the right-hand side of (23) are, respectively, the tth site and tth prediction associated with the scandictor (Ψ,F) when operating on $x(B)$ and the subtractions on the right-hand side of (113) are in the group sense of (111). 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 (autoregressively) determined (recall analogous discussion following the proof of Theorem 5). Hence we have the following fact.

Fact 2: For any scandictor $(\Psi,F)\in\mathcal{S}(B)$, the transformation $T_{(\Psi,F)}\colon A^B\to A^{|B|}$ defined in (113) is one-to-one.

An immediate consequence of Fact 2, which is key to the results of this section, is the following discrete analog of equation (38):

$$\begin{aligned}&\left|\left\{e_1^{|B|}\in A^{|B|}\colon\sum_{i=1}^{|B|}\rho(e_i)\le nd\right\}\right|\\ &\qquad=\left|\left\{x(B)\in A^B\colon L_{(\Psi,F)}(x(B))\le d\right\}\right|,\\ &\qquad\qquad\forall\, B\in\mathcal{V},\ (\Psi,F)\in\mathcal{S}(B).\end{aligned} \tag{114}$$

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

$$\lambda(s)=-\log\left[\sum_{z\in A}e^{-s\rho(z)}\right],\qquad s>0 \tag{115}$$

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

$$\gamma(d)=\inf_{s>0}[sd-\lambda(s)],\qquad d>0. \tag{116}$$

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

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

$$q_s(z) = e^{-s\rho(z)+\lambda(s)}, \qquad z \in A \tag{117}$$

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

$$E_q\rho(Z) \leq E_s\rho(Z) \Rightarrow H(q) \leq H(q_s) \tag{118}$$

with equality on the right-hand side if and only if $q = q_s$. Hence, the first item of Proposition 4 holds *verbatim* for this case as well. Furthermore, (118) implies that the right-hand side of (116) is the explicit expression for the more qualitative form

$$\gamma(d) = \max_{p:\, E_p\rho(Z)\leq d} H(p). \tag{119}$$

Since $\gamma(\cdot)$ is continuous (as is seen from its definition (116)) it follows, e.g., by combining the relation (119) with a typical-sequences analysis [24], that

$$\lim_{n\to\infty}\frac{1}{n}\log\left|\left\{e_1^n \in A^n\colon \sum_{i=1}^n \rho(e_i) =\leq nd\right\}\right| = \gamma(d) \tag{120}$$

which is the discrete-alphabet analog of (22). Equation (120), combined with (114), implies that for large $B \in \mathcal{V}$ and any scandictor $(\Psi, F) \in \mathcal{S}(B)$

$$|\{x(B) \in A^B\colon L_{(\Psi,F)}(x(B)) \leq d\}| \approx e^{|B|\gamma(d)}.$$

Thus, if $G_B \subseteq A^B$ is a set of size which is exponentially larger than $e^{|B|\gamma(d)}$, 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, (120) and (114) lead to the following.

Theorem 18: For any $a > \gamma(d)$ and $\varepsilon < a-\gamma(d)$, 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 $(\Psi, F) \in \mathcal{S}(B)$

$$L_{(\Psi,F)}(x(B)) > d, \qquad \forall\, x(B) \in G_B \setminus J_B \tag{121}$$

where

$$\frac{|J_B|}{|G_B|} \leq e^{-|B|\varepsilon}. \tag{122}$$

Theorem 18 is an "individual-sequence" type of result. For the probabilistic setting, we have the following analog of Theorem 9.

Theorem 19: Let $\{B_n\}_{n\geq 1}$, $B_n \in \mathcal{V}$, be an arbitrary sequence satisfying $|B_n| \to \infty$. Let $\boldsymbol{W} = \{W_i\}_{i\geq 1}$ 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 autoregressively generated by the scandictor $(\Psi^{(n)}, F^{(n)})$ with the innovation process $W^{(n)} = (W_1, W_2, \ldots, W_{|B_n|})$, i.e.,

$$x(B_n) = T^{-1}_{(\Psi^{(n)}, F^{(n)})}\left(W^{(n)}\right). \tag{123}$$

[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.

Then

$$\begin{aligned}\gamma^{-1}(H(W_*)) &\leq \liminf_{n\to\infty} U(\rho,\, x(B_n)) \\ &\leq \limsup_{n\to\infty} U(\rho,\, x(B_n)) \leq E\rho(W_*).\end{aligned} \tag{124}$$

The proof of Theorem 19 is analogous to (though simpler than) that of Theorem 9.

Proof Sketch: The upper bound in (124) 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

$$J_{B_n} = \{x(B_n) \in A^{B_n}\colon L_{(\Psi,F)}(x(B_n)) \leq d\}.$$

Since, as discussed above, $J_{B_n} \approx e^{|B_n|\gamma(d)}$, if $H(W_*) > \gamma(d)$ 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 $d < \gamma^{-1}(H(W_*))$, which implies the lower bound in (124). □

For simplicity, in the hypotheses of Theorem 19 we have required the convergence in distribution of $\{W_i\}_{i\geq 1}$ to W_*, which implies in the present finite-alphabet setting that (42) and (43) hold.[7]

Since, as discussed earlier, the first item of Proposition 4 holds for the current setting, Theorem 19 implies, similarly as Corollary 11 from the continuous case, the following.

Corollary 20: Let the setting of Theorem 19 hold and suppose further that W_* has a maximum-entropy distribution (i.e., of the form (117)) w.r.t. ρ. Then

$$\lim_{n\to\infty} U(\rho,\, x(B_n)) = E\rho(W_*). \tag{125}$$

In what follows we apply Theorem 19 and Corollary 20 to a few concrete cases of special interest.

Let S be any half-plane (so that inequalities among elements of $\mathbb{Z}^2$ appearing henceforth 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 autoregressively represented as

$$X_t = f(X(\{s < 0\} + t)) + W_t \tag{126}$$

where $W = \{W_t\}_{t\in\mathbb{Z}^2}$, $W_t \in A$, is an i.i.d. field (the innovation process), $f\colon A^{\{s<0\}} \to A$ is a given mapping, and addition in the right-hand side of (126) is in the group sense of this section.

[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 III, where convergence in distribution does not imply that (42) and (43) hold.

In other words, the conditional distribution of X_0 based on its past is given by

$$Q(X_0 = a|X(\{s < 0\})) = p_W(a - f(X(\{s < 0\}))) \quad \text{a.s.} \tag{127}$$

where p_W is the PMF of W_1. For this case, one can use Theorem 19, very similarly to (yet even more simply than) the way that Theorem 9 was used to establish Theorem 17, to show that

$$\gamma^{-1}(H(W_1)) \leq U(\rho, Q) \leq E\rho(W_1) \tag{128}$$

with equality when W_1 has a maximum-entropy distribution w.r.t. ρ. Furthermore, the upper bound on $U(\rho, Q)$ in (128) 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

$$\rho_H(a) = \begin{cases} 0, & \text{if } a = 0 \\ 1, & \text{otherwise} \end{cases} \tag{129}$$

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

$$p_W(i) = \begin{cases} 1 - p, & \text{if } i = 0 \\ p/(M-1), & \text{otherwise} \end{cases} \tag{130}$$

for $p < (M-1)/M$. For an MRF characterized by (126) or, equivalently, by (127), where W_1 is distributed according to p_W, we thus have

$$U(\rho_H, X) = p. \tag{131}$$

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 (127) holds for some f if and only if

$$Q(X_0 = 1|X(\{s < 0\})) \in \{\varepsilon, 1 - \varepsilon\} \quad \text{a.s.} \tag{132}$$

We thus have the following.

Corollary 21: Let $Q \in \mathcal{M}_s(\Omega)$ be a binary field satisfying (132) (w.r.t. any half-plane S). Then

$$U(\rho_H, Q) = \varepsilon \tag{133}$$

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

The following are examples for special cases covered by (131).

Symmetric First-Order Markov Source in One Dimension: This case was mentioned in Section I. If X is a first-order Markov process (on $\mathbb{Z}$) with the autoregressive representation (1), (131) implies that when $p < (M-1)/M$, the optimal scandictor (for Hamming loss, i.e., minimum expected number of errors) is the trivial one, namely, that which scans the data from left to right and predicts the previously observed value. Note that the line of argumentation leading to (131) (and hence to the optimality of trivial scandiction for the autoregressive process under discussion) is no longer valid for the case $p > (M-1)/M$, as for this case the distribution (130) is no longer max-entropy with respect to Hamming loss. Indeed, it is beyond the scope of this work but can be shown [25] that it is the "odds-then-evens" scandictor which is optimal for this range of p (trivial scandiction being strictly suboptimal for this case).

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

$$Q(X_0 = 1|X(\{s < 0\})) = Q\left(X_0 = 1 \,\middle|\, X_{(-1,0)}, X_{(0,-1)}\right) \text{ a.s.} \tag{134}$$

Suppose further that

$$Q\left(X_0 = 1 \,\middle|\, X_{(-1,0)} = a, X_{(0,-1)} = b\right) = p_{ab} \tag{135}$$

where $p_{ab} \in \{\varepsilon, 1 - \varepsilon\}$. The presentation (135) has an equivalent eight-nearest-neighbors presentation, cf. [23, Sec. 2.2.5] for details. Corollary 21 implies that $U(\rho_H, Q) = \varepsilon$ for this case, which can be achieved via the lexicographic scan.

Unfortunately, general MRFs (even as simple as four-nearest-neighbor ones) do not adhere to an autoregressive representation of the type in (126), for which the results of this section hold. Even standard fields susch as the Ising and the Potts model do not have such a representation, and the characterization of their scandictability remains an open problem.

VI. Concluding Remarks and Future Directions

The bottom line of this work is the following conclusion. If a stochastic process or field can be autoregressively represented with a max-entropy innovation process, then it is optimally scandicted using the scandictor associated with the said representation. The optimality criterion discussed in this work for the stochastic setting was expected normalized scandiction loss. The volume-preservation argument used, however, can actually be shown to lead to the following much stronger conclusion. The scandictor associated with the autoregressive representation (assuming a max-entropy innovation process) is optimal also in the error-exponent sense (i.e., has the best large deviations performance) for all threshold values. The interested reader is referred to [27] for the details.

Suppose that rather than a single loss function we are presented with a list of loss functions with respect to which scandiction performance is to be evaluated. 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 $(d_1, \ldots, d_k)$. Analogs 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 III. The interested reader is referred to [26, Sec. 6].

In the remainder of this section, we outline future research directions related to this work. The first direction pertains to assessing the tightness of the upper and lower bounds in Theorem 9 (see (45)) for the case where the distribution of W_* is not maximum entropy. Suppose, for example, that the field $X(B_n)$ is autoregressively 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 (45) lies in the lower bound and that, in fact, (47) holds in cases other than when the distribution of W_* is maximum entropy.

Another direction is that of 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.

An additional direction for future research is that of scandiction under the large-deviations performance criterion. Is there no loss of optimality in restricting attention to deterministic (given the observations) scandictors for this case? Is it still true that an autoregressively generated field is best scandicted the way it was generated? A partial answer (in the affirmative) to the latter question was given in the recent work [27].

Finally, we mention the problem of noisy scandiction. Suppose that a scandictor is to operate on a noise-corrupted image (e.g., a Gaussian image corrupted by additive white Gaussian noise), yet its performance is evaluated relative to the clean image (cf., e.g., [28], [29], 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 IV carry over to the case of a Gaussian image corrupted by additive white Gaussian noise?

Some of the above issues are under current research.

Appendix A
Proof of Proposition 4

Proof of Item 1): According to [18, Theorem 11.1.1],

$$E_q\rho(Z) \le E_s\rho(Z) \Rightarrow H(q) \le H(q_s) \tag{A1}$$

where the right-hand side of (A1) holds with equality if and only if $q(\cdot) = q_s(\cdot)$. To see why this implies item 1) let $E_q\rho(Z) = d$ and recall that $\gamma(d)$ is the differential entropy of q_s, where $E_s\rho(Z) = d$. □

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

$$\Pr\left\{\sum_{i=1}^n \rho(X_i) \le nd\right\} = \frac{\mathrm{Vol}\left(\left\{\sum_{i=1}^n \rho(X_i) \le nd\right\} \cap [-a, a]^n\right)}{(2a)^n} \tag{A2}$$

so that

$$\frac{1}{n}\log\Pr\left\{\sum_{i=1}^n \rho(X_i) \le nd\right\} = \frac{1}{n}\log\mathrm{Vol}\left(\left\{\sum_{i=1}^n \rho(X_i) \le nd\right\} \cap [-a, a]^n\right) - \log(2a). \tag{A3}$$

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

$$\Pr\left\{\sum_{i=1}^n \rho(X_i) \le nd\right\} \le 2e^{-n\inf_{x\le d}\Lambda_a^*(x)} \tag{A4}$$

where

$$\begin{aligned}\Lambda_a(\lambda) &= \log Ee^{\lambda\rho(X_1)} = \log\left[\frac{1}{2a}\int_{-a}^{a} e^{\lambda\rho(z)}\,dz\right] \\ &= -\log(2a) + \log\left[\int_{-a}^{a} e^{\lambda\rho(z)}\,dz\right] \end{aligned}\tag{A5}$$

and

$$\begin{aligned}\Lambda_a^*(x) &= \sup_{\lambda\in\mathbb{R}}[\lambda x - \Lambda_a(\lambda)] \\ &= \log(2a) + \sup_{\lambda\in\mathbb{R}}\left[\lambda x - \log\left[\int_{-a}^{a} e^{\lambda\rho(z)}\,dz\right]\right] \\ &\ge \log(2a) + \sup_{\lambda\in\mathbb{R}}\left[\lambda x - \log\left[\int_{-\infty}^{\infty} e^{\lambda\rho(z)}\,dz\right]\right] \\ &= \log(2a) - \inf_{\lambda\in\mathbb{R}}\left[-\lambda x - \left(-\log\left[\int_{-\infty}^{\infty} e^{\lambda\rho(z)}\,dz\right]\right)\right] \\ &= \log(2a) - \inf_{s\in\mathbb{R}}\left[sx - \left(-\log\left[\int_{-\infty}^{\infty} e^{-s\rho(z)}\,dz\right]\right)\right] \\ &\ge \log(2a) - \inf_{s>0}\left[sx - \left(-\log\left[\int_{-\infty}^{\infty} e^{-s\rho(z)}\,dz\right]\right)\right] \\ &= \log(2a) - \gamma(x). \end{aligned}\tag{A6}$$

Combining inequality (A4) with (A6) we obtain

$$\begin{aligned}\frac{1}{n}\log\Pr\left\{\sum_{i=1}^n \rho(X_i)\le nd\right\} &\le \frac{\log 2}{n} - \inf_{x\le d}\Lambda_a^*(x) \\ &\le \frac{\log 2}{n} - \inf_{x\le d}[\log(2a) - \gamma(x)] \\ &= \frac{\log 2}{n} - \log(2a) + \sup_{x\le d}\gamma(x) \\ &= \frac{\log 2}{n} - \log(2a) + \gamma(d) \end{aligned}\tag{A7}$$

where the last equality follows by the fact that $\gamma(x)$ is monotonically increasing in x. Combining equality (A3) with (A7) gives

$$\frac{1}{n}\log\mathrm{Vol}\left(\left\{\sum_{i=1}^n \rho(X_i) \le nd\right\} \cap [-a, a]^n\right) \le \frac{\log 2}{n} + \gamma(d). \tag{A8}$$

Finally, taking the limit of the left-hand side of (A8) 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 PDF $q_s(\cdot)$ (recall (19)), where s is tuned so that $E_s\rho(X_1) = d$. It is then easy to verify that the differential entropy of X_1 is $H(X_1) = \gamma(d)$. Furthermore, letting $A_n(d) = \{x_1^n\colon \sum_{i=1}^n \rho(x_i) \le nd\}$, the weak law of large numbers implies

$$\lim_{n\to\infty}\Pr\{X^n \in A_n(d+\varepsilon)\} = 1, \qquad \forall\varepsilon > 0. \tag{A9}$$

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

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathrm{Vol}(A_n(d+\varepsilon)) \geq H(X_1) = \gamma(d), \qquad \forall\, \varepsilon > 0. \tag{A10}$$

Combining inequality (A10) with the continuity of $\gamma(d)$ and item 2) of the proposition completes the proof. □

Appendix B
Proof of Equation (53)

By the hypotheses of Theorem 9 we have

$$\mathrm{Var}(\log f_{W_i}(W_i)) \leq C < \infty$$
$$\frac{1}{|B_n|} \sum_{i=1}^{|B_n|} H(W_i) \to H_*$$

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 $\mathrm{Var}(X_i) \leq C < \infty$. If $S_n = \sum_{i=1}^{n} X_i$ then $S_n/n \to \mu$ in L^2.

Proof:

$$\begin{aligned} E(S_n/n - \mu)^2 &= E\left(\frac{S_n}{n} - \frac{1}{n}\sum_{i=1}^{n} EX_i + \frac{1}{n}\sum_{i=1}^{n} EX_i - \mu\right)^2 \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) + \left[\frac{1}{n}\sum_{i=1}^{n} EX_i - \mu\right]^2 \\ &\leq \frac{C}{n} + \left[\frac{1}{n}\sum_{i=1}^{n} EX_i - \mu\right]^2 \to 0. \end{aligned} \tag{B1}$$

□

Acknowledgment

Interesting discussions with Shie Mannor are gratefully acknowledged. The final version benefitted greatly from the insightful comments of the anonymous referees.

References

[1] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.
[2] M. Weinberger, G. Seroussi, and G. Sapiro, "The large LOCI-I lossless image compression algorithm: Principles and standardization into JPEG-LS," *IEEE Trans. Image Processing*, vol. 9, pp. 1309–1324, Aug. 2000.
[3] H. Li, S. Sun, and H. Derin, *Video Data Compression for Multimedia Computing*. Norwell, MA: Kluwer, Jan. 1997.
[4] M. Seul, L. O'Gorman, and M. J. Sammon, *Practical Algorithms for Image Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1999.
[5] T. Sikora, "MPEG digital video-coding standards," *IEEE Signal Processing Mag.*, vol. 14, pp. 82–100, Sept. 1997.
[6] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 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*, vol. 41, pp. 714–722, May 1995.
[8] J. Rissanen, "Fisher information and stochastic complexity," *IEEE Trans. Inform. Theory*, vol. 42, pp. 40–47, Jan. 1996.
[9] N. Merhav and M. Feder, "Universal prediction," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2124–2147, Oct. 1998.
[10] M. Gardner, "Mathematical games," *Sci. Amer.*, pp. 124–133, 1976.
[11] B. B. Mandelbrot, *Fractals: Form, Chance, and Dimension*. San Francisco, CA: Freeman, 1977.
[12] A. Lempel and J. Ziv, "Compression of two-dimensional data," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 2–8, Jan. 1986.
[13] T. Weissman and S. Mannor, "On universal compression of multidimensional data arrays using self-similar curves," in *Proc. 38th Annu. Allerton Conf. Communication, Control, and Computing*, vol. I, Oct. 2000, pp. 470–479.
[14] H. O. Georgii, *Gibbs Measures and Phase Transitions*. Berlin: Germany/New York: Walter de Gruyter, 1988.
[15] N. Merhav and M. Feder, "Universal schemes for sequential decision from individual data sequences," *IEEE Trans. Inform. Theory*, vol. 39, pp. 1280–1292, July 1993.
[16] Z. Ye and T. Berger, *Information Measures for Discrete Random Fields*. Beijing: China/New York: Science, 1998.
[17] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. New York: Springer-Verlag, 1998.
[18] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[19] G. Schwartz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 1978.
[20] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods*, 2nd ed. New York: Springer-Verlag, 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 Math.*, vol. 99, pp. 165–202, 1958.
[23] X. Guyon, *Random Fields on a Network*. New York: Springer-Verlag, 1995.
[24] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.
[25] T. Weissman, "Optimal scandiction for Markov processes," in preparation.
[26] N. Merhav and T. Weissman, "Scanning and prediction in multidimensional data arrays," Technion-I.I.T., CCIT Pub. 349, Aug. 2001.
[27] T. Weissman and N. Merhav. (2002, Feb.) On competitive prediction and its relationship to rate–distortion theory and to channel capacity theory. [Online]. Available: http://tiger.technion.ac.il/users/merhav/
[28] ——, "Universal prediction of individual binary sequences in the presence of noise," *IEEE Trans. Inform. Theory*, vol. 47, pp. 2151–2173, Sept. 2001.
[29] T. Weissman, N. Merhav, and A. Baruch, "Twofold universal pediction schemes for achieving the finite-state predictability of a noisy individual binary sequence," *IEEE Trans. Inform. Theory*, vol. 47, pp. 1849–1866, July 2001.