Generalized Forms of the Kraft Inequality for Finite-State Encoders

Neri Merhav

The Andrew & Erna Viterbi Faculty of Electrical Engineering
Technion - Israel Institute of Technology
Technion City, Haifa 3200003, ISRAEL
E–mail: `merhav@ee.technion.ac.il`

Abstract

We derive a few extended versions of the Kraft inequality for information lossless finite-state encoders. The main basic contribution is in defining a notion of a Kraft matrix and in establishing the fact that a necessary condition for information losslessness of a finite-state encoder is that none of the eigenvalues of this matrix have modulus larger than unity, or equivalently, the generalized Kraft inequality asserts that the spectral radius of the Kraft matrix cannot exceed one. For the important special case where the FS encoder is irreducible, we derive several equivalent forms of this inequality, which are based on well known formulas for spectral radius. It also turns out that in the irreducible case, Kraft sums are bounded by a constant, independent of the block length, and thus cannot grow even in any subexponential rate. Finally, two extensions are outlined - one concerns the case of side information available to both encoder and decoder, and the other is for lossy compression.

1 Introduction

Kraft's inequality plays a pivotal role in information theory. It provides a complete and elegant characterization of the feasibility of variable–length uniquely decodable (UD) codes by imposing a simple constraint on codeword lengths. In 1949, Kraft [1] introduced this inequality for prefix codes, establishing a condition on codeword lengths necessary for prefix decodability. Seven years later, McMillan [2] generalized this to UD codes, leading to the Kraft-McMillan inequality, which is widely used in information theory, first and foremost, to furnish a necessary and sufficient condition for the existence of a UD code with a given code-length function, and thereby also to prove the converse to the lossless source coding theorem, asserting that no UD source code can yield a coding rate below the entropy rate of the source. Once this necessary and sufficient condition is satisfied,

1

there exists, not only a general UD code, but also more specifically, a prefix code with that length function. Beyond its immediate operational meaning, Kraft's inequality underlies many fundamental principles in lossless compression, such as the equivalence between lossless source coding and probability assignment. In general, its importance stems from the fact that it connects combinatorial properties of codes with analytical bounds in a precise and tractable manner. Classical treatments can be found in standard texts such as [3] and [4].

When memory is introduced into the encoder, however, the classical Kraft inequality (CKI) no longer applies directly. Finite–state encoders constitute a natural and widely studied model for compression with memory, arising in universal source coding, individual–sequence coding, and finite–state prediction. In this setting, the encoder's output depends, not only on the current source symbol, but also on an internal state that evolves over time in a manner that depends on past inputs. As a result, the set of admissible codeword length assignments is no longer characterized by a single scalar inequality, and the extension of Kraft's condition becomes substantially more subtle.

Early progress in this direction was made by Ziv and Lempel [5], who derived a generalized Kraft inequality (GKI) for information–lossless (IL) finite–state (FS) encoders by considering blocks over large super-alphabets, see Lemma 2 in [5]. When reading Ziv and Lempel's article, the reader is under the impression is that their GKI was established merely an auxiliary result needed on their way of proving that the finite-state compressibility of a sequence is lower bounded by its asymptotic empirical entropy. Their focus was not on the Kraft inequality on its own right. Consequently, their formulation of Kraft's inequality suffers from two main limitations: (i) it does not reduce exactly to the CKI when the encoder has merely one state, and (ii) it is based on super-alphabet extensions to long blocks rather than being formulated in a single-letter manner, or at the level in which the encoder is defined in the first place. More precisely, while the inequality remains valid even for short block lengths, it yields tight results only asymptotically for long blocks.

Subsequent work has explored various aspects of finite–state coding, including irreducibility, asymptotic equipartition properties, and connections to entropy and prediction. Nevertheless, a direct, state–level generalization of Kraft's inequality that mirrors the simplicity and sharpness of the classical result has remained elusive.

In this paper, we present several new forms of GKI's for IL FS encoders. Our approach associates

with every given IL FS encoder a nonnegative matrix, termed the *Kraft matrix*, whose entries are determined by the encoder's single–symbol output lengths and state transitions. We show that information losslessness imposes a spectral–radius constraint on this matrix, which serves as a natural analogue of Kraft's inequality. Unlike Ziv and Lempel's GKI mentioned above, this inequality, as well as its several equivalent forms presented herein, reduces exactly to the CKI in the single–state case and avoids the use of super-alphabet extensions.

We then further refine the analysis for irreducible FS encoders, where Perron–Frobenius theory yields stronger, uniform bounds on matrix powers. These results lead to transparent lower bounds on achievable compression rates for both stochastic sources and individual sequences. In addition, we extend the framework to settings with side information (SI) available at both the encoder and decoder, where the relevant constraint is expressed in terms of the joint spectral radius (JSR) of a finite set of Kraft matrices [6]. This extension clarifies the structural limitations imposed by SI and highlights the role of common sub-invariant vectors. Finally, another extension is associated with lossy source coding in the spirit of those of [7], [8], and [9].

Overall, the proposed framework provides a unified and exact characterization of feasibility conditions for FS encoders, sharpening existing results and offering new tools for the analysis of compression and prediction under finite–memory constraints.

The outline of the remaining part of this article is as follows. In Section 2, we establish notation conventions, define the setting, and provide some background on the GKI of Ziv and Lempel. In Section 3, we present our basic GKI asserting that the spectral radius of the Kraft matrix must not exceed unity for an IL FS encoder. Stronger and more explicit statements are then provided for irreducible encoders in Section 4. In Section 5, we apply the GKI of Section 4 to obtain converse bounds on compression and prediction of irreducible machines, both in the probabilistic setting and for individual sequences. Finally, in Section 6, we extend the GKI to the case of availability of SI, and in Section 7, we extend it to the lossy case.

2   Notation, Setting and Background

Throughout this paper, scalar random variables (RV's) will be denoted by capital letters, their sample values will be denoted by the respective lower case letters, and their alphabets will be

denoted by the respective calligraphic letters. A similar convention will apply to random vectors and their sample values, which will be denoted with same symbols superscripted by the dimension. Thus, for example, X^n (n – positive integer) will denote a random n-vector $(X_1, \ldots, X_n)$, and $x^n = (x_1, \ldots, x_n)$ is a specific vector value in $\mathcal{X}^n$, the n–th Cartesian power of $\mathcal{X}$, which is the alphabet of each component of x^n. For two positive integers, i and j, where $i \le j$, x_i^j and X_i^j will designate segments $(x_i, \ldots, x_j)$ and $(X_i, \ldots, X_j)$, respectively, where for $i = 1$, the subscript will be omitted (as above). For $i > j$, x_i^j (or X_i^j) will be understood as the null string. An infinite sequence $(x_1, x_2, \ldots)$ will be denoted by $\boldsymbol{x}$. Logarithms and exponents, throughout this paper, will be understood to be taken to the base 2 unless specified otherwise. The indicator of an event $\mathcal{A}$ will be denoted by $\mathcal{I}\{\mathcal{A}\}$, i.e., $\mathcal{I}\{\mathcal{A}\} = 1$ if $\mathcal{A}$ occurs and $\mathcal{I}\{\mathcal{A}\} = 0$ if not.

Following the the finite-state encoding model of [5], an FS encoder is defined by the quintuple, $E = (\mathcal{X}, \mathcal{Y}, \mathcal{Z}, f, g)$, whose five ingredients are defined as follows: $\mathcal{X}$ is the finite alphabet of each symbol of the source sequence to be compressed. The cardinality of $\mathcal{X}$ will be denoted by α. $\mathcal{Y}$ is a finite collection of binary variable-length strings, which is allowed to consist of empty string, denoted 'null' (whose length is zero); $\mathcal{Z}$ is a finite set of s states of the encoder; $f : \mathcal{Z} \times \mathcal{X} \to \mathcal{Y}$ is the output function, and $g : \mathcal{Z} \times \mathcal{X} \to \mathcal{Z}$ is the next-state function.

Given an infinite source sequence to be compressed, $\boldsymbol{x} = (x_1, x_2, \ldots)$, with $x_i \in \mathcal{X}$, the FS encoder E produces an infinite output sequence, $\boldsymbol{y} = (y_1, y_2, \ldots)$ with $y_i \in \mathcal{Y}$, forming the compressed bit-stream, while passing through a sequence of states $\boldsymbol{z} = (z_1, z_2, \ldots)$ with $z_i \in \mathcal{Z}$, $i = 1, 2, \ldots$. The encoder is governed by the recursive equations:

$$y_i = f(z_i, x_i), \tag{1}$$

$$z_{i+1} = g(z_i, x_i), \tag{2}$$

for $i = 1, 2, \ldots$, with a fixed initial state $z_1 \in \mathcal{Z}$. If at any step $y_i = \text{null}$, this is referred to as idling as no output is generated, but only the state evolves in response to the input. At each time instant i, the encoder emits $L(y_i) = L[f(z_i, x_i)]$ bits, and it is understood that $L(\text{null}) = 0$.

An encoder with s states, henceforth called an s-state encoder, is one for which $|\mathcal{Z}| = s$. For the sake of simplicity, we adopt a few notation conventions from [5]: Given a segment of input symbols x_i^j, where i and j are positive integers with $i \le j$, and an initial state z_i, we use $f(z_i, x_i^j)$ to denote the corresponding output segment y_i^j produced by E. Similarly, $g(z_i, x_i^j)$ will denote the final state

z_{j+1} after processing the inputs x_i^j, beginning from state z_i. Thus, in response to an input x^n, the encoder produces a compressed bit string of length $L(y^n) = L[f(z_1, x^n)] = \sum_{i=1}^{n} L[f(z_i, x_i)] = \sum_{i=1}^{n} L(y_i)$ bits. An FS encoder E is called *information lossless* (IL) if, given any initial state $z_i \in \mathcal{Z}$, any positive integer n, and any input string, x_i^{i+n}, the triplet $(z_i, f(z_i, x_i^{i+n}), g(z_i, x_i^{i+n}))$ uniquely determines the corresponding input string x_i^{i+n}.

For example, a fixed-to-variable block encoder of length k can be viewed as an IL FS encoder with $s = \sum_{j=0}^{k-1} \alpha^j = \frac{\alpha^k - 1}{\alpha - 1}$ states. The initial state designates the beginning of each block. At each time instant, the state of the encoder is simply the contents of the part of the current input block received so far. In general, as long as the input block has not been completed, the encoder idles and upon receiving the last input symbol of the block, the encoder produces the compressed codeword for that block and it returns to its initial state, ready to receive the next input block. In some cases, if there is enough structure, the encoder does not necessarily have to idle until the very end of the block. For instance, if at a certain time before the end of the block, the contents of the beginning of the block read so far already determines the beginning of the compressed representation, the encoder can start to output these compressed bits before the end of the block. As an instance of such a block code of length $k = 2$, see Example 1 below.

In Lemma 2 of [5], Ziv and Lempel presented a GKI for IL FS encoders. It asserts that for every IL encoder with s states and every positive integer ℓ,

$$\sum_{x^\ell \in \mathcal{X}^\ell} 2^{-\min_{z \in \mathcal{Z}} L[f(z, x^\ell)]} \leq s^2 \left[1 + \log\left(1 + \frac{\alpha^\ell}{s^2}\right) \right]. \tag{3}$$

Ziv and Lempel's GKI was a perfect tool for their purpose of proving that the compression ratio achieved by an IL FS encoder cannot be smaller than the asymptotic empirical entropy rate (defined in [5]) for any infinite source sequence $\boldsymbol{x}$. However, when examined for finitely long sequences, and from the perspective of serving as a necessary condition for information losslessness, this inequality suffers from two main weaknesses.

1. It does not exactly recover the CKI for the special case, $s = 1$, as in that case, the r.h.s. becomes $1 + \log(1 + \alpha^\ell) > 1$. Moreover, even if $\ell = 1$, the right-hand side (r.h.s.), which is $1 + \log_2(1 + \alpha)$, is even larger than 2 for every $\alpha \geq 2$. On a related note, a close inspection of the proof of Lemma 2 in [5] reveals that the inequality in eq. (3) is actually a strong inequality ($<$), in other words, this inequality is always loose.

2. It is significant only upon an extension from single symbols into the super-alphabet of ℓ-strings for large ℓ, unlike the ordinary Kraft inequality, which is asserted in the same level that code is defined. For example, the CKI for a code that is defined in the level of single symbols of $\mathcal{X}$ is asserted in that level, i.e., $\sum_{x \in \mathcal{X}} 2^{-L[f(x)]} \leq 1$.

Our objective in this work is first and foremost, to establish another GKI for IL FS encoders that is free of the above mentioned drawbacks. In other words, for the case $s = 1$, it would recover the traditional Kraft inequality exactly, and it will be posed in the single-letter level without recourse to alphabet extensions. The latter property will enable one to verify relatively easily that this inequality holds in a given situation.

Our first proposed GKI serves as the basis for our subsequent derivations. Having derived it, we then confine attention to the subclass of irreducible IL FS encoders, namely, FS encoders for which every state can be reached from every state in a finite number of steps. For this important subclass of encoders, we provide several alternative formulations of the GKI and provide a stronger upper bound to the growth rate of the Kraft sum as function of the block length. Again, all these forms are smooth extensions of the CKI in the sense that in the special case $s = 1$, they degenerate to the CKI. Finally, we consider extensions in two directions (one at a time): the first is the case where SI is available to both encoder and decoder, and the second is the case of lossy compression.

3 The Basic Generalized Kraft Inequality

For a given IL FS encoder E with s states, let us define an $s \times s$ *Kraft matrix* K, whose (z, z') entry is given by

$$K_{zz'} = \sum_{\{x:\ g(z,x)=z'\}} 2^{-L[f(z,x)]}, \qquad (z, z') \in \mathcal{Z}^2, \tag{4}$$

where the summation over an empty set is understood as zero. Since K is a non-negative matrix, then according to Theorem 8.3.1 in [10], the spectral radius of K, $\rho(K)$, is an eigenvalue of K. Our first form of a GKI is the following.

Theorem 1. *For every IL FS encoder,*

$$\rho(K) \leq 1. \tag{5}$$

6

As can be seen, this GKI has the two desired properties we mentioned above:

1. The case $s = 1$ obviously recovers the CKI, since in this case, K degenerates to a scalar, which is nothing but the Kraft sum, $\sum_{x \in \mathcal{X}} 2^{-L[f(x)]}$, and then eq. (5) asserts that $\sum_{x \in \mathcal{X}} 2^{-L[f(x)]} \leq 1$, as desired.

2. The matrix K is defined in terms of the functions f and g only. These functions are defined in the level of the single symbols and states.

The first property sets the stage of establishing the condition $\rho(K) \leq 1$ as a *necessary condition for information losslessness* of a given FS encoder, in analogy to the fact that ordinary Kraft inequality is a necessary (and sufficient) condition for the existence of unique decodability in the case $s = 1$. Since there is no involvement of summations over super-alphabets of long vectors, this condition is relatively easy to check, similarly as the CKI, which is a necessary condition for the unique decodability (UD) property of ordinary lossless source codes.

Proof. Let $L_{\max} \triangleq \max_{z,x} L[f(z,x)]$. For every positive integer ℓ, the (z, z') entry of the ℓ-th order power, K^ℓ, is given by

$$
\begin{aligned}
[K^\ell]_{zz'} &= \sum_{z_2 \in \mathcal{Z}} \sum_{z_3 \in \mathcal{Z}} \cdots \sum_{z_\ell \in \mathcal{Z}} \prod_{i=1}^{\ell} \left(\sum_{\{x_i: \ g(z_i,x_i)=z_{i+1}\}} 2^{-L[f(z_i,x_i)]} \right) \\
&= \sum_{\{x^\ell: \ g(z,x^\ell)=z'\}} 2^{-L[f(z,x^\ell)]} \\
&= \sum_{l=0}^{\ell \cdot L_{\max}} 2^{-l} \cdot |\{x^\ell : \ L[f(z,x^\ell)] = l, \ g(z,x^\ell) = z'\}| \\
&\leq \sum_{l=0}^{\ell \cdot L_{\max}} 2^{-l} \cdot 2^l \\
&= 1 + \ell \cdot L_{\max}, \quad (6)
\end{aligned}
$$

where in the first line, $z_1 = z$ and $z_{\ell+1} = z'$, and the inequality is due to the postulated IL property (as z and z' are fixed). Alternatively, we can also bound $[K^\ell]_{zz'}$ by $1 + \log(1 + \alpha^\ell)$ using the same considerations as in the proof of Lemma 2 in [5], except that the factor s^2 is missing since z and z' are fixed. Which bound is tighter depends on $L_{\max}$. In any case, both expressions are essentially

linear in ℓ. Continuing with the first bound, it follows that

$$\sum_{z' \in \mathcal{S}} [K^\ell]_{zz'} = \sum_{x^\ell \in \mathcal{X}^\ell} 2^{-L[f(z,x^\ell)]} \leq s(1 + \ell \cdot L_{\max}). \tag{7}$$

Let $\boldsymbol{e}_z$ be a column vector of dimension s whose entries are all zero except the entry corresponding to state z, which is 1, and let $\mathbf{1}$ denote the all-one column vector of dimension s. Then, eq. (7) can be rewritten as

$$\boldsymbol{e}_z^\top K^\ell \mathbf{1} \leq s(1 + \ell \cdot L_{\max}). \tag{8}$$

To prove that $\rho(K) \leq 1$, assume conversely, that $\lambda \overset{\triangle}{=} \rho(K) > 1$. Since K has non-negative entries, the Perron–Frobenius theorem (see again Theorem 8.3.1 in [10]) guarantees that the right eigenvector v corresponding to λ has non-negative components and at least one strictly positive component. Since $\mathbf{1} = (1, \dots, 1)^T$ has strictly positive components, there exists a constant $\delta > 0$ such that $\mathbf{1} \geq \delta v$ component-wise. Multiplying by K^ℓ from the left and using the non-negativity of K, we obtain

$$K^\ell \mathbf{1} \geq \delta K^\ell v = \delta \lambda^\ell v. \tag{9}$$

Taking the z-th component yields

$$\boldsymbol{e}_z^\top K^\ell \mathbf{1} \geq \delta \lambda^\ell v_z. \tag{10}$$

For any index z with $v_z > 0$, the r.h.s. grows exponentially in ℓ since $\lambda > 1$, but this contradicts eq. (8) which establishes an upper bound that grows only linearly in ℓ. Therefore the postulate $\rho(K) > 1$ cannot hold true, and we conclude that $\rho(K) \leq 1$, which completes the proof. $\qquad\square$

Since $\rho(K) \leq 1$, it is clear that for every natural ℓ, $\rho(K^\ell) = [\rho(K)]^\ell \leq \rho(K) \leq 1$. In other words, the spectral radius of

$$K^\ell = \left\{ \sum_{\{x^\ell: \, g(z,x^\ell)=z'\}} 2^{-L[f(z,x^\ell)]} \right\}_{z,z' \in \mathcal{Z}} \tag{11}$$

is also never larger than unity, which is an extension of our GKI to super-alphabets, which is again, a smooth extension that degenerates to the CKI for $s = 1$.

Example 1. *Consider a binary source sequence and a block code of length 2, which maps the source strings 00, 01, 10, and 11, into 0, 10, 110, and 111, respectively. This code can be implemented by*

a FS encoder with $s = 3$ states, labeled 'S', 'O', and 'I', using the following functions, f and g (see also Fig. 1):

$$
\begin{aligned}
g(S,0) &= O, \\
g(S,1) &= I, \\
g(O,0) &= g(O,1) = g(I,0) = g(I,1) = S,
\end{aligned}
$$

and

$$
\begin{aligned}
f(S,0) &= null, \\
f(S,1) &= 11, \\
f(O,0) &= 0, \\
f(O,1) &= 10, \\
f(I,0) &= 0, \\
f(I,1) &= 1.
\end{aligned}
$$

State 'S' designates the start of a block. State 'O' remembers that the first input of the block was '0' and state 'I' remembers that the first input was '1'. Upon moving to state 'I', the encoder can already output '11', because the entire codeword will be either '110' or '111' if the first source symbol is '1', so the first two coded bits are '11' in either case. After state 'I', the encoder can complete the codeword according to the second input in the block. After state 'O', outputs are generated only upon receiving the second symbol. After both states 'O' and 'I', the encoder must return to state 'S' in order to start the next block. The corresponding Kraft matrix (with row and column indexing in the order of (S,O,I)) is given by:

$$
K = \begin{pmatrix} 0 & 2^{-0} & 2^{-2} \\ 2^{-1}+2^{-2} & 0 & 0 \\ 2^{-1}+2^{-1} & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0.25 \\ 0.75 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \tag{12}
$$

whose eigenvalues are 1, 0, and -1, and so the spectral radius is $\rho(K) = 1$. As can be seen, the sums of the second and third rows do not exceed unity, so when the initial state is either 'O' or 'I', the Kraft sum does not exceed 1. On the other hand, the Kraft sum corresponding to the first row (pertaining to 'S') exceeds unity. This demonstrates an important observation: The model of a general IL FS encoder is broader and more general than a model of a FS encoder for which

given every state, the encoder implements a certain prefix (or UD) code for the variety of incoming symbols. For $\ell = 100$, we find that

$$K^{100} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.75 & 0.1875 \\ 0 & 1 & 0.25 \end{pmatrix}$$

with eigenvalues are 0, 1, and 1. Here, the sums of the first and the second rows do not exceed unity, so when the initial state is either 'S' or 'O', the Kraft sum does not exceed 1. On the other hand, the Kraft sum corresponding to the third row exceeds unity, and so, the above comment with regard to K applies here too. This concludes the Example 1.
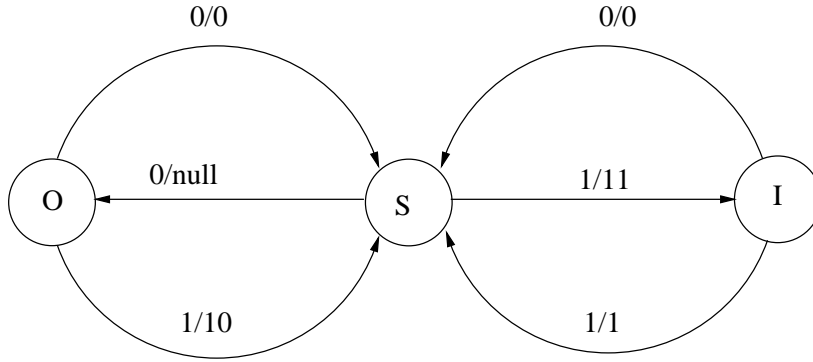

Figure 1: State transition diagram of the encoder in Example 1. The various state transitions are labeled in a form x/y, where x denotes the input and $y = f(z, x)$ denotes the output.

Earlier we said that $\rho(K) \leq 1$ is a necessary condition for a given code with next-state function g and code-lengths $\{L[f(z, x)]\}$ to be IL. One might naturally wonder whether it is also a sufficient condition. This question is open in general, but we have two comments related to this issue.

The first is that the answer is obviously affirmative for the subclass of IL encoders, which satisfies the CKI for each and every state, i.e., $\sum_{z' \in \mathcal{Z}} K_{zz'} = \sum_{x \in \mathcal{X}} 2^{-L[f(z,x)]} \leq 1$: Simply construct a separate prefix code with length function $\{L[f(z, x)], \ x \in \mathcal{X}\}$ for each $z \in \mathcal{Z}$. However, in general, an IL code does not necessarily satisfy the ordinary Kraft inequality for each z. Indeed, in Example 1, the sum of the first row of K is larger than 1.

The second comment is that we can give an affirmative answer in the level of longer blocks. Let $z_1 \in \mathcal{Z}$ be an arbitrary initial state and consider the lengths, $l(z_1, x^n) = \sum_{i=1}^{n} L[f(z_i, x_i)]$. Then,

as we have seen in (6):

$$\sum_{x^n} 2^{-L[f(z_1, x^n)]} \leq s(1 + nL_{\max}), \tag{13}$$

where the factor of s stems from taking the sum of $K_{zz'}$ over $z' \in \mathcal{Z}$. Equivalently,

$$\sum_{x^n} 2^{-[l[f(z_1, x^n)] + \log[s(1 + nL_{\max})]]} \leq 1, \tag{14}$$

and so, there exists a prefix code with lengths $l'(x^n) = l(z_1, x^n) + \log[s(1 + nL_{\max})] + \log s$, which are relatively only slightly longer than those of the original code. Here, the additional $\log s$ term is a header that notifies z_1.

4 Irreducible FS Encoders

IL FS encoders for which the next-state function g allows transition from every state to every state within a finite number of steps, are henceforth referred to as *irreducible FS encoders*. Equivalently, defining the $s \times s$ adjacency matrix A such that $A_{zz'} = 1$ whenever $\exists x \in \mathcal{X}$ such that $g(z, x) = z'$ and $A_{zz'} = 0$ otherwise, then an IL FS encoder is irreducible iff the matrix A is irreducible. Likewise, an IL FS encoder is irreducible iff the matrix K is irreducible. For an irreducible FS encoder, the shortest path from every state z to every state z' lasts no longer than $s - 1$ steps, because any longer path must visit a certain state z'' at least twice, meaning that this path contains a loop starting and ending at z'', which can be eliminated. Clearly, the encoder of Example 1 is irreducible.

Intuitively, it makes sense to use irreducible encoders, because for reducible ones, once the machine leaves a certain subset of transient states, it can never return, and so, effectively, a reducible encoders uses eventually a smaller number of states after finite time. Specifically, given a reducible machine and an infinite individual sequence $x_1, x_2, \ldots$, suppose the machine starts at a transient state. Then, there are two possibilities: either the machine quits the subset of transient states after finite time, or it stays in that subset forever. In the former case, the transient states are in use for finite time only and then never used again. In the latter case, the recurrent states are never used. In either case, asymptotically, only a subset of the available states are used, and so, effectively, the number of states actually used is smaller than s. Let $\mathcal{Z}_\infty$ denote the set of states visited infinitely many times along the sequence. This set is necessarily closed and induces a strongly connected subgraph. Consequently, the asymptotic behavior of the encoder along the given

sequence is governed entirely by its restriction to $\mathcal{Z}_\infty$, which constitutes an irreducible finite-state encoder with strictly fewer than s states. Therefore, reducible encoders cannot offer asymptotic advantages over irreducible ones, even for individual sequences.

Assume next that the next-state function g induces an irreducible matrix K^ℓ, where ℓ be an arbitrary positive integer. Since K^ℓ is non-negative and irreducible, the Collatz-Wielandt formulas [11], [12] for the spectral radius of K^ℓ hold true. These are given by

$$\rho(K^\ell) = \max_{w \in \mathcal{W}_+} \min_{\{z:\ w_z > 0\}} \frac{[K^\ell w]_z}{w_z} = \min_{w \in \mathcal{W}_+} \max_{\{z:\ w_z > 0\}} \frac{[K^\ell w]_z}{w_z}. \tag{15}$$

where w is an s-dimensional column vector and $\mathcal{W}_+$ is the set of all such vectors with non-negative components not all of which are zero. These lead to the two following GKI's:

$$\forall\ w \in \mathcal{W}_+ \ \exists\ z \text{ such that } w_z > 0 \text{ and } \sum_{z' \in \mathcal{S}} w_{z'} \cdot \sum_{\{x^\ell:\ g(z,x^\ell)=z'\}} 2^{-L[f(z,x^\ell)]} \leq w_z, \tag{16}$$

and

$$\exists\ w \in \mathcal{W}_+ \ \forall\ z \text{ such that } w_z > 0: \sum_{z' \in \mathcal{S}} w_{z'} \cdot \sum_{\{x^\ell:\ g(z,x^\ell)=z'\}} 2^{-L[f(z,x^\ell)]} \leq w_z, \tag{17}$$

The first formulation can be simplified at the price of a possible loss of tightness, by selecting w to be the all-one vector and thereby bounding $\rho(K^\ell)$ from below. This results in the conclusion that an IL FS encoder always satisfies yet another GKI:

$$\exists z \in \mathcal{S} \sum_{x^\ell \in \mathcal{X}^\ell} 2^{-L[f(z,x^\ell)]} \leq 1. \tag{18}$$

In words, for every given irreducible FS encoder, (f,g), and for every natural ℓ, there is at least one initial state, $z \in \mathcal{Z}$, for which the Kraft sum is less than unity, but again, not all states must satisfy this condition (as we saw in Example 1, the Kraft sum exceeds unity when the initial state is 'S'). All these are also smooth extensions of the CKI in the sense that for $s = 1$ we are back to the CKI.

But there is an even stronger GKI that applies to irreducible encoders. It asserts that in the irreducible case, K^n does not even grow linearly as in (6), but is rather bounded by a constant, independent of n. For $s = 1$, this constant is 1, again in agreement with the CKI.

Theorem 2. *Let K be an irreducible Kraft matrix. Then, for all $z, z' \in \mathcal{Z}$ and for every natural n,*

$$(K^n)_{zz'} = \sum_{\{x^n:\ g(z,x^n)=z'\}} 2^{-L[f(z,x^n)]} \leq 2^{(s-1)L_{\max}}. \tag{19}$$

12

Consequently, for every $z \in \mathcal{S}$,

$$\sum_{x^n \in \mathcal{X}^n} 2^{-L[f(z,x^n)]} \le s \cdot 2^{(s-1)L_{\max}}, \tag{20}$$

and

$$\sum_{z \in \mathcal{S}} \sum_{x^n \in \mathcal{X}^n} 2^{-L[f(z,x^n)]} \le s^2 \cdot 2^{(s-1)L_{\max}}. \tag{21}$$

Proof. It is sufficient to prove the first inequality, as the two other ones will follow trivially by a summation over $z' \in \mathcal{Z}$ and then also over $z \in \mathcal{Z}$, respectively. Since K is non-negative and irreducible, the Perron-Frobenius theorem applies. This theorem asserts that the spectral radius, $\rho(K)$, is positive and simple, with left and right eigenvectors, u and v, respectively, that have only strictly positive components. In Theorem 1 we have already proved that $\rho(K) \le 1$. Assume first that $\rho(K) = 1$. Then, $u^T K^n = u^T$, or, equivalently,

$$\sum_{z \in \mathcal{S}} u_z (K^n)_{zz'} = u_{z'} \quad \forall \, z' \in \mathcal{Z}. \tag{22}$$

Since all terms are non-negative, the left-hand side is lower bound by $u_z (K^n)_{zz'}$ for any $z \in \mathcal{Z}$. This implies for every $z, z' \in \mathcal{Z}$

$$(K^n)_{zz'} \le \frac{u_{z'}}{u_z} \le \frac{\max_{z \in \mathcal{Z}} u_z}{\min_{z \in \mathcal{Z}} u_z}. \tag{23}$$

Let $z_\star \in \mathcal{Z}$ and $z^\star \in \mathcal{Z}$ be achievers of $\min_{z \in \mathcal{Z}} u_z$ and $\max_{z \in \mathcal{Z}} u_z$, respectively. Then, for every $z, z' \in \mathcal{Z}$,

$$(K^n)_{zz'} \le \frac{u_{z^\star}}{u_{z_\star}}. \tag{24}$$

Since K is irreducible, the exists a path of length $\ell \le s - 1$ from $z^\star$ to $z_\star$, say, $z^\star \to z_1 \to \cdots \to z_{\ell-1} \to z_\star$ such that

$$(K^\ell)_{z^\star z_\star} \ge K_{z^\star z_1} \cdot K_{z_1 z_2} \cdots K_{z_{\ell-1} z_\star} > 0. \tag{25}$$

Since all positive entries of K are at least as large as $2^{-L_{\max}}$, this product is at least as large as $2^{-\ell L_{\max}} \ge 2^{-(s-1)L_{\max}}$. It follows then that

$$(K^\ell)_{z^\star z_\star} \ge 2^{-(s-1)L_{\max}}. \tag{26}$$

Now,

$$u_{z_\star} = \sum_{z \in \mathcal{S}} u_z (K^\ell)_{zz_\star} \ge u_{z^\star} (K^\ell)_{z^\star z_\star} \ge u_{z^\star} 2^{-(s-1)L_{\max}}, \tag{27}$$

13

which implies that

$$2^{(s-1)L_{\max}} \geq \frac{u_{z^\star}}{u_{z_\star}} \geq (K^n)_{zz'}, \tag{28}$$

for every $z, z' \in \mathcal{S}$. This completes the proof for the case $\rho(K) = 1$. The case $\rho(K) < 1$ is obtained from the case $\rho(K) = 1$ by simply defining $\hat{K} = K/\rho(K)$ and using the fact that all non-negative entries of $\hat{K}$ are lower bounded by $2^{-L_{\max}}/\rho(K)$. Since $\hat{K}$ is also irreducible and since $\rho(\hat{K}) = 1$, we now have

$$(\hat{K}^n)_{zz'} \leq [\rho(K)2^{L_{\max}}]^{s-1}. \tag{29}$$

But $\hat{K}^n = K^n/[\rho(K)]^n$, and so,

$$(K^n)_{zz'} \leq [\rho(K)]^{n+s-1} \cdot 2^{(s-1)L_{\max}} < 2^{(s-1)L_{\max}}. \tag{30}$$

This completes the proof of Theorem 2. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ $\square$

5 Converse Bounds Derived from the GKI

In this section, we demonstrate how the GKI of Section 4 can be used to obtain lower bounds on the performance of irreducible machines in compression and in prediction problems. For compression, both probabilistic sources and individual sequences are considered. For prediction, only the individual sequence version is presented, but the probabilistic counterpart can also be derived straightforwardly using the same ideas.

Let $\{P(z, x^\ell), \ z \in \mathcal{Z}, \ x^\ell \in \mathcal{X}^\ell\}$ be a joint probability distribution of random variables Z and X^ℓ. Then,

$$
\begin{aligned}
s^2 \cdot 2^{(s-1)L_{\max}} \ \geq \ & \sum_{z \in \mathcal{Z}} \sum_{x^\ell \in \mathcal{X}^\ell} 2^{-L[f(z,x^\ell)]} \\
= \ & \sum_{z \in \mathcal{Z}} \sum_{x^\ell \in \mathcal{X}^\ell} P(z, x^\ell) \cdot 2^{-L[f(z,x^\ell)] - \log P(z,x^\ell)} \\
\geq \ & \exp_2 \left\{ -\sum_{z \in \mathcal{Z}} \sum_{x^\ell \in \mathcal{X}^\ell} P(z, x^\ell) L[f(z, x^\ell)] + H(Z, X^\ell) \right\} \\
= \ & \exp_2 \left[-\boldsymbol{E}\{L[f(Z, X^\ell)]\} + H(Z, X^\ell) \right],
\end{aligned} \tag{31}
$$

where the inequality follows from Jensen's inequality and the convexity of the exponential function.

14

By taking logarithms of both sides, rearranging terms, and normalizing by ℓ, we get

$$
\begin{aligned}
R &= \frac{\boldsymbol{E}\{L[f(Z, X^\ell)]\}}{\ell} \\
&\geq \frac{H(Z, X^\ell)}{\ell} - \frac{\log_2\left(s^2 \cdot 2^{(s-1)L_{\max}}\right)}{\ell} \\
&\geq \frac{H(X^\ell)}{\ell} - \frac{2\log_2 s + (s-1)L_{\max}}{\ell},
\end{aligned}
\tag{32}
$$

and if the source P is stationary, $H(X^\ell)/\ell$ can be further lower bounded by $H(X_\ell|X^{\ell-1})$, to obtain

$$
R \geq H(X_\ell|X^{\ell-1}) - \frac{2\log_2 s + (s-1)L_{\max}}{\ell}.
\tag{33}
$$

Since this bound applies to every positive integer ℓ, we may maximize the lower bound over ℓ, and obtain

$$
R \geq \sup_{\ell \geq 1}\left\{H(X_\ell|X^{\ell-1}) - \frac{2\log_2 s + (s-1)L_{\max}}{\ell}\right\}.
\tag{34}
$$

We see that thanks to Theorem 2, the vanishing term subtracted from the entropy decays at the rate of $1/\ell$ as opposed to the $(\log \ell)/\ell$ rate that stems from Lemma 2 of [5] as well as from the more general inequality of $1 + \ell \cdot L_{\max}$, that is obtained when reducible machines are allowed.

In the context of individual sequences, we can arrive at an analogous lower bound, provided that we define a shift-invariant empirical distribution. Specifically, let x^n be a given individual sequence, let ℓ be a positive integer smaller than n, and let z_1 be a given initial state of the encoder. We assume that x^n cyclic with respect to (w.r.t.) g in the sense that $g(z_n, x_n) = z_1$. If this is not the case, consider an extension of x^n by concatenating a suffix x_{n+1}^{n+m} such that the extended sequence would be cyclic w.r.t. g. Since g is assumed irreducible, this is always possible and the length m of the extension need not be larger than $s-1$. To avoid cumbersome notation, we redefine x^n to be the sequence after the cyclic extension (if needed), and we shall keep in mind that this cyclic extension adds no more than $m \cdot L_{\max} \leq (s-1)L_{\max}$ bits to the compressed description, or equivalently, $(s-1)L_{\max}/n$ to the compression ratio, and so, this extra rate should be subtracted back upon returning to the original sequence before the cyclic extension. For every $w^\ell \in \mathcal{X}^\ell$ and $z \in \mathcal{S}$, let

$$
\delta(z_i, x_i^{((i-1)\oplus(\ell-1))+1}; z, w^\ell) = \begin{cases} 1 & z_i = z \text{ and } x_i^{((i-1)\oplus(\ell-1))+1} = w^\ell \\ 0 & \text{elsewhere} \end{cases}
\tag{35}
$$

where $\oplus$ denotes modulo-n addition. Next, define the empirical distribution

$$
\hat{P}(z, w^\ell) = \frac{1}{n}\sum_{i=1}^{n}\delta(z_i, x_i^{((i-1)\oplus(\ell-1))+1}; z, w^\ell),
\tag{36}
$$

15

Now,

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}L[f(z_i,x_i)] &= \frac{1}{n\ell}\sum_{i=1}^{n}\ell\cdot L[f(z_i,x_i)] \\
&= \frac{1}{n\ell}\sum_{i=1}^{n}\sum_{j=0}^{\ell-1}L[f(z_i,x_{((i-1)\oplus j)+1})] \\
&= \frac{1}{n\ell}\sum_{i=1}^{n}L[f(z_i,x_i^{((i-1)\oplus(\ell-1))+1})] \\
&= \frac{1}{n\ell}\sum_{i=1}^{n}\sum_{z\in\mathcal{S}}\sum_{w^\ell\in\mathcal{X}^\ell}\delta(z_i,x_i^{((i-1)\oplus(\ell-1))+1};z,w^\ell)L[f(z,w^\ell)] \\
&= \frac{1}{n\ell}\sum_{z\in\mathcal{S}}\sum_{w^\ell\in\mathcal{X}^\ell}\sum_{i=1}^{n}\delta(z_i,x_i^{((i-1)\oplus(\ell-1))+1};z,w^\ell)L[f(z,w^\ell)] \\
&= \frac{1}{\ell}\sum_{z\in\mathcal{S}}\sum_{w^\ell\in\mathcal{X}^\ell}\hat{P}(z,w^\ell)L[f(z,w^\ell)] \\
&\geq \hat{H}(X_\ell|X^{\ell-1})-\frac{2\log_2 s+(s-1)L_{\max}}{\ell}, \tag{37}
\end{aligned}
$$

where $\hat{H}(X_\ell|X^{\ell-1})$ is the empirical conditional entropy derived from the shift-invariant distribution $\hat{P}$. Using the fact that this is true for every natural $\ell<n$ and returning to the original sequence before the cyclic extension, we find that

$$
\frac{1}{n}\sum_{i=1}^{n}L[f(z_i,x_i)]\geq\max_{1\leq\ell<n}\left\{\hat{H}(X_\ell|X^{\ell-1})-\frac{2\log s+(s-1)L_{\max}}{\ell}\right\}-\frac{(s-1)L_{\max}}{n}. \tag{38}
$$

Furthermore, invoking Ziv's inequality (see eq. (13.125) in [4]), this can be further lower bounded in terms of the LZ complexity. Specifically, according to eq. (13.125) in [4], for every Markov source, $Q_{\ell-1}$, of order $\ell-1$ and every $x^n\in\mathcal{X}^n$,

$$
c(x^n)\log c(x^n)\leq-\log Q_{\ell-1}(x^n|x^0_{-(\ell-2)})+\epsilon_\ell(n), \tag{39}
$$

where $c(x^n)$ is the maximum number of distinct phrases whose concatenation forms x^n, and where $\epsilon_\ell(n)$ tends to zero at the rate of $O(\log(\log n)/\log n)$ for every fixed ℓ. By minimizing the r.h.s. w.r.t. $Q_{\ell-1}$, we get

$$
c(x^n)\log c(x^n)\leq n\hat{H}(X_\ell|X_0^{\ell-1})+n\cdot\epsilon_\ell(n), \tag{40}
$$

and so,

$$
\frac{1}{n}\sum_{i=1}^{n}L[f(z_i,x_i)]\geq\frac{c(x^n)\log c(x^n)}{n}-\min_\ell\left[\epsilon_\ell(n)+\frac{2\log s+(s-1)L_{\max}}{\ell}\right]-\frac{(s-1)L_{\max}}{n}. \tag{41}
$$

16

The minimizing ℓ can be found to be proportional to $\sqrt{n}$, but the dominant term of $\epsilon_\ell(n)$ remains of the order of $\frac{\log(\log n)}{\log n}$.

We next derive a lower bound to the prediction error of any FS predictor that is based on an irreducible FS machine. Consider a finite-state (FS) predictor with q states, defined by the following recursion, for $i = 1, 2, \ldots$

$$\begin{aligned} \hat{x}_{i+1} &= u(x_i, \sigma_i), \\ \sigma_{i+1} &= v(x_i, \sigma_i), \end{aligned} \tag{42}$$

where $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots)$, $\sigma_i \in \Sigma$, $i = 1, 2, \ldots$, is a corresponding infinite state sequence, whose alphabet, Σ, is a finite set of states of cardinality q, and $\hat{\boldsymbol{x}} = (\hat{x}_1, \hat{x}_2, \ldots)$, $\hat{x}_i \in \mathcal{X}$, $i = 1, 2, \ldots$, is the resulting predictor output sequence. Without loss of generality, the initial state, σ_1, and the initial prediction, $\hat{x}_1$, are assumed fixed members, $\sigma_\star \in \Sigma$ and $\hat{x}_\star \in \mathcal{X}$, respectively. Here, $u : \mathcal{X} \times \Sigma \to \mathcal{X}$ is the predictor output function and $v : \mathcal{X} \times \Sigma \to \Sigma$ is the next-state function.

It is assumed that $\mathcal{X}$ is a group with well-defined addition and subtraction operations. For example, if $\mathcal{X} = \{0, 1, \ldots, \alpha - 1\}$ then it is natural to equip $\mathcal{X}$ with addition and subtraction modulo α. Let $\rho : \mathcal{X} \to \mathbb{R}^+$ denote a given loss function. Then, the performance of a predictor across the time range, $1 \leq t \leq n$ is measured in terms of the time-average,

$$\frac{1}{n} \sum_{i=1}^{n} \rho(x_i - \hat{x}_i). \tag{43}$$

Given an arbitrary irreducible FS predictor (u, v) as defined above, consider the auxiliary conditional probability distribution,

$$Q_\theta(x_{i+1}|x_i, \sigma_i) = \frac{e^{-\rho(x_{i+1} - u(x_i, \sigma_i))/\theta}}{Z(\theta)}, \quad \theta \geq 0, \tag{44}$$

where

$$Z(\theta) = \sum_{x \in \mathcal{X}} e^{-\rho(x)/\theta}. \tag{45}$$

Define also the function

$$\Delta(R) = \sup_{\theta \geq 0} \theta \cdot [R - \log Z(\theta)], \quad R \geq 0. \tag{46}$$

Now, define

$$Q_\theta(x^n) = \prod_{i=1}^{n} Q_\theta(x_i|x_{i-1}, \sigma_{i-1}) \tag{47}$$

17

where σ_0 and x_0 are arbitrary members of Σ and $\mathcal{X}$, respectively, such that $\sigma_1 = v(x_0, \sigma_0) = s_\star$, and $\sigma_2, \sigma_3, \ldots, \sigma_{n-1}$ are generated from $x_1, x_2, \ldots, x_{n-1}$ as in (42).

Let k divide n and consider the lossless compression of x_0^{n-1} in blocks of length k, $\boldsymbol{x}_j = x_{jk+1}^{jk+k}$, $j = 0, 1, \ldots, n/k - 1$, by using the Shannon code, whose length function for a vector x^k is $\lceil -\log Q_\theta(x^k) \rceil$. This is equivalent to predictive coding, where the prediction error signal, $z_n = x_n - f(x_{n-1}, s_{n-1})$ is compressed losslessly under a model of a memoryless source with a marginal $Q_\theta(z)$ (see Fig. 2 for illustration). In this case, since the ceiling operation is carried over k-blocks, and there are n/k such k-blocks, the upper bound to $L(x^n)$ becomes

$$L(x^n) = \sum_{i=0}^{n/k-1} \lceil -\log Q_\theta(x_{ik+1}^{ik+k}) \rceil \leq \frac{1}{\theta} \cdot \sum_{i=1}^{n} \rho(x_i - u(x_{i-1}, \sigma_{i-1})) + n \log Z(\theta) + \frac{n}{k}. \tag{48}$$

On the other hand, the corresponding encoder of Fig. 2 can be viewed as an encoder with $q \cdot M_k$ states, where $M_k = (\alpha^k - 1)/(\alpha - 1)$, since this is the number of combinations of a state of the q-state predictor and a state of the lossless block encoder, whose number of states is $\sum_{j=0}^{k-1} \alpha^j = M_k$. Thus,

$$\frac{L(x^n)}{n} \geq \hat{H}(X^\ell | X^{\ell-1}) - \frac{2 \log(q M_k) + (q M_k - 1) L_{\max}}{\ell} - \frac{L_{\max}}{n} \tag{49}$$

where it should be kept in mind that $L_{\max}$ is expected to grow linearly with k. Thus, by comparing the upper bound and the lower bound to $L(x^n)$, we have

$$\frac{1}{n\theta} \cdot \sum_{i=1}^{n} \rho(x_i - u(x_{i-1}, \sigma_{i-1})) + \log Z(\theta) + \frac{1}{k}$$
$$\geq \quad \hat{H}(X^\ell | X^{\ell-1}) - \frac{2 \log(q M_k) + (q M_k - 1) L_{\max}}{\ell} - \frac{L_{\max}}{n}. \tag{50}$$

or, equivalently,

$$\frac{1}{n} \sum_{i=1}^{n} \rho(x_i - u(x_{i-1}, \sigma_{i-1}))$$
$$\geq \quad \theta \left[\hat{H}(X^\ell | X^{\ell-1}) - \frac{2 \log(q M_k) + (q M_k - 1) L_{\max}}{\ell} - \frac{L_{\max}}{n} - \frac{1}{k} - \log Z(\theta) \right]. \tag{51}$$

Maximizing the r.h.s over $\theta \geq 0$, we get

$$\frac{1}{n} \sum_{i=1}^{n} \rho(x_{i+1} - u(x_i, \sigma_i)) \geq \Delta \left(\hat{H}(X_\ell | X^{\ell-1}) - \frac{2 \log(q M_k) + (q M_k - 1) L_{\max}}{\ell} - \frac{L_{\max}}{n} - \frac{1}{k} \right). \tag{52}$$

The bound is meaningful if $k \gg 1$ and $\ell \gg q M_k$, so that the two subtracted terms in the argument of the function $\Delta(\cdot)$ are small compared to the main term, $\hat{H}(X_\ell | X^{\ell-1})$. It is tight essentially for

18

sequences of the form $x_i = u(x_{i-1}, \sigma_{i-1}) + z_i$, $i = 1, 2, \ldots$, where $z^n = (z_1, \ldots, z_n)$ is typical to an i.i.d. source and where the marginal empirical distribution of each z_i is close to $e^{-\rho(z)/\theta}/Z(\theta)$ for some $\theta \geq 0$.
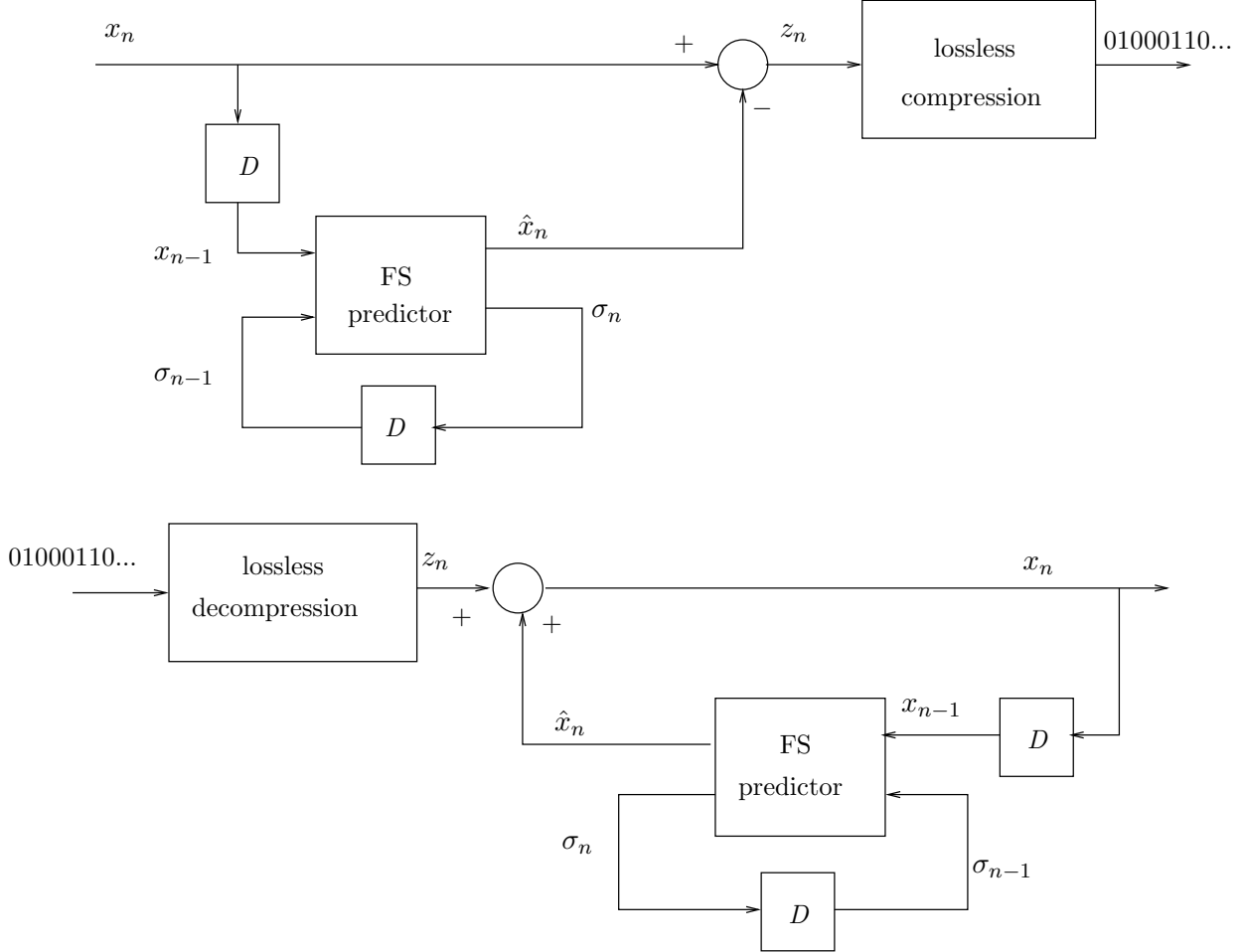

Figure 2: Auxiliary predictive encoder and decoder. The upper block diagram depicts the encoder that losslessly compresses the prediction error signal, z_n, which is the difference between the input signal, x_n, and its prediction, $\hat{x}_n$ obtained using a FS predictor. The lower block diagram stands for the corresponding decoder.

6 GKI in the Presence of Side Information

We now discuss briefly an extension of the GKI for IL FS encoders in the case where SI is available at both the encoder and the decoder. The resulting condition is expressed in terms of the *joint spectral radius* (JSR) of a finite set of nonnegative matrices indexed by the various side–information

symbols. We identify verifiable sufficient conditions for subexponential growth of Kraft sums and discuss the limitations inherent in the presence of SI.

Let $\mathcal{X}$ be the source alphabet as before and let $\mathcal{W}$ denote the finite alphabet of the SI sequence, $w_1, w_2, \ldots$, whose symbols are synchronized with the corresponding source symbols. As before, let $\mathcal{Z}$ be the finite set of states with $|\mathcal{Z}| = s$. An FS encoder with SI is specified by an output function $f : \mathcal{Z} \times \mathcal{X} \times \mathcal{W} \to \mathcal{Y}$, ($\mathcal{Y}$ being defined as a subset of $\{0, 1\}^*$, similarly as before) and a next–state function $g : \mathcal{Z} \times \mathcal{X} \times \mathcal{W} \to \mathcal{Z}$. Given an initial state, $z_1 = z$, a source sequence, $\boldsymbol{x} = (x_1, x_2, \ldots)$, and a SI sequence, $\boldsymbol{w} = (w_1, w_2, \ldots)$, the encoder implements the equations:

$$
\begin{aligned}
y_i &= f(z_i, x_i, w_i), \\
z_{i+1} &= g(z_i, x_i, w_i),
\end{aligned}
\tag{53}
$$

for $i = 1, 2, \ldots$, and the total code-length produced by the encoder after n steps is

$$
L[f(z, x^n, w^n)] = \sum_{i=1}^{n} L\big(f(z_i, x_i, w_i)\big).
\tag{54}
$$

Definition 1. *An FS encoder is said to be* information–lossless with side information *if for every n, the quadruple $(z_1, y^n, w^n, z_{n+1}) \in \mathcal{Z} \times \mathcal{Y}^n \times \mathcal{W}^n \times \mathcal{Z}$ dictates $x^n \in \mathcal{X}^n$.*

For each SI symbol, $w \in \mathcal{W}$, define the corresponding Kraft matrix

$$
[K(w)]_{zz'} = \sum_{\{x \in \mathcal{X} : \; g(z,x,w)=z'\}} 2^{-L[f(z,x,w)]}, \qquad z, z' \in \mathcal{Z}.
\tag{55}
$$

Each $K(w)$ is a nonnegative $s \times s$ matrix. For a given SI sequence, w^n, define the product matrix

$$
K(w^n) = K(w_1) \cdot K(w_2) \cdots K(w_n).
\tag{56}
$$

Now, let $\mathcal{K} = \{K(w), \; w \in \mathcal{W}\}$. The growth rate of the Kraft products, $K(w^n)$, over arbitrary SI sequences, $\{w^n\}$, is governed by the JSR of $\mathcal{K}$, which is defined as follows.

Definition 2. *The JSR of $\mathcal{K}$ is defined as*

$$
\rho_{\mathrm{JSR}}(\mathcal{K}) = \lim_{n \to \infty} \max_{w^n \in \mathcal{W}^n} \|K(w^n)\|^{1/n},
\tag{57}
$$

where $\|\cdot\|$ is any matrix norm.

It is a classical result that this limit exists and is independent of the chosen norm. The GKI in the presence of SI can be formulated as follows.

Theorem 3. *For an IL FS encoder with SI,*

$$\rho_{\text{JSR}}(\mathcal{K}) \le 1. \tag{58}$$

Proof. Fix an arbitrary SI sequence, $w^n \in \mathcal{W}^n$ and states $z, z' \in \mathcal{Z}$. The (z, z') entry of $K(w^n)$ is given by

$$\sum_{\{x^n:\ g(z, x^n, w^n) = z'\}} 2^{-L[f(z, x^n, w^n)]}. \tag{59}$$

Since the encoder is IL for the fixed sequence w^n, the mapping between x^n and (z, x^n, w^n, z') is injective over all paths from z to z'. Grouping sequences according to their total code-length (similarly as before) and using a standard counting argument yields a linear upper bound (in n) on each matrix entry of $K(w^n)$, uniformly over w^n. Exponential growth of $\|K(w^n)\|$ is therefore impossible, and the JSR must satisfy $\rho_{\text{JSR}}(\mathcal{K}) \le 1$. $\qquad\square$

The following proposition can sometimes help.

Proposition 1. *If there exists a vector $v \in \mathbb{R}^s$ with strictly positive components such that $K(w)v \le v$ for every $w \in \mathcal{W}$, then for every SI sequence w^n, $K(w^n)v \le v$, and hence the family $\{K(w^n)\}$ is uniformly bounded.*

Proof. The claim follows by induction on n. Since $v > 0$, uniform boundedness of all products implies $\rho_{\text{JSR}}(\mathcal{K}) \le 1$. $\qquad\square$

For example, if $v = \mathbf{1}$ satisfies proposition 1, this means that the Kraft sum is less than or equal to unity for every initial state and every SI sequence. In such a case, one can simply design a separate prefix code for every combination of initial state and SI sequence.

In contrast to the case without SI, bounding the spectral radius of each individual Kraft matrix $K(w)$ is necessary but insufficient to control the growth rate of arbitrary products. In other words, even if $\rho[K(w)] \le 1$ for every $w \in \mathcal{W}$ individually, the JSR may exceed unity, and in fact, may

be arbitrarily large. As an example, let ϵ be an arbitrarily small positive real and consider the matrices

$$A = \begin{pmatrix} \epsilon & \frac{1}{\epsilon} \\ 0 & \epsilon \end{pmatrix}$$

and $B = A^T$. While $\rho(A) = \rho(B) = \epsilon$, which is arbitrarily small, it turns out that $\rho(A \cdot B) = \epsilon^2 + \frac{1}{2\epsilon^2} + \sqrt{1 + \frac{1}{4\epsilon^4}} \approx \frac{1}{\epsilon^2}$, which is accordingly, arbitrarily large. The JSR is therefore the correct quantity governing feasibility.

Exact computation of the JSR is undecidable in general, even for nonnegative rational matrices. Consequently, the above result should be interpreted as a structural constraint rather than a computational criterion. Nonetheless, there is a plethora of upper and lower bounds to the JSR. Also, as mentioned earlier, the existence of a common positive sub-invariant vector provides a meaningful and verifiable sufficient condition for subexponential growth.

7   GKI for Lossy Compression

For lossy compression, we adopt a simple encoder model, where each source vector $x^\ell \in \mathcal{X}^\ell$ is first mapped into a reproduction vector $\hat{x}^\ell = Q(x^\ell) \in \hat{\mathcal{X}}^\ell$ within distortion ℓD and then $\hat{x}^\ell$ is losslessly compressed by an IL FS encoder with s states exactly as before. The latter may work in the level of single letters or in the level of ℓ-blocks. Let us define $\mathcal{B}(\hat{x}^\ell) = \{x^\ell \in \mathcal{X}^\ell : \ d(x^\ell, \hat{x}^\ell) \le \ell D\}$ and let $B_\ell = \max_{\hat{\boldsymbol{x}}} |\mathcal{B}(\hat{\boldsymbol{x}})|$. Now,

$$
\begin{aligned}
K_{zz'} &\triangleq \sum_{\{x^\ell:\ g(z,Q(x^\ell))=z'\}} 2^{-L[f(z,Q(x^\ell))]} \\
&= \sum_{\{\hat{x}^\ell:\ g(z,\hat{x}^\ell))=z'\}} \sum_{\{x^\ell:\ Q(x^\ell)=\hat{x}^\ell\}} 2^{-L[f(z,\hat{x}^\ell)]} \\
&\le \sum_{\{\hat{x}^\ell:\ g(z,\hat{x}^\ell))=z'\}} |\mathcal{B}(\hat{x}^\ell)| \cdot 2^{-L[f(z,\hat{x}^\ell)]} \\
&\le B_\ell \cdot \sum_{\{\hat{x}^\ell:\ g(z,\hat{x}^\ell))=z'\}} 2^{-L[f(z,\hat{x}^\ell)]} \\
&\triangleq B_\ell \cdot \hat{K}_{zz'},
\end{aligned}
\tag{60}
$$

and so, $K \leq B_\ell \cdot \hat{K}$ entry-wise. Now, $\hat{K}$ has all the properties that we have proved for the lossless case, it is just defined in the super-alphabet of ℓ-blocks. Since $\rho(\hat{K}) \leq 1$, we readily have:

$$\rho(K) = \rho(B_\ell \cdot \hat{K}) = B_\ell \cdot \rho(\hat{K}) \leq B_\ell. \tag{61}$$

For additive distortion measures, the quantity B_ℓ can be estimated using the method of types [13], or the Chernoff bound, or saddle-point integration [14], [15]. It is is upper bounded by $2^{\ell\Phi(D)}$, where

$$\Phi(D) = \max_{\{P_{X\hat{X}}: \; d(X,\hat{X}) \leq D\}} H(X|\hat{X}). \tag{62}$$

Thus, the corresponding GKI reads

$$\rho(K) \leq 2^{\ell\Phi(D)}. \tag{63}$$

References

[1] Kraft, L. G. *A device for quantizing, grouping, and coding amplitude modulated pulses*, Master's thesis, Department of Electrical Engineering, MIT, Cambridge, MA, U.S.A., 1949.

[2] McMillan, B. "Two inequalities implied by unique decipherability," *IEEE Trans. Inform. Theory* **1956**, vol. IT-2, pp. 115–116.

[3] Gallager, R. G. *Information Theory and Reliable Communication*, John Wiley & Sons, New York, 1968.

[4] Cover, T. M.; Thomas, J. A. *Elements of Information Theory*, John Wiley & Sons, Hoboken, NJ, U.S.A., 2006.

[5] Ziv, J.; Lempel, A. "Compression of individual sequences via variable-rate coding," *IEEE Trans. Inform. Theory* **1978**, vol. IT–24, no. 5, pp. 530–536.

[6] Rota, G. C.; Strang, G. "A note on the joint spectral radius" **1960**, *Proceedings of the Netherlands Academy*, 22: 379–381.

[7] Campbell, L. L. "Kraft inequality for decoding with respect to a fidelity criterion," *IEEE Trans. Inform. Theory* **1973**, vol. IT–19, no. 1, pp. 68–73.

[8] Merhav, N. "A comment on "A rate of convergence result for a universal D-semifaithful code"," *IEEE Trans. Inform. Theory* **1995**, vol. 41, no. 4, pp. 1200–1202.

[9] Merhav, N. "'Refinements and generalizations of the Shannon lower bound via extensions of the Kraft inequality," *Entropy* **2026**, 28(1), 76; `https://doi.org/10.3390/e28010076`

[10] Horn, R. A.; Johnson, C. R. *Matrix Analysis*, Cambridge University Press, New York, NY, U.S.A., 1985.

[11] Collatz, L. "Einschlieβungssatz für die charakteristischen Zahlen von Matrizen", *Mathematische Zeitschrift* **1942**, vol. 48, no. 1, pp. 221-226.

[12] Wielandt, H. "Unzerlegbare, nicht negative Matrizen," *Mathematische Zeitschrift* **1950**, vol. 52, no. 1, 642–648.

[13] Csiszár, I.; Körner, J. *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Second Edition, Cambridge University Press, 2011.

[14] de Bruijn, N. G. *Asymptotic Methods in Analysis*, Dover Publications, New York, 2nd Edition, 1981.

[15] Merhav, N.; Weinberger, N. "A toolbox for refined information-theoretic analyses with applications," *Foundations and Trends in Communications and Information Theory* **2025**, vol. 22, no. 1, pp. 1–184.