Fano's Inequality and Channel Capacity in the Individual-Sequence Regime

Neri Merhav

The Viterbi Faculty of Electrical and Computer Engineering
Technion – Israel Institute of Technology
Haifa 3200003, Israel
merhav@technion.ac.il

Abstract

We derive an individual-sequence analogue of Fano's inequality, replacing entropy by Lempel–Ziv (LZ) complexity and probability of error by empirical error frequency. In the setting considered, an arbitrary encoder maps a message to a channel input sequence $\{x_t\}$, the channel maps the input to an output $\{y_t\}$ in a completely arbitrary manner, and a block finite-state decoder produces an estimate of the channel input from the output. The individual-sequence Fano inequality lower-bounds the empirical symbol-error frequency in terms of the LZ complexity of the channel input conditioned on the channel output. This leads to a channel coding converse: reliable communication at a given rate is possible only if that rate does not exceed the individual-sequence LZ channel capacity, defined as the LZ mutual information maximized over channel inputs and minimized (worst-case) over a noise set. We also prove a matching achievability theorem for the class of *symbol-wise invertible* channels, in which the channel output at each time t is $y_t = f(x_t, v_t)$, v_t being a 'noise' variable, and $f(\cdot, v)$ is a bijection for every fixed v. This class includes group-additive channels $y_t = x_t \oplus v_t$ studied by Feder and coworkers. For such a channel, any rate below the LZ capacity is achievable with vanishing per-message error frequency, simultaneously for every v^n in a given noise set $\mathcal{N}$, provided $\mathcal{N}$ is closed under block permutations; this condition is satisfied by any noise set that depends on v^n only via its type class with respect to such blocks. The achievability uses a universal decoder that requires no knowledge of the channel and no feedback.

Index Terms: Lempel-Ziv algorithm, channel capacity, adversarial channel, Fano's inequality, channel coding.

1 Introduction

1.1 Background and Motivation

Fano's inequality [1] is one of the cornerstones of classical information theory. It states that for any communication system in which a message X is transmitted over a channel with output Y and decoded as $\hat{X}$, the error probability $P_e \stackrel{\Delta}{=} \Pr\{\hat{X} \neq X\}$ satisfies

$$\phi(P_e) \stackrel{\Delta}{=} h_2(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X|Y), \tag{1}$$

where $H(X|Y)$ is the conditional entropy of X given Y, $h_2(\cdot)$ is the binary entropy function, and $|\mathcal{X}|$ is the cardinality of the alphabet, $\mathcal{X}$, of the random variable X. This is the most common tool for proving converses to channel coding theorems.

In the *individual-sequence* (or *universal*) framework of Ziv and Lempel [2, 3, 4], sequences are individual deterministic objects and no probabilistic model is assumed on either the source or the channel. The role of conditional entropy rate is played by the *conditional LZ complexity* — the normalized cost of compressing $x^n = (x_1, \ldots, x_n)$ given $y^n = (y_1, \ldots, y_n)$ using the conditional version of the LZ78 algorithm [4, 5, 6, 7].

No individual-sequence analogue of Fano's inequality has appeared in the prior literature; this paper fills that gap and then derives a complete channel coding theorem — converse and achievability — in the individual-sequence framework.

1.2 Main Contributions

The main contributions of this paper are as follows.

1. Individual-sequence Fano inequality (Theorem 3.1). Let x^n be a channel input sequence that produces a channel output y^n, and let $\hat{x}^n$ be the reconstruction of x^n produced by a block finite-state decoder of block length k (to be defined precisely). Let $\hat{P}_e = \frac{1}{n}\sum_{t=1}^{n} \mathbf{1}\{\hat{x}_t \neq x_t\}$ denote the fraction of incorrectly decoded symbols of x^n. Then,

$$\phi(\hat{P}_e) \geq \bar{\rho}_k(x^n|y^n) - \Gamma_0, \tag{2}$$

where $\bar{\rho}_k(x^n|y^n)$ is the block-average conditional LZ complexity, namely, the average compression ratio of conditional LZ compression [5] of x^n in blocks of length k using y^n as side information (to be defined precisely in the sequel), $\Gamma_0 \to 0$ as $n \to \infty$ for fixed k and a fixed number of states of the decoder. We believe that this individual-sequence analogue of Fano's inequality is interesting in its own right as a universal lower bound on empirical error frequency, and we will have more to say about it later.

2. Channel coding converse (Theorem 5.2). The aforementioned Fano inequality sets the stage for a converse to a channel coding theorem that applies to a completely arbitrary channel. The channel model is $y^n = q(x^n, v^n)$ for an arbitrary deterministic function q, with $v^n \in \mathcal{N}$ a fixed but unknown individual noise sequence. As mentioned above, the decoder is assumed to be a block finite-state machine (block-FSM); Under these conditions, define the *individual-sequence LZ channel capacity* $C_k(\mathcal{N})$ as the maximin LZ mutual information over input k-block types and noise sequences in $\mathcal{N}$ (see Definition 5.1). Then, for any rate-R channel code,

$$R \leq C_k(\mathcal{N}) + \phi(\hat{P}_e) + \Gamma_0. \tag{3}$$

This is the individual-sequence analogue of the classical converse $R \leq C + \phi(P_e)$.

3. Achievability theorem (Theorem 5.3). For the class of *symbol-wise invertible* channels $y_t = f(x_t, v_t)$ (Definition 2.1) with a fixed but unknown individual noise sequence $v^n \in \mathcal{N}$, and for noise sets $\mathcal{N}$ that are closed under k-block permutations, we prove that any rate strictly below $C_k(\mathcal{N})$ (Definition 5.1) is achievable using a universal decoder that requires no knowledge of the

channel and without any feedback. The achievability is proved via a random coding argument whose conclusion is the existence of a deterministic code with vanishing per-message error frequency simultaneously for every $v^n \in \mathcal{N}$. The converse holds for the general channel model, so the capacity formula $C_k(\mathcal{N})$ is the same in both bounds; they match exactly for symbol-wise invertible channels with permutation-closed noise sets.

Together, results 1–3 constitute an individual-sequence analogue of the classical channel coding theorem: reliable transmission is possible if and only if the code rate does not exceed the LZ channel capacity. The individual-sequence framework for channel coding relates to joint source-channel coding studied in [8, 9, 10], though in those works the channel is still probabilistic (memoryless), whereas here the channel noise is treated as an individual sequence.

Fixed noise versus worst-case guarantee. The noise sequence v^n is an individual, fixed sequence that is committed before any transmission takes place — in particular, before the encoder selects a codeword. However, neither the encoder nor the decoder knows which element of $\mathcal{N}$ was selected; both know only the set $\mathcal{N}$. The code must therefore be designed to work reliably for every possible $v^n \in \mathcal{N}$ simultaneously. For this reason, the converse and achievability are stated in terms of the worst-case noise in $\mathcal{N}$: not because v^n varies during transmission, but because the code is chosen in ignorance of which fixed v^n will occur. The fixed-in-advance nature of v^n corresponds to an *oblivious* adversary: since v^n is committed before the codeword is transmitted, the adversary cannot adapt the noise to the actual transmitted sequence. It is the encoder's ignorance of which $v^n \in \mathcal{N}$ was committed — not any adaptive power of the adversary — that forces the maximin structure of the capacity: the encoder optimizes the codebook first, against the worst-case noise in $\mathcal{N}$.

1.3 Related Work

The individual-sequence approach to source-channel coding was initiated by Ziv [8], who established a separation theorem for individual source sequences transmitted over a probabilistic discrete memoryless channel (DMC). This was extended and further developed in [9, 10]. The source coding theorems of [3] are also foundational to this line of work. The conditional LZ complexity [5] and its application to universal decoding over finite-state channels [6] are the direct predecessors of this work.

The most closely related body of work is due to Feder and his collaborators [11, 12, 13, 14]. Shayevitz and Feder [11] and Lomnitz and Feder [12, 13, 14] study individual-sequence channel coding for the *modulo-additive* channel $y_t = x_t \oplus v_t$, where $\oplus$ is group addition, x_t is the channel input at time t, y_t is the output, and v^n is a fixed but unknown individual noise sequence. They characterize the capacity as the logarithm of the output alphabet size (in bits) minus the LZ complexity of the noise, and construct achievability schemes, some requiring feedback (see point (iii) below for details).

The present paper differs from this line of work in several respects.

(i) General channel model. Our channel model $y^n = q(x^n, v^n)$ strictly generalizes the modulo-additive model $y^n = x^n \oplus v^n$ of Feder *et al.*, and their results can be recovered as a special case of ours (see Section 4). In particular, our adversary model is compatible with theirs: in all these works,

the noise sequence v^n is committed before the codeword is transmitted (an *oblivious* adversary in the terminology of Langberg [25]), and does not depend on the codebook or the transmitted sequence.

(ii) Fano inequality and converse. Shayevitz and Feder [11] prove achievability only and explicitly leave the converse as an open problem. Lomnitz and Feder [12, 13, 14] prove a converse via a counting argument: fixing y^n and showing that too many codewords have high empirical mutual information with y^n forces a decoding error. This yields the qualitative conclusion that $R > C_k$ implies $\hat{P}_e$ is bounded away from zero, but does not give the quantitative form $R \leq C_k + \phi(\hat{P}_e)$. To our knowledge, no individual-sequence analogue of Fano's inequality appears in the prior literature; our Fano inequality is what enables the quantitative converse.

(iii) Lack of feedback. The achievability schemes of [11, 12] require feedback from the decoder to the encoder, while the converse of [12, 13] allows no feedback. This creates a gap in their framework: since feedback can increase capacity for channels with memory (as discussed in Section 6), their achievability may be operating above the non-feedback capacity, and the converse of [12, 13] does not bound the feedback capacity. Whether the two sides of their coding theorem actually match is therefore unclear. Our achievability requires no feedback, so our converse and achievability are directly comparable.

(iv) Relation to the general probabilistic capacity formula. Verdú and Han [15] established a general formula for the capacity of any channel (with no assumptions of stationarity, ergodicity, or memorylessness) in terms of the limsup in probability of the normalized information density $\frac{1}{n} \log \frac{dP_{Y^n|X^n}}{dP_{Y^n}}$. Their setting is probabilistic but otherwise assumption-free on the channel side, which is the closest classical analogue of our converse, where the channel is likewise an arbitrary deterministic mapping with no structural assumptions. The normalized information density of Verdú-Han is the probabilistic counterpart of our individual-sequence LZ mutual information: both measure the reduction in uncertainty about the output when the input is known, one in a distributional sense and the other in a compressibility sense. The key difference is that Verdú-Han work with a channel distribution and characterize capacity as a probabilistic limit, while here there are no distributions anywhere: the channel is a fixed deterministic mapping, and the bound holds for every individual sequence pair (x^n, y^n) without any probabilistic averaging. Our converse also requires a finite-state decoder, a condition that has no counterpart in the Verdú-Han framework.

(v) Relation to the classical AVC literature. The arbitrarily varying channel (AVC) is a classical model in which the channel is selected, for each transmitted symbol or block, by an adversary from a known family. Blackwell, Breiman, and Thomasian [19] established the random coding capacity of the AVC, showing it equals the capacity of the compound channel formed by the worst-case member of the family. Ahlswede [21] proved the fundamental *AVC dichotomy*: under the average-error criterion, the deterministic code capacity is either zero or equals the random coding capacity. The boundary between the two regimes is characterized by *symmetrizability*: Csiszár and Narayan [22] showed that the deterministic capacity equals the random coding capacity if and only if the AVC is not symmetrizable. Hughes and Thomas [24] and Csiszár and Narayan [23] extended these results to AVCs with input and state cost constraints. Langberg [25] studied the distinction between oblivious and non-oblivious adversaries, showing that obliviousness strictly limits the adversary's power.

The present paper departs from the AVC framework in two fundamental respects. First, in

the classical AVC the adversary selects the channel state according to some strategy from a *set of distributions*; here the noise v^n is a completely individual deterministic sequence with *no distribution assumed anywhere* — not even the existence of a probability space is required. Second, and more importantly, the converse is established via an *individual-sequence Fano inequality* (Theorem 3.2), a new tool with no direct counterpart in the AVC literature (where converses rely on probabilistic arguments such as the blowing-up lemma or the method of types). Symmetrizability plays no role in our setting: the maximin capacity $C_k(\mathcal{N})$ is well-defined and positive whenever the channel is symbol-wise invertible and the noise budget ρ_v is smaller than the channel input entropy, with no symmetrizability condition needed.

Eswaran, Sarwate, Sahai, and Gastpar [16] study an AVC driven by an individual state sequence z^n: for each state value $z \in \mathcal{Z}$, the channel is a DMC $W(y|x,z)$, and the state sequence is arbitrary and unknown. They show that zero-rate feedback combined with common randomness suffices to achieve rates approaching the empirical mutual information $I(P, W_{z^n})$, where W_{z^n} is the state-averaged channel. The present paper differs from [16] in three respects. First, our channel model is strictly more general: we impose no probabilistic transition structure — the channel is a fully deterministic mapping $x^n \mapsto y^n$ with an individual noise sequence, whereas their channel is a DMC for each state value. Second, we require no feedback: common randomness alone suffices, which strictly dominates their requirement of zero-rate feedback plus common randomness. Third, we provide a matching converse (Theorem 5.2), making ours a complete coding theorem; [16] establishes only an achievability result.

1.4 Paper Organization

The paper is structured around a two-tier proof strategy. Rather than working directly at the LZ complexity level, we first establish a complete coding theorem — converse and achievability — at the level of *empirical entropies*, a more tractable complexity measure defined by the empirical distribution of non-overlapping blocks of a fixed order. The converse is then lifted to the LZ level via an upper sandwiching inequality that relates block-average conditional LZ complexity to empirical conditional entropy; the achievability is proved at both levels — first at the empirical level (Theorem 4.2), and then directly at the LZ level. This two-tier structure is the reason Section 4 (empirical-entropy coding theorem) precedes Section 5 (LZ coding theorem); the latter builds directly on the former.

Section 2 establishes the model, notation, LZ complexity background, and the upper sandwiching bound of [17]. Section 3 proves the individual-sequence Fano inequality (Theorems 3.1 and 3.2) and its generalization to difference distortion measures (Theorem 3.3). Section 4 establishes the coding theorem at the empirical-entropy level: it defines the empirical capacity $\hat{C}_m(\mathcal{N})$ for general block-permutation-invariant noise sets $\mathcal{N}$ (Definition 4.1), proves the Fano-based converse (Corollary 4.1), and proves achievability at the empirical-entropy level (Theorem 4.2), with specialization to the empirical entropy ball $\mathcal{N}_{\text{emp}}(H_v)$ in Remark 4.2. Section 5 lifts these results to the LZ level: it defines $C_k(\mathcal{N})$ (Definition 5.1), proves the LZ converse (Theorem 5.2), and proves LZ achievability (Theorem 5.3). Section 5 also contains two subsections with additional results: Section 5.4 applies the full coding theorem to joint source-channel coding, establishing an individual-sequence source-channel separation theorem at both the empirical-entropy level (Lemma 5.2, Theorem 5.4) and the LZ complexity level (Remark 5.3); Section 6 discusses open problems.

2 Model, Notation, and Background

2.1 Channel Model and Problem Formulation

We consider a communication system in which a channel input sequence $x^n \in \mathcal{X}^n$ is transmitted over a channel, and the channel output is $y^n \in \mathcal{Y}^n$. The alphabets $\mathcal{X}$, $\mathcal{Y}$ are finite with sizes α, β, respectively. The noise alphabet $\mathcal{V}$ has size ν. All logarithms are base 2.

Notation. Given x^n and integers $1 \le i \le j \le n$, we use x_i^j to denote the segment $(x_i, x_{i+1}, \ldots, x_j)$. When $i = 1$, the subscript is omitted, thus $x_1^n = x^n$. The same convention applies to y^n and v^n. Calligraphic letters ($\mathcal{X}$, $\mathcal{Y}$, $\mathcal{V}$) denote the finite alphabets introduced above. The indicator function of an event $\mathcal{A}$ is written $\mathbf{1}\{\mathcal{A}\}$, and $h_2(u) \stackrel{\Delta}{=} -u \log u - (1-u)\log(1-u)$, $u \in [0,1]$, denotes the binary entropy function.

Encoder. For the converse, the encoder is an *arbitrary* mapping that maps a message into a channel input vector, $x^n \in \mathcal{X}^n$. For the achievability, the encoder is defined by a codebook (see Definition 5.3).

Channel model (converse). For the converse, the channel is a fully general deterministic function $q : \mathcal{X}^n \times \mathcal{V}^n \to \mathcal{Y}^n$, with a fixed but unknown individual noise sequence $v^n \in \mathcal{N}$:

$$y^n = q(x^n, v^n), \tag{4}$$

with no structural assumptions on q beyond determinism.

The role of the noise set $\mathcal{N}$. The noise sequence v^n is an individual, fixed sequence that does not depend on the codebook, the message, or the transmitted codeword. The noise set $\mathcal{N} \subseteq \mathcal{V}^n$ is a design parameter known to both encoder and decoder: it specifies the class of noise sequences against which the code is required to be reliable.

This model is non-trivial despite v^n being fixed, because the decoder is constrained to a block-FSM with s states of inter-block memory (see below). Had the decoder been unconstrained, then for fixed v^n, the optimal reference decoder could have canceled the noise and thereby recovered the message x^n (at least under many channel models, like the modulo-additive channel), rendering the problem trivial and uninteresting. The block-FSM constraint prevents this: for large n, the decoder cannot store v^n across blocks or exploit its structure globally. It is this computational limitation on the decoder, not uncertainty about v^n, that makes reliable communication over the channel a genuinely non-trivial problem. The maximin capacity formula (60) reflects this structure: the maximum over x^n represents the encoder's commitment based on $\mathcal{N}$ alone, and the minimum over $v^n \in \mathcal{N}$ is the worst-case guarantee the code must meet simultaneously for every possible realization of the fixed noise sequence. In the language of the AVC literature (Section 1, point (v)), this corresponds to an *oblivious* adversary: v^n is committed before the codeword is transmitted and does not depend on it. As mentioned before, this is also the model of Shayevitz and Feder [11] and Lomnitz and Feder [12, 13, 14], with whom our adversarial assumptions are compatible.

Channel model (achievability). For the achievability, two conditions are required. First, the channel must be *symbol-wise invertible* (Definition 2.1 below): the output at each time t depends only on the current input and noise symbols, and the noise can be recovered from input and output.

Second, the noise set $\mathcal{N}$ must be *closed under k-block permutations*: for every $v^n \in \mathcal{N}$ and every permutation π of the n/k consecutive k-blocks of v^n, the permuted sequence v_π^n also belongs to $\mathcal{N}$ (this is satisfied by the LZ ball $\mathcal{N}_{\mathrm{LZ}}(\rho_v)$ and by Hamming balls).

Definition 2.1 (Symbol-wise invertible channel)**.** *A deterministic channel is* symbol-wise invertible *if there exists a function $f : \mathcal{X} \times \mathcal{V} \to \mathcal{Y}$ such that $y_t = f(x_t, v_t)$ for every $t = 1, \ldots, n$, and for every fixed $v \in \mathcal{V}$ the map $f(\cdot, v) : \mathcal{X} \to \mathcal{Y}$ is a bijection.*

This class is broader than it may first appear. Three families of practical interest are:

1. **Group-additive channels.** $y_t = x_t \oplus v_t$, where $\mathcal{X} = \mathcal{Y}$ is a finite abelian group and $\oplus$ is the group operation. For every fixed v, the map $x \mapsto x \oplus v$ is a group translation, hence a bijection. This is the model studied in [11, 12, 13].
2. **Affine channels over a finite field.** $\mathcal{X} = \mathcal{Y} = \mathbb{F}_q$, $y_t = a \cdot x_t + v_t$ for a fixed nonzero field element $a \in \mathbb{F}_q^*$ and noise $v_t \in \mathbb{F}_q$. For every fixed v, the map $x \mapsto ax + v$ is an affine bijection. This subsumes the group-additive case ($a = 1$) and captures channels with a multiplicative gain.
3. **Permutation channels.** $\mathcal{X} = \mathcal{Y}$, $y_t = \pi_{v_t}(x_t)$, where $\{\pi_v\}_{v \in \mathcal{V}}$ is an arbitrary family of permutations of $\mathcal{X}$ (not necessarily forming a group or even a coset). The noise symbol v_t selects which permutation is applied at time t, independently of all others. This is strictly more general than group-additive channels, since the permutations π_v need not be group translations.

Decoder (converse vs. achievability). A *decoder* maps the channel output sequence $y^n \in \mathcal{Y}^n$ to an estimate $\hat{x}^n \in \mathcal{X}^n$ of the transmitted channel input sequence. For the *converse*, the decoder is a *block finite-state machine (block-FSM)* with s states and block length m. At each block step $i = 0, 1, \ldots, n/m - 1$, it implements the pair of equations:

$$\hat{x}_{im+1}^{(i+1)m} = f(z_i, y_{im+1}^{(i+1)m}), \qquad z_{i+1} = g(z_i, y_{im+1}^{(i+1)m}), \tag{5}$$

where $z_0, z_1, z_2, \ldots$ take values in a finite set $\mathcal{Z}$ of size s, and z_0 is fixed.

For the *achievability*, the decoder is the k-block resetting conditional LZ decoder (Definition 5.2), which selects the codeword minimizing the block-average conditional LZ complexity and requires no knowledge of the channel.

Decoder state count. The converse decoder (block length m) gives a state-count correction term of the form $(\log s)/m$, which vanishes as $m \to \infty$ for any fixed s. The achievability decoder (block length k) has $s = 1$, so its correction term vanishes identically.

The block-FSM model strictly generalizes the classical symbol-by-symbol FSM: any ordinary FSM run for m steps can be viewed as an m-block FSM (by taking the m-step product of its transition and output functions), but the converse is false — a block-FSM output function $f : \mathcal{Z} \times \mathcal{Y}^m \to \mathcal{X}^m$ can implement arbitrary block-to-block mappings that no symbol-by-symbol FSM with the same state space can realize. At first glance, the condition $\log s \ll m$ might appear to limit the converse's generality, since a machine that accumulates memory over past blocks would

need state cardinality exponential in m, making $(\log s)/m = O(1)$ and the correction term non-vanishing. The key point is that the condition is not a restriction on the decoder's computational power *within* a block — the output function f may be arbitrarily complex — but solely on the amount of *inter-block memory*.

Error measure and key function. The *empirical symbol-error frequency* is:

$$\hat{P}_e \triangleq \frac{1}{n}\sum_{t=1}^{n} \mathbf{1}\{\hat{x}_t \neq x_t\}. \tag{6}$$

The function ϕ, which plays the role of the right-hand side of Fano's inequality (1), is:

$$\phi(p) \triangleq h_2(p) + p\log(|\mathcal{X}| - 1). \tag{7}$$

It is strictly increasing on $[0, 1 - 1/|\mathcal{X}|]$, with maximum $\log\alpha$ and hence invertible in this range, with inverse denoted $\phi^{-1} : [0, \log|\mathcal{X}|] \to [0, 1 - 1/|\mathcal{X}|]$.

2.2 Normalized Empirical Entropies

Proofs use normalized empirical entropies as intermediate quantities. For positive integer m dividing n, let $\hat{P}(x^m, y^m) = \frac{m}{n}\sum_{i=0}^{n/m-1} \mathbf{1}\{x_{im+1}^{im+m} = x^m\} \cdot \mathbf{1}\{y_{im+1}^{im+m} = y^m\}$ and define

$$\hat{H}_m(x^n) \triangleq -\frac{1}{m}\sum_{x^m} \hat{P}(x^m)\log\hat{P}(x^m). \tag{8}$$

The conditional version is $\hat{H}_m(x^n|y^n) \triangleq -\frac{1}{m}\sum_{x^m,y^m} \hat{P}(x^m, y^m)\log\hat{P}(x^m|y^m)$, and the mutual information is $\hat{I}_m(x^n; y^n) \triangleq \hat{H}_m(x^n) - \hat{H}_m(x^n|y^n)$. Sometimes a more convenient notation would be the following: let X^m and Y^m be an auxiliary random m-vectors jointly distributed according to $\hat{P}$. Then $\hat{H}_m(x^n) = \hat{H}(X^m)/m$ and $\hat{H}_m(x^n|y^n) = \hat{H}(X^m|Y^m)/m$, where $\hat{H}(\cdot)$ and $\hat{H}(\cdot|\cdot)$ denote respectively the Shannon entropy and conditional entropy of the auxiliary random vectors under $\hat{P}$. Similar notation rules will be applied also for other random vectors.

3 An Individual-Sequence Analogue of Fano's Inequality

3.1 Basic Form

Theorem 3.1 (Individual-sequence Fano inequality, basic form)**.** *Given $x^n, \hat{x}^n \in \mathcal{X}^n$ and a positive integer m dividing n:*

$$\hat{H}_m(x^n|\hat{x}^n) \leq \phi(\hat{P}_e). \tag{9}$$

Proof. Let

$$\hat{P}_e[j] \triangleq \frac{m}{n}\sum_{i=0}^{n/m-1} \mathbf{1}\{x_{im+j} \neq \hat{x}_{im+j}\}, \quad j = 1, \ldots, m, \tag{10}$$

so that $\frac{1}{m}\sum_{j=1}^{m}\hat{P}_e[j] = \hat{P}_e$. Using the auxiliary-vector notation, the chain rule and the fact that conditioning reduces entropy yield

$$\hat{H}_m(x^n|\hat{x}^n) = \frac{\hat{H}(X^m|\hat{X}^m)}{m} = \frac{1}{m}\sum_{j=1}^{m}\hat{H}(X_j|X_1^{j-1},\hat{X}^m) \le \frac{1}{m}\sum_{j=1}^{m}\hat{H}(X_j|\hat{X}_j). \tag{11}$$

Applying the classical Fano inequality to each term:

$$\begin{aligned}
\hat{H}_m(x^n|\hat{x}^n) &\le \frac{1}{m}\sum_{j=1}^{m}\hat{H}(X_j|\hat{X}_j) \\
&\le \frac{1}{m}\sum_{j=1}^{m}\phi(\hat{P}_e[j]) \\
&\le \phi\left(\frac{1}{m}\sum_{j=1}^{m}\hat{P}_e[j]\right) \\
&= \phi(\hat{P}_e), \quad \square
\end{aligned} \tag{12}$$

where the last inequality is by the concavity of $\phi(\cdot)$.

3.2 Fano's Inequality in Terms of the Channel Output

Theorem 3.2 (Individual-sequence Fano inequality, output form)**.** *Let $x^n \in \mathcal{X}^n$, $y^n \in \mathcal{Y}^n$ be given and let $\hat{x}^n$ be the output of a block-FSM decoder with block length m (dividing n) and s states, applied to y^n. Then:*

$$\phi(\hat{P}_e) \ge \hat{H}_m(x^n|y^n) - \frac{\log s}{m}. \tag{13}$$

For $m \gg \log s$, the term $(\log s)/m$ is small and the main term remains $\hat{H}_m(x^n|y^n)$.

Proof. At step i the decoder reads $y_{im+1}^{(i+1)m}$, produces $\hat{x}_{im+1}^{(i+1)m} = f(z_i, y_{im+1}^{(i+1)m})$, and updates $z_{i+1} = g(z_i, y_{im+1}^{(i+1)m})$. Define

$$\hat{P}(z, x^m, y^m, \hat{x}^m) \triangleq \frac{m}{n}\sum_{i=0}^{n/m-1}\mathbf{1}\{z_i = z,\ x_{im+1}^{(i+1)m} = x^m,\ y_{im+1}^{(i+1)m} = y^m,\ \hat{x}_{im+1}^{(i+1)m} = \hat{x}^m\}. \tag{14}$$

Let $(Z, X^m, Y^m, \hat{X}^m)$ denote the corresponding random variables under $\hat{P}$ using the alternative notation rules. By the relation $\hat{X}^m = f(Z, Y^m)$

$$\phi(\hat{P}_e) \ge \frac{\hat{H}(X^m|\hat{X}^m)}{m} \tag{15}$$

$$\ge \frac{\hat{H}(X^m|\hat{X}^m, Z, Y^m)}{m} \tag{16}$$

$$= \frac{\hat{H}(X^m|Z, Y^m)}{m} \tag{17}$$

$$= \frac{\hat{H}(X^m|Y^m)}{m} - \frac{\hat{I}(Z;\, X^m \mid Y^m)}{m} \tag{18}$$

$$\geq \frac{\hat{H}(X^m|Y^m)}{m} - \frac{\hat{H}(Z|Y^m)}{m} \tag{19}$$

$$\geq \frac{\hat{H}(X^m|Y^m)}{m} - \frac{\log s}{m} \tag{20}$$

$$= \hat{H}_m(x^n|y^n) - \frac{\log s}{m}. \tag{21}$$

□

3.3 Generalized Fano Inequality for Difference Distortion Measures

The individual-sequence Fano inequality established above is interesting in its own right, beyond its role in the channel coding converse. Before proceeding to the converse, we pause to examine it from a somewhat broader perspective by generalizing it to the setting of general difference distortion measures. This generalization shows that the Fano inequality captures a structural property of empirical entropy that goes beyond the channel coding context.

The Hamming distortion frequency, $\hat{P}_e$, is a natural distortion measure when $\mathcal{X}$ is an unstructured alphabet, but in many settings the alphabet members carry quantitative values — for example, $\mathcal{X} = \mathbb{Z}_m$ with a meaningful notion of distance between symbols. In such cases a finer distortion measure, one that penalizes large errors more than small ones, is more informative than the raw fraction of incorrect symbols. We now extend Theorem 3.2 to a general *difference distortion measure*. Suppose $\mathcal{X}$ is a finite abelian group and $\Delta : \mathcal{X} \to [0, \infty)$ satisfies $\Delta(0) = 0$. The distortion between x and $\hat{x}$ is defined as $\Delta(x - \hat{x})$. The empirical average distortion is

$$\bar{\Delta}(x^n, \hat{x}^n) \triangleq \frac{1}{n} \sum_{t=1}^{n} \Delta(x_t - \hat{x}_t). \tag{22}$$

Define the *maximum-entropy function*

$$\Phi(D) \triangleq \max\big\{H(Z) : Z \text{ r.v. on } \mathcal{X},\ \mathbb{E}[\Delta(Z)] \leq D\big\}, \tag{23}$$

which is concave and non-decreasing in $D \geq 0$, with $\Phi(0) = 0$.

Theorem 3.3 (Generalized Fano inequality for difference distortion)**.** *Under the same hypotheses as Theorem 3.2:*

$$\Phi(\bar{\Delta}(x^n, \hat{x}^n)) \geq \hat{H}_m(x^n|y^n) - \frac{\log s}{m}. \tag{24}$$

Proof. Write $e_t \triangleq x_t - \hat{x}_t$ for the error symbol at time t, so $\bar{\Delta}(x^n, \hat{x}^n) = \frac{1}{n}\sum_t \Delta(e_t)$. Since $x_t = \hat{x}_t + e_t$, knowing $\hat{x}^n$ reduces estimating x^n to estimating e^n:

$$\hat{H}_m(x^n|\hat{x}^n) = \hat{H}_m(e^n|\hat{x}^n) \leq \hat{H}_m(e^n) \leq \hat{H}_1(e^n), \tag{25}$$

where: the first step is an equality because $x^n \mapsto e^n = x^n - \hat{x}^n$ is a bijection given $\hat{x}^n$; the second inequality is data-processing (dropping conditioning); the third uses $\hat{H}_m \leq \hat{H}_1$, which follows by

the facts that conditioning reduces entropy and concavity of the entropy. Since $\frac{1}{n}\sum_t \Delta(e_t) = \frac{1}{n}\sum_t \Delta(x_t - \hat{x}_t)$, by the definition of $\Phi(\cdot)$,

$$\hat{H}_m(x^n|\hat{x}^n) \le \hat{H}_1(e^n) \le \Phi\big(\bar{\Delta}(x^n, \hat{x}^n)\big). \tag{26}$$

But from the proof of Theorem 3.2,

$$\hat{H}_m(x^n|\hat{x}^n) \ge \hat{H}_m(x^n|y^n) - \frac{\log s}{m}, \tag{27}$$

and so,

$$\hat{H}_m(x^n|y^n) \le \Phi\big(\bar{\Delta}(x^n, \hat{x}^n)\big) + \frac{\log s}{m}. \tag{28}$$

□

4 Empirical Channel Capacity

4.1 Capacity Definition and Converse

Definition 4.1 (Empirical channel mutual information and capacity)**.** *For channel input x^n, output y^n, and block length m:*

$$\hat{I}_m(x^n; y^n) \triangleq \hat{H}_m(x^n) - \hat{H}_m(x^n|y^n). \tag{29}$$

For a given noise set $\mathcal{N}$, the empirical channel capacity of order m *is:*

$$\hat{C}_m(\mathcal{N}) \triangleq \max_{x^n \in \mathcal{X}^n} \min_{v^n \in \mathcal{N}} \hat{I}_m(x^n; y^n), \tag{30}$$

where $y^n = q(x^n, v^n)$ is the channel output (4)*.* ◊

This is the maximin of the empirical mutual information over inputs and noise sequences in $\mathcal{N}$. The per-realization capacity $C_m(v^n) \triangleq \max_{x^n} \hat{I}_m(x^n; y^n)$ for the specific v^n satisfies $C_m(v^n) \ge \hat{C}_m(\mathcal{N})$ for every $v^n \in \mathcal{N}$.

Remark 4.1 (Input constraints)**.** *Input constraints are handled by restricting the maximum in* (30) *to a set $\mathcal{A} \subseteq \mathcal{X}^n$:*

$$\hat{C}_m(\mathcal{A}, \mathcal{N}) \triangleq \max_{x^n \in \mathcal{A}} \min_{v^n \in \mathcal{N}} \hat{I}_m(x^n; y^n). \tag{31}$$

The converse (Corollary 4.1) extends to the constrained case with the same proof, since Theorem 3.2 holds for any x^n regardless of $\mathcal{A}$. The LZ-level extension is discussed in Section 5.

For the achievability, the constraint must be compatible with the probabilistic method: the codebook is drawn from the appropriate type class of the capacity-achieving x^n (first-order for memoryless constraints, r-th order for memory-r constraints, or the finite-state type of Csiszár [18] for automaton-defined constraints such as (d, k) RLL codes). In every case the type class size is exponential in n (with at most a polynomial correction factor), so the achievability proof goes through with the polynomial prefactor replacing $(n+1)^\alpha$. ◊

Theorem 4.1 (Converse Theorem in empirical-entropy terms)**.** *Under the conditions of Theorem 3.2, suppose a codebook $\mathcal{C}_n$ of $M = 2^{nR}$ codewords achieves*

$$\max_{x^n \in \mathcal{C}_n} \max_{v^n \in \mathcal{N}} \frac{1}{n} \sum_{t=1}^{n} \mathbf{1}\{\hat{x}_t(q(x^n, v^n)) \neq x_t\} \;\leq\; \delta, \tag{32}$$

i.e., the worst-case symbol-error frequency over all codewords and all $v^n \in \mathcal{N}$ is at most δ. Then:

$$R \leq \hat{C}_m(\mathcal{N}) + \phi(\delta) + \frac{\log s}{m} + o(1), \tag{33}$$

where $o(1) \to 0$ as $n \to \infty$ for fixed m.

Proof. Any codebook of $M = 2^{nR}$ distinct codewords over $\mathcal{X}^n$ has at most $(n/m+1)^{\alpha^m}$ distinct order-m empirical types. By the pigeonhole principle, there exists a type P_{X^m} such that at least $2^{nR-o(n)}$ codewords share type P_{X^m}. Since all these codewords are distinct elements of the type class $\mathcal{T}_m(P_{X^m})$, and $|\mathcal{T}_m(P_{X^m})| \leq 2^{nH(P_{X^m})/m}$ by the standard multinomial estimate, we get:

$$R \leq \frac{H(P_{X^m})}{m} + o(1). \tag{34}$$

Fix this type P_{X^m} and any codeword x^n of this type. Let $v_*^n \triangleq \arg\min_{v^n \in \mathcal{N}} \hat{I}_m(x^n; q(x^n, v^n))$. Since the code achieves $\hat{P}_e \leq \delta$ for every $v^n \in \mathcal{N}$, applying Theorem 3.2 to $(x^n, q(x^n, v_*^n))$:

$$\begin{aligned}
\phi(\delta) \geq \phi(\hat{P}_e(v_*^n)) &\geq \hat{H}_m(x^n|y_*^n) - \frac{\log s}{m} \\
&= \hat{H}_m(x^n) - \hat{I}_m(x^n; y_*^n) - \frac{\log s}{m} \\
&\geq \frac{H(P_{X^m})}{m} - \hat{C}_m(\mathcal{N}) - \frac{\log s}{m} \\
&\geq R - \hat{C}_m(\mathcal{N}) - \frac{\log s}{m} - o(1),
\end{aligned} \tag{35}$$

□

4.2 Achievability in Empirical-Entropy Terms

Before we state and prove achievability, a few preparatory steps are in order. The first is the following simple lemma.

Lemma 4.1. *Let the channel be symbol-wise invertible (Definition 2.1). For every $x^n \in \mathcal{X}^n$ and $v^n \in \mathcal{V}^n$ (with $y^n = q(x^n, v^n)$):*

$$\hat{H}_m(v^n|x^n) = \hat{H}_m(y^n|x^n), \tag{36}$$

and consequently:

$$\hat{I}_m(x^n; y^n) = \hat{H}_m(y^n) - \hat{H}_m(v^n|x^n). \tag{37}$$

Proof. For fixed x^n, the map $v_t \mapsto y_t = f(x_t, v_t)$ is a bijection for each x_t (Definition 2.1). Within any m-block, $(v^m, x^m) \mapsto (y^m, x^m)$ acts symbol-wise as a bijection in $v \leftrightarrow y$ for fixed x, so $H(\hat{P}_{V^m|X^m}) = H(\hat{P}_{Y^m|X^m})$. Averaging over blocks gives (36), and (37) follows from $\hat{I}_m(x^n; y^n) = \hat{H}_m(y^n) - \hat{H}_m(y^n|x^n)$. □

The second preparatory step is provided by the following lemma.

Lemma 4.2 (Constancy of the minimum empirical mutual information)**.** *Let the channel be symbol-wise invertible (Definition 2.1), let $\mathcal{N} \subseteq \mathcal{V}^n$ be any noise set that is closed under m-block permutations (i.e., if $v^n \in \mathcal{N}$ and π is any permutation of the n/m blocks of size m, then $\pi(v^n) \in \mathcal{N}$), and let P_{X^m} be any order-m type on $\mathcal{X}^m$. For every $x^n \in \mathcal{T}_m(P_{X^m})$, define*

$$\hat{I}^*_m(x^n) \triangleq \min_{v^n \in \mathcal{N}} \hat{I}_m(x^n; q(x^n, v^n)). \tag{38}$$

Then $\hat{I}^*_m(x^n)$ takes the same value for all $x^n \in \mathcal{T}_m(P_{X^m})$; we denote this common value $\hat{I}^*_m(\mathcal{N}, P_{X^m})$.

Proof. For any $x^n, \tilde{x}^n \in \mathcal{T}_m(P_{X^m})$, there is an m-block permutation π with $\pi(x^n) = \tilde{x}^n$. For any $v^n \in \mathcal{N}$, set $\tilde{v}^n = \pi(v^n)$; by the permutation-closure assumption, $\tilde{v}^n \in \mathcal{N}$ as well. Write $\tilde{y}^n = q(\tilde{x}^n, \tilde{v}^n)$. Since q acts symbol-wise, applying π to every block gives $\tilde{y}^n = \pi(y^n)$ where $y^n = q(x^n, v^n)$. The empirical mutual information depends only on the joint empirical distribution of m-blocks, which is preserved by any m-block permutation, so $\hat{I}_m(\tilde{x}^n; \tilde{y}^n) = \hat{I}_m(x^n; y^n)$. The map $v^n \mapsto \tilde{v}^n = \pi(v^n)$ is a bijection on $\mathcal{N}$, so

$$\hat{I}^*_m(\tilde{x}^n) = \min_{\tilde{v}^n \in \mathcal{N}} \hat{I}_m(\tilde{x}^n; \tilde{y}^n) = \min_{v^n \in \mathcal{N}} \hat{I}_m(x^n; y^n) = \hat{I}^*_m(x^n). \tag{39}$$

Since x^n and $\tilde{x}^n$ were arbitrary elements of $\mathcal{T}_m(P_{X^m})$, the minimum is constant on $\mathcal{T}_m(P_{X^m})$. □

Third, we need to define the decoder. The decoder to be used in the achievability is the *Minimum Empirical Conditional Entropy (MECE) decoder*:

$$\hat{w} \triangleq \arg \min_{w \in \mathcal{M}_n} \hat{H}_m(x^n(w) \mid y^n), \tag{40}$$

where $x^n(w)$ is the codeword pertaining to message number w and $\mathcal{M}_n = \{1, 2, \ldots, 2^{nR}\}$.

We are now ready to assert and prove the achievability theorem.

Theorem 4.2 (Achievability)**.** *Let the channel be symbol-wise invertible (Definition 2.1), and let $\mathcal{N} \subseteq \mathcal{V}^n$ be any noise set that is closed under m-block permutations. Let P_{X^m} be any order-m type on $\mathcal{X}^m$ (the capacity-achieving type $P^*_{X^m}$ is the natural choice, but the achievability holds for any P_{X^m}), and let $\hat{I}^*_m(\mathcal{N}, P_{X^m})$ be the constant value guaranteed by Lemma 4.2. Suppose $R < \hat{I}^*_m(\mathcal{N}, P_{X^m})$. A randomized code of rate R with $M = 2^{nR}$ codewords drawn independently and uniformly from $\mathcal{T}_m(P_{X^m})$, decoded by the MECE decoder* (40)*, achieves:*

$$\mathbb{E}_\omega\{\hat{P}_e\} \to 0 \quad \text{as } n \to \infty, \tag{41}$$

simultaneously for every $v^n \in \mathcal{N}$ and every message w. Consequently, $\hat{C}_m(\mathcal{N})$ is achievable.

Proof. By Lemma 4.2, $\hat{I}_m^*(\mathcal{N}, P_{X^m})$ is well-defined as a constant on $\mathcal{T}_m(P_{X^m})$. Let $R = \hat{I}_m^*(\mathcal{N}, P_{X^m}) - \epsilon$ for some $\epsilon > 0$. For the transmitted codeword $x^n = x^n(w, \omega) \in \mathcal{T}_m(P_{X^m})$ and the actual $v^n \in \mathcal{N}$, the MECE decoder assigns it score $\hat{H}_m(x^n \mid q(x^n, v^n))$. By the chain rule:

$$\hat{H}_m(x^n \mid y^n) = \hat{H}_m(x^n) - \hat{I}_m(x^n; y^n) \le \frac{H(P_{X^m})}{m} - \hat{I}_m^*(\mathcal{N}, P_{X^m}). \tag{42}$$

For fixed y^n, define the confusable set with the same threshold:

$$\mathcal{C}(y^n) \triangleq \big\{\tilde{x}^n \in \mathcal{T}_m(P_{X^m}) : \hat{H}_m(\tilde{x}^n \mid y^n) \le \tfrac{1}{m} H(P_{X^m}) - \hat{I}_m^*(\mathcal{N}, P_{X^m})\big\}. \tag{43}$$

By the method of types,

$$|\mathcal{C}(y^n)| \le \exp_2\{n[H(P_{X^m})/m - \hat{I}_m^*(\mathcal{N}, P_{X^m})) + o(n)]\} \tag{44}$$

and $|\mathcal{T}_m(P_{X^m})| \ge \exp_2\{nH(P_{X^m})/m\}/(n/m+1)^{\alpha^m}$. For any wrong codeword $x^n(w', \omega)$, $w' \ne w$, drawn uniformly from $\mathcal{T}_m(P_{X^m})$ independently of y^n:

$$\begin{aligned}
\Pr_{\omega}\big\{x^n(w', \omega) \in \mathcal{C}(y^n)\big\} &= \frac{|\mathcal{C}(y^n)|}{|\mathcal{T}_m(P_{X^m})|} \\
&\le \frac{2^{n(\frac{1}{m}H(P_{X^m}) - \hat{I}_m^*(\mathcal{N}, P_{X^m}))} \cdot (n/m+1)^{\alpha^m \beta^m}}{2^{n \cdot \frac{1}{m} H(P_{X^m})}/(n/m+1)^{\alpha^m}} \\
&= 2^{-n\hat{I}_m^*(\mathcal{N}, P_{X^m})} \cdot (n/m+1)^{\alpha^m(\beta^m+1)}.
\end{aligned} \tag{45}$$

Since m is fixed, $(n/m+1)^{\alpha^m(\beta^m+1)} = 2^{o(n)}$. A decoding error requires at least one wrong codeword in $\mathcal{C}(y^n)$, so by the union bound over $M = 2^{nR} - 1$ codewords:

$$\begin{aligned}
\Pr_{\omega}\{\text{error} \mid w, v^n\} &\le 2^{nR} \cdot 2^{-n\hat{I}_m^*(\mathcal{N}, P_{X^m})} \cdot (n/m+1)^{\alpha^m(\beta^m+1)} \\
&= 2^{-n(\hat{I}_m^*(\mathcal{N}, P_{X^m}) - R)} \cdot (n/m+1)^{\alpha^m(\beta^m+1)} \\
&= 2^{-n\epsilon} \cdot 2^{o(n)} \to 0.
\end{aligned} \tag{46}$$

The bound holds for every $v^n \in \mathcal{N}$ (since $\hat{I}_m(x^n; y^n) \ge \hat{I}_m^*(\mathcal{N}, P_{X^m})$ for all such v^n) and every message w, so $\mathbb{E}_\omega\{\hat{P}_e\} \to 0$. □

The coding theorem identifies the empirical capacity $\hat{C}_m(\mathcal{N})$ as the threshold for reliable communication, but leaves it as an implicit maximum over input types and minimum over noise sequences. However, in certain special cases, it is possible to derive a simpler and more explicit capacity formula. One such special case is the case where $\mathcal{N}$ is defined by the set of all $\{v^n\}$ whose empirical entropy does not exceed a given threshold, say,

$$\mathcal{N} = \mathcal{N}_{\text{emp}}(H_v) \triangleq \{v^n \in \mathcal{V}^n : \ \hat{H}_m(v^n) \le H_v\}, \tag{47}$$

where H_v is a given constant in the interval $[0, \log \mathcal{V}]$. In this case, we next show that

$$\lim_{m\to\infty} \lim_{n\to\infty} \hat{C}_m(\mathcal{N}_{\text{emp}}(H_v)) = \log \alpha - H_v. \tag{48}$$

To prove this statement, the following lemma is the key tool for evaluating this maximum explicitly for the empirical entropy noise ball: it shows that for a fixed x^n, almost all noise sequences v^n of a given type have empirical conditional entropy $\hat{H}_m(v^n|x^n)$ close to the unconditional type entropy $H(P_{V^m})$, i.e., conditioning on x^n essentially does not reduce the empirical entropy. This decoupling of v^n from x^n at the empirical level is what allows the explicit formula $\hat{C}_m(\mathcal{N}_{\text{emp}}(H_v)) = \log\alpha - H_v$ to be derived in Remark 4.2 below.

Lemma 4.3 (Empirical Independence)**.** *Let $\mathcal{X}$, $\mathcal{V}$ be finite alphabets. Fix any $x^n \in \mathcal{X}^n$ and any m-block type P_{V^m} on $\mathcal{V}^m$. Let $N = n/m$ (assumed integer). For any $\varepsilon > 0$:*

$$\frac{\#\{v^n \in \mathcal{T}(P_{V^m}) : \hat{H}_m(v^n \mid x^n) < H(P_{V^m}) - \varepsilon\}}{|\mathcal{T}(P_{V^m})|} \;\le\; (N+1)^{|\mathcal{X}|^m|\mathcal{V}|^m+|\mathcal{V}|^m} \cdot 2^{-N\varepsilon}, \tag{49}$$

which vanishes exponentially in N. Moreover, $\hat{H}_m(v^n \mid x^n) \le H(P_{V^m})$ holds for every *$v^n \in \mathcal{T}(P_{V^m})$ without exception.*

Proof. Treat m-blocks as super-letters; there are $N = n/m$ super-letter positions.

Upper bound (exact, no exceptions). Since conditioning reduces empirical entropy, $\hat{H}_m(v^n \mid x^n) \le \hat{H}_m(v^n) = H(P_{V^m})$ for every $v^n \in \mathcal{T}(P_{V^m})$.

Lower bound (for most v^n). By the method of types [18], the number of $v^n \in \mathcal{T}(P_{V^m})$ sharing a given joint type $\hat{P}_{X^mV^m}$ (with V^m-marginal P_{V^m}) is at most $2^{NH(\hat{Q}|\hat{P}_{X^m})} = 2^{N\hat{H}_m(v^n|x^n)}$, and the number of distinct joint types is at most $(N+1)^{|\mathcal{X}|^m|\mathcal{V}|^m}$. Hence the number of $v^n \in \mathcal{T}(P_{V^m})$ with $\hat{H}_m(v^n|x^n) < H(P_{V^m}) - \varepsilon$ is at most $(N+1)^{|\mathcal{X}|^m|\mathcal{V}|^m} \cdot 2^{N(H(P_{V^m})-\varepsilon)}$. Since $|\mathcal{T}(P_{V^m})| \ge 2^{NH(P_{V^m})}/(N+1)^{|\mathcal{V}|^m}$ by the method of types, taking the ratio gives (49). □

Remark 4.2 (Explicit capacity formula)**.** *Under symbol-wise invertibility (Definition 2.1) and with no input constraints:*

$$\lim_{m\to\infty}\lim_{n\to\infty} \hat{C}_m(\mathcal{N}_{\text{emp}}(H_v)) = \log\alpha - H_v, \tag{50}$$

achieved by any x^n with uniform empirical distribution ($\hat{H}_m(x^n) = \log\alpha$). This holds for all symbol-wise invertible channels, including group-additive channels $y = x \oplus v$.

Proof of Remark 4.2. By the relabeling lemma (37), $\hat{I}_m(x^n; y^n) = \hat{H}_m(y^n) - \hat{H}_m(v^n \mid x^n)$ for every x^n, v^n.

Lower bound (exact). Choose x^n uniform over $\mathcal{X}^m$-blocks, so $\hat{H}_m(x^n) = \log\alpha$. Symbol-wise invertibility and uniformity of x^n imply $\hat{H}_m(y^n) = \log\alpha$. For every $v^n \in \mathcal{N}_{\text{emp}}(H_v)$, $\hat{H}_m(v^n \mid x^n) \le \hat{H}_m(v^n) \le H_v$, so

$$\hat{I}_m(x^n; y^n) = \hat{H}_m(y^n) - \hat{H}_m(v^n \mid x^n) \ge \log\alpha - H_v,$$

and therefore $\hat{C}_m(\mathcal{N}_{\text{emp}}(H_v)) \ge \log\alpha - H_v$.

Upper bound (asymptotic). For any x^n and any type P_{V^m} with $H(P_{V^m}) = H_v$, Lemma 4.3 gives $\hat{H}_m(v^n \mid x^n) \ge H_v - \varepsilon$ for all but exponentially few $v^n \in \mathcal{T}(P_{V^m})$. Any such v^n lies in $\mathcal{N}_{\text{emp}}(H_v)$ (since $\hat{H}_m(v^n) = H_v$), and

$$\hat{I}_m(x^n; y^n) = \hat{H}_m(y^n) - \hat{H}_m(v^n \mid x^n) \le \log\alpha - H_v + \varepsilon.$$

Since x^n is arbitrary and $\varepsilon \to 0$ as $n \to \infty$ then $m \to \infty$, $\lim_{m\to\infty}\lim_{n\to\infty} \hat{C}_m(\mathcal{N}_{\text{emp}}(H_v)) \le \log\alpha - H_v$, which together with the lower bound gives (50). □

The expression $\log \alpha - H_v$ essentially agrees with the capacity formula of Shayevitz and Feder [11] and Lomnitz and Feder [12, 13] for group-additive channels. There is a slight difference that in their case, the limiting empirical entropy of the noise, H_v, is replaced by the LZ complexity of the noise, but the two measures are essentially equivalent for large m as will be discussed in the sequel (see also [17]).

Input constraints. When x^n is restricted to a set $\mathcal{X}_{\text{cst}}$ (e.g., a power constraint or a composition constraint), the capacity becomes

$$\hat{C}_m(\mathcal{N}_{\text{emp}}(H_v), \mathcal{X}_{\text{cst}}) \;=\; \max_{x^n \in \mathcal{X}_{\text{cst}}} \min_{v^n \in \mathcal{N}_{\text{emp}}(H_v)} \hat{I}_m(x^n; y^n). \tag{51}$$

By (37) this equals

$$\max_{x^n \in \mathcal{X}_{\text{cst}}} \min_{v^n \in \mathcal{N}_{\text{emp}}(H_v)} \Big[\hat{H}_m(y^n) - \hat{H}_m(v^n|x^n)\Big], \tag{52}$$

where $y_t = f(x_t, v_t)$ symbol-wise. Both terms depend on the noise v^n: the output entropy $\hat{H}_m(y^n)$ depends on how v^n interacts with x^n through the channel, while $\hat{H}_m(v^n|x^n)$ measures the residual noise entropy given the input. The adversary minimizes the difference by choosing v^n that simultaneously reduces the output entropy and concentrates the noise given the input. This is the individual-sequence analogue of the classical formula $C = \max_{P_X} I(X;Y) = \max_{P_X}[H(Y) - H(Y|X)]$, where the output entropy $H(Y)$ is jointly determined by the input distribution and the noise statistics. In particular, an input constraint that forces x^n away from uniform will in general reduce $\hat{H}_m(y^n)$ below $\log \alpha$, lowering the capacity below $\log \alpha - \rho_v$.

5 LZ Channel Capacity

The coding theorem of the preceding section is formulated in terms of the empirical entropies of m-blocks, where m is the block length of the converse's block-FSM decoder and s is its state count. These parameters determine the class of reference decoders: larger m and s yield a richer class and hence a tighter converse bound. However, m and s are not intrinsic parameters of the channel or the noise set $\mathcal{N}$, and it is desirable to pass to a characterization that is free of them without taking the explicit limits of m and s tending to infinity. In this context, prior works of Shayevitz and Feder [11] and Lomnitz and Feder [12, 13, 14] essentially suggest a language that is free of those parameters as their results are stated directly in terms of LZ complexities. This suggests that we attempt to reformulate our results in this LZ language as well. This will also allow a direct comparison with those works.

It turns out that the replacement of empirical entropies by LZ complexities require a considerable level of caution. The correct passage to LZ turns out to be selective. We were not able to replace both the conditional and the unconditional entropies by the respective LZ and conditional LZ complexities, but the resulting capacity $C_k(\mathcal{N})$ turns out to admit a *hybrid* form: essentially only the conditional entropy of the empirical mutual information can be replaced by conditional LZ complexity, whereas the unconditional entropy remains essentially intact (except the block length). This hybrid form — input entropy minus conditional LZ complexity — reflects exactly what the type-class random coding argument requires. In the further limit $k \to \infty$, both $H(P_{X^k})/k$ and $\bar{\rho}_k(x^n|y^n)$ converge to the same LZ complexities, giving the parameter-free expression $C_{\text{LZ}}(\mathcal{N})$.

5.1 LZ Complexity Background

5.1.1 LZ78 Complexity

The *LZ78 incremental parsing* [4] of x^n proceeds greedily: at each step, find the shortest extension p_j of a previously seen phrase $p'_j \in \{p_1, \ldots, p_{j-1}\}$ by one new symbol, such that $p_j \notin \{p_1, \ldots, p_{j-1}\}$. Let $c(x^n)$ denote the number of phrases. The *LZ78 complexity* is:

$$\rho_{\mathrm{LZ}}(x^n) \triangleq \frac{c(x^n) \log c(x^n)}{n}. \tag{53}$$

LZ78 is universally optimal over all finite-state compressors [4]: $\rho_{\mathrm{LZ}}(x^n) \leq \rho_s(x^n) + \epsilon(n, s)$ where $\epsilon(n, s) \to 0$ for fixed s, and $\rho_s(x^n)$ is the s-state compressibility of x^n.

5.1.2 Conditional LZ Complexity

Given side information y^n, the *conditional LZ complexity* $\rho_{\mathrm{LZ}}(x^n|y^n)$ is defined via the conditional incremental parsing of [5]. Concretely, apply the LZ78 incremental parsing algorithm to the sequence of pairs $(x_1, y_1), \ldots, (x_n, y_n)$ over $\mathcal{X} \times \mathcal{Y}$, obtaining $c(x^n, y^n)$ joint phrases. Denote by $c'(y^n)$ the number of phrases in the LZ78 parsing of y^n alone, and by $c_l(x^n|y^n)$ the number of x-components (sub-phrases) within the l-th y-phrase. Then:

$$\rho_{\mathrm{LZ}}(x^n|y^n) \triangleq \frac{1}{n} \sum_{l=1}^{c'(y^n)} c_l(x^n|y^n) \log c_l(x^n|y^n). \tag{54}$$

For fixed y^n, $x^n \mapsto \rho_{\mathrm{LZ}}(x^n|y^n)$ is approximately the normalized length of a UD code (up to a lower-order redundancy term [5]), a property used in the achievability proof.

5.1.3 Block-Average LZ Complexity

For a given positive integer k dividing n (the LZ-scale block length), we define the following quantities:

$$\bar{\rho}_k(x^n) \triangleq \frac{k}{n} \sum_{i=0}^{n/k-1} \rho_{\mathrm{LZ}}(x_{ik+1}^{ik+k}), \tag{55}$$

$$\bar{\rho}_k(x^n|y^n) \triangleq \frac{k}{n} \sum_{i=0}^{n/k-1} \rho_{\mathrm{LZ}}(x_{ik+1}^{ik+k}|y_{ik+1}^{ik+k}), \tag{56}$$

$$\bar{I}_k(x^n; y^n) \triangleq \bar{\rho}_k(x^n) - \bar{\rho}_k(x^n|y^n). \tag{57}$$

The *k-block resetting scheme* refers throughout the sequel to the application of the LZ78 algorithm independently in successive blocks, with the LZ dictionary reset at the start of each k-block.

5.1.4 Sandwiching Bounds

The following theorem gives the upper sandwiching bound that connects $\bar{\rho}_k$ to empirical entropy, used in the converse proof below.

Theorem 5.1 (Upper sandwiching bound [17])**.** *For every $x^n \in \mathcal{X}^n$, positive integer k dividing n, and $m \geq 1$:*

$$\bar{\rho}_k(x^n) \leq \hat{H}_m(x^n) + \frac{1}{m} + \frac{2(m+1)\alpha^{m+1}}{n} \log \frac{n}{m} + \epsilon_2(k, \alpha^{2m}), \tag{58}$$

where $\epsilon_2(k, s) \to 0$ as $k \to \infty$ for fixed s. An analogous bound holds for conditional quantities, with $\alpha\beta$ replacing α in the exponents.

We abbreviate (58) as:

$$\bar{\rho}_k(x^n) \leq \hat{H}_m(x^n) + \delta_+(k, m, n, \alpha), \quad \bar{\rho}_k(x^n|y^n) \leq \hat{H}_m(x^n|y^n) + \delta_+(k, m, n, \alpha\beta), \tag{59}$$

where $\delta_+(k, m, n, \cdot) \to 0$ as $n \gg k \gg m \gg 1$ with $\log n \gg m$. Here '$k \gg m$' means $k/m \to \infty$, which is needed for $\epsilon_2(k, \alpha^{2m}) \to 0$: since $\epsilon_2(k, s) = O((\log s)/k)$ (by the LZ redundancy bound [4]), we have $\epsilon_2(k, \alpha^{2m}) = O(m \log \alpha / k) \to 0$ whenever $k/m \to \infty$. The condition $\log n \gg m$ (equivalently $m = o(\log n)$) is needed for the term $2(m+1)\alpha^{m+1}/n \cdot \log(n/m) \to 0$.

The conditional bound (59) is the sole bridge used in the converse proof (Step 4 of Theorem 5.2 below): it converts the Fano-based bound on $\hat{H}_m(x^n|y^n)$ into a bound on $\bar{\rho}_k(x^n|y^n)$, under the hierarchy $n \gg k \gg m \gg 1$, $m = o(\log n)$. No sandwiching is needed for the achievability, which operates directly at the LZ level.

5.2 LZ Capacity: Definition and Converse

Definition 5.1 (Individual-sequence LZ channel capacity)**.** *Let $\mathcal{N} \subseteq \mathcal{V}^n$ be any noise set. For a channel (arbitrary deterministic mapping $x^n \mapsto y^n$, parameterized by $v^n \in \mathcal{N}$) and LZ-scale block length k (with $k|n$), the* block-average LZ channel capacity *is:*

$$C_k(\mathcal{N}) \triangleq \sup_{x^n \in \mathcal{X}^n} \min_{v^n \in \mathcal{N}} \left[\frac{H(P_{X^k})}{k} - \bar{\rho}_k(x^n|y^n) \right], \tag{60}$$

where P_{X^k} is the k-block empirical distribution of x^n, and $y^n = q(x^n, v^n)$ is the channel output (4). This is well-defined for any $\mathcal{N} \subseteq \mathcal{V}^n$: the inner expression is a real number for each pair (x^n, v^n), and the minimum over the finite set $\mathcal{N}$ and supremum over the finite set $\mathcal{X}^n$ are unambiguous.

This is the individual-sequence analogue of the compound channel capacity: maximize over input distributions (here, k-block types), minimize over noise sequences in $\mathcal{N}$. No cardinality condition on $\mathcal{N}$ is needed for the converse to hold. The choice of $\mathcal{N}$ is flexible; the converse (Theorem 5.2) applies to *any* $\mathcal{N} \subseteq \mathcal{V}^n$. The achievability (Theorem 5.3) additionally requires $\mathcal{N}$ to be closed under k-block permutations: for any $v^n \in \mathcal{N}$ and any k-block permutation π, the permuted sequence v^n_π must also belong to $\mathcal{N}$. This condition is satisfied by $\mathcal{N}_{\mathrm{LZ}}(\rho_v)$ (since $\bar{\rho}_k$ is permutation-invariant), by Hamming balls, and by any set defined by a permutation-invariant criterion. The classical LZ ball $\{v^n : \rho_{\mathrm{LZ}}(v^n) \leq \rho_v\}$ is *not* permutation-closed in general (a permutation may increase ρ_{LZ} by disrupting the global parsing structure), so achievability for that set remains an open problem.

Theorem 5.2 (Channel capacity converse in LZ terms)**.** *Let the channel be $y^n = q(x^n, v^n)$ for an arbitrary deterministic function q, with $v^n \in \mathcal{N}$ for some noise set $\mathcal{N} \subseteq \mathcal{V}^n$. Let the encoder be arbitrary (no finite-state constraint) and the decoder be a block-FSM with s states and block length m. Suppose a codebook of $M = 2^{nR}$ codewords achieves small $\hat{P}_e$ (average symbol-error frequency, as in (6)) for every $v^n \in \mathcal{N}$. Then for positive integers $k \gg m \gg 1$ (with $m|n$, $k|n$, and $\log n \gg m$):*

$$\phi(\hat{P}_e) \geq \bar{\rho}_k(x^n|y^n) - \Gamma_0, \tag{61}$$

and consequently:

$$R \leq C_k(\mathcal{N}) + \phi\left(\max_{v^n \in \mathcal{N}} \hat{P}_e(v^n)\right) + \Gamma_0, \tag{62}$$

where $C_k(\mathcal{N})$ is defined in Definition 5.1, and

$$\Gamma_0 = \frac{\log s}{m} + \delta_+(k, m, n, \alpha\beta), \tag{63}$$

with $\Gamma_0 \to 0$ as $k \gg m \gg 1$ with $\log n \gg m$ and s, α, β fixed.

Proof. **Step 1: Dominant type and rate bound.** By the same pigeonhole argument as in the proof of Corollary 4.1 (with m replaced by k and order-m types replaced by k-block types), there exists a k-block type P_{X^k} such that at least $2^{nR-o(n)}$ codewords share type P_{X^k}, and

$$R \leq \frac{H(P_{X^k})}{k} + o(1). \tag{64}$$

Fix this dominant type P_{X^k} and any codeword x^n of this type for the remainder of the proof.

Step 2: Worst noise and Fano. Let v_n^* be a maximizer of $\bar{\rho}_k(x^n|q(x^n, v^n))$ across $\mathcal{N}$ for the fixed codeword x^n of Step 1. Since ϕ is non-decreasing,

$$\phi\left(\max_{v^n \in \mathcal{N}} \hat{P}_e(v^n)\right) \geq \phi(\hat{P}_e(v_*^n)) \tag{65}$$

so it suffices to lower-bound $\phi(\hat{P}_e(v_*^n))$. In what follows, $y^n = q(x^n, v_*^n)$.

Step 3: Apply Theorem 3.2. Theorem 3.2 (with block length m, decoder state count s) gives:

$$\phi(\hat{P}_e(v_*^n)) \geq \hat{H}_m(x^n|y^n) - \frac{\log s}{m}. \tag{66}$$

Step 4: Convert empirical entropy to $\bar{\rho}_k$ via sandwiching. By the upper sandwiching bound (59) applied to $(x^n \mid y^n)$ (joint alphabet size $\alpha\beta$):

$$\bar{\rho}_k(x^n|y^n) \leq \hat{H}_m(x^n|y^n) + \delta_+(k, m, n, \alpha\beta), \tag{67}$$

so $\hat{H}_m(x^n|y^n) \geq \bar{\rho}_k(x^n|y^n) - \delta_+(k, m, n, \alpha\beta)$. Substituting into (66):

$$\phi(\hat{P}_e(v_*^n)) \geq \bar{\rho}_k(x^n|y^n) - \frac{\log s}{m} - \delta_+(k, m, n, \alpha\beta) = \bar{\rho}_k(x^n|y^n) - \Gamma_0, \tag{68}$$

where $\Gamma_0 = (\log s)/m + \delta_+(k, m, n, \alpha\beta) \to 0$ as $k \gg m \gg 1$ with $\log n \gg m$ and s fixed. Combined with (65), this gives (61).

Step 5: Upper-bound R using $C_k(\mathcal{N})$. Combining (64) and (68) with (65):

$$\begin{aligned} R &\le \frac{H(P_{X^k})}{k} + o(1) \\ &= \frac{H(P_{X^k})}{k} - \bar{\rho}_k(x^n|y^n) + \bar{\rho}_k(x^n|y^n) + o(1) \\ &\le \frac{H(P_{X^k})}{k} - \bar{\rho}_k(x^n|y^n) + \phi\left(\max_{v^n \in \mathcal{N}} \hat{P}_e(v^n)\right) + \Gamma_0 + o(1). \end{aligned} \tag{69}$$

By definition of $C_k(\mathcal{N})$ (Definition 5.1),

$$\begin{aligned} \frac{H(P_{X^k})}{k} - \bar{\rho}_k(x^n|y^n) &= \frac{H(P_{X^k})}{k} - \max_{v^n \in \mathcal{N}} \bar{\rho}_k(x^n|q(x^n, v^n)) \\ &= \min_{v^n \in \mathcal{N}} \left[\frac{H(P_{X^k})}{k} - \bar{\rho}_k(x^n|q(x^n, v^n))\right] \\ &\le \max_{\tilde{x}^n \in \mathcal{X}^n} \min_{v^n \in \mathcal{N}} \left[\frac{H(P_{\tilde{X}^k})}{k} - \bar{\rho}_k(\tilde{x}^n|q(\tilde{x}^n, v^n))\right] \\ &= C_k(\mathcal{N}). \end{aligned} \tag{70}$$

Absorbing the $o(1)$ into Γ_0 gives (62). □

Remark 5.1 (Operational interpretation via predictive coding)**.** *Theorem 5.2 admits a direct operational interpretation in terms of* predictive coding with a finite-state predictor*, offering a perspective complementary to the sandwich-based proof above.*

Consider any two-stage encoder for x^n with side information y^n:

1. Prediction: *Apply a block-FSM filter with s states to y^n to produce a prediction $\hat{x}^n$ of x^n.*
2. Residual coding: *Compress the error sequence $e^n = x^n - \hat{x}^n$ (well-defined w.l.o.g., since Hamming distance is a difference distortion measure) using a memoryless empirical entropy coder at rate $\hat{H}_1(e^n)$, plus a redundancy term that vanishes with n.*

Since $e_t = 0$ whenever $\hat{x}_t = x_t$, the symbol e_t takes value 0 with empirical frequency $1 - \hat{P}_e$. The empirical entropy $\hat{H}_1(e^n)$ is therefore at most $\phi(\hat{P}_e) = h_2(\hat{P}_e) + \hat{P}_e \log(\alpha - 1)$, the maximum entropy of any distribution on $\mathcal{X}$ with $P(e = 0) = 1 - \hat{P}_e$.

On the other hand, by [6], no finite-state encoder with side information y^n can compress x^n to a rate below $\bar{\rho}_k(x^n|y^n)$ minus a redundancy term (arising from finite-k and finite-n effects, of the same character as δ_+ in (59)*). Since the two-stage encoder is a particular finite-state encoder, its rate $\hat{H}_1(e^n)$ cannot be below this fundamental limit. Absorbing all redundancy terms into a single Γ_0 (of the same character as in* (63)*) then yields a bound of the same form as* (61)*, though the precise redundancy term may differ.*

Thus, Theorem 5.2 is the statement that no finite-state predictor followed by memoryless residual coding can achieve a rate below $\phi(\hat{P}_e)$ *— regardless of the predictor used.*

5.3 Achievability

We prove achievability for the channel model $y^n = q(x^n, v^n)$ when the encoder knows q and $\mathcal{N}$ but not the specific $v^n \in \mathcal{N}$. Throughout this subsection the noise set $\mathcal{N}$ is assumed to be *closed under k-block permutations*: for any $v^n \in \mathcal{N}$ and any permutation π of the n/k consecutive k-blocks of v^n, we have $v^n_\pi \in \mathcal{N}$. This is the structural condition needed for the random-coding argument; it holds for all noise sets listed in the examples following Definition 5.1. The main result uses capacity $C_k(\mathcal{N})$ (Definition 5.1).

Definition 5.2 (k-block resetting conditional LZ decoder)**.** *Fix a positive integer k dividing n. Given a codebook $\mathcal{C}_n \subset \mathcal{X}^n$ and channel output y^n, the* k-block resetting conditional LZ decoder *computes, for each $x^n \in \mathcal{C}_n$, the block-average conditional LZ complexity $\bar{\rho}_k(x^n|y^n)$ as defined in* (56)*, where each block's conditional LZ complexity is computed independently (resetting to a fresh empty dictionary at each block boundary), and outputs:*

$$\hat{x}^n \triangleq \arg\min_{x^n \in \mathcal{C}_n} \bar{\rho}_k(x^n \mid y^n), \tag{71}$$

with ties broken arbitrarily. ◊

Remark 5.2 (Type-independence of the inner minimum for permutation-closed $\mathcal{N}$)**.** *Since $\mathcal{N}$ is closed under k-block permutations (as assumed throughout this subsection), the inner minimum in* (60) *depends only on the k-block type of x^n, not on the specific representative. Indeed, if x^n and $\tilde{x}^n$ share the same k-block type P_{X^k}, then $\tilde{x}^n = x^n_\pi$ for some k-block permutation π, and the map $v^n \mapsto v^n_\pi$ is a bijection on $\mathcal{N}$, so*

$$\min_{v^n \in \mathcal{N}} \bar{\rho}_k(\tilde{x}^n \mid q(\tilde{x}^n, v^n)) = \min_{v^n \in \mathcal{N}} \bar{\rho}_k(x^n_\pi \mid q(x^n_\pi, v^n_\pi)).$$

Since $\bar{\rho}_k(\cdot|\cdot)$ is a block average, it is invariant under joint k-block permutation of both arguments: $\bar{\rho}_k(x^n_\pi \mid q(x^n_\pi, v^n_\pi)) = \bar{\rho}_k(x^n \mid q(x^n, v^n))$. Hence the minimum over $\mathcal{N}$ is the same for all representatives of a given k-block type, and the supremum in (60) *reduces equivalently to a supremum over k-block types.*

Definition 5.3 (Randomized codebook ensemble)**.** *Fix a channel q, noise set $\mathcal{N}$, LZ-scale block length k, and rate $R < C_k(\mathcal{N})$. Since the number of k-block types is finite, the supremum in* (60) *is attained; let $P^*_{X^k}$ be any maximizing type:*

$$\min_{v^n \in \mathcal{N}} \left[\frac{H(P^*_{X^k})}{k} - \bar{\rho}_k(x^n | q(x^n, v^n)) \right] = C_k(\mathcal{N}),$$

for any x^n with k-block empirical distribution $P^*_{X^k}$ (by Remark 5.2, the value does not depend on the choice of representative). The k-type class *is*

$$\mathcal{T}_k(P^*_{X^k}) \triangleq \left\{ x^n \in \mathcal{X}^n : P_{X^k}(x^n) = P^*_{X^k} \right\},$$

i.e., all sequences whose k-block empirical distribution equals $P^*_{X^k}$. Set $M \triangleq \lfloor 2^{nR} \rfloor$. A randomized code *consists of:*

- *A* message set $\mathcal{M}_n \triangleq \{1, 2, \ldots, M\}$.
- *An* ensemble of codebooks $\{\mathcal{C}_n(\omega)\}_\omega$*, each consisting of* M *codewords drawn independently and uniformly from* $\mathcal{T}_k(P^*_{X^k})$*, indexed by a shared random key* ω *drawn uniformly from a finite index set.*
- *A* decoder*: the* k*-block resetting conditional LZ decoder (Definition 5.2), applied to the codebook* $\mathcal{C}_n(\omega)$*.*

The key ω *is known to both encoder and decoder.* $\Diamond$

Lemma 5.1 (Kraft counting lemma)**.** *For any fixed* $y^n \in \mathcal{Y}^n$*, positive integer* k *dividing* n*, and any* $L \geq 0$*:*

$$\left|\left\{\tilde{x}^n \in \mathcal{X}^n : \bar{\rho}_k(\tilde{x}^n|y^n) \leq \frac{L}{n}\right\}\right| \leq 2^{L+o(n)}, \tag{72}$$

where $o(n)$ *depends only on* n*,* k*, and* $|\mathcal{X}|$ *(not on* y^n *or* $\tilde{x}^n$*).*

Proof. The k-block resetting conditional LZ scheme (Definition 5.2) encodes $\tilde{x}^n$ given y^n block by block, each block independently. For block i (indices $ik+1$ to $(i+1)k$), the conditional LZ algorithm [5] on k symbols produces a uniquely decodable code for $\tilde{x}_{ik+1}^{(i+1)k}$ given $y^n{}_{ik+1}^{(i+1)k}$ with codelength $\ell_i(\tilde{x}^n) \triangleq k\rho_{\mathrm{LZ}}(\tilde{x}_{ik+1}^{(i+1)k} \mid y^n{}_{ik+1}^{(i+1)k}) + O(c_i)$, where $c_i \leq k/\log k$ is the number of phrases and $O(c_i) = o(k)$. Concatenating the n/k independent block codes gives a UD code for $\tilde{x}^n$ given y^n with total codelength $\ell(\tilde{x}^n) = \sum_{i=0}^{n/k-1} \ell_i(\tilde{x}^n) = n\bar{\rho}_k(\tilde{x}^n|y^n) + o(n)$. By the Kraft inequality applied independently within each block (the blocks are decoded independently):

$$\sum_{\tilde{x}^n \in \mathcal{X}^n} 2^{-\ell(\tilde{x}^n)} \leq 1. \tag{73}$$

For $\tilde{x}^n$ with $\bar{\rho}_k(\tilde{x}^n|y^n) \leq L/n$: $\ell(\tilde{x}^n) \leq L + o(n)$. Restricting the Kraft sum:

$$\left|\left\{\tilde{x}^n : \bar{\rho}_k(\tilde{x}^n|y^n) \leq \tfrac{L}{n}\right\}\right| \cdot 2^{-(L+o(n))} \leq \sum_{\tilde{x}^n : \bar{\rho}_k(\tilde{x}^n|y^n) \leq L/n} 2^{-\ell(\tilde{x}^n)} \leq 1, \tag{74}$$

giving (72). $\square$

Theorem 5.3 (Achievability)**.** *Let the channel be symbol-wise invertible (Definition 2.1) and let* $\mathcal{N}$ *be a noise set closed under* k*-block permutations: for every* $v^n \in \mathcal{N}$ *and every permutation* π *of the* n/k *consecutive* k*-blocks of* v^n*, the permuted sequence* v^n_π *belongs to* $\mathcal{N}$*. The capacity* $C_k(\mathcal{N})$ *is defined in Definition 5.1. For any rate* $R < C_k(\mathcal{N})$ *and all sufficiently large* n*, there exists a deterministic code of at least* $2^{n(R-n^{-1})}$ *messages and block length* n *(with the* k*-block resetting conditional LZ decoder of Definition 5.2) such that the worst-case per-symbol error frequency satisfies*

$$\max_{i \in \mathcal{M}_n} \hat{P}_e(i) \;\to\; 0 \tag{75}$$

simultaneously for every $v^n \in \mathcal{N}$*, where* $\epsilon \triangleq C_k(\mathcal{N}) - R > 0$*. (The* n^{-1} *rate loss is due to expurgation; it is negligible for large* n*.)*

Proof of Theorem 5.3. Let $\epsilon \triangleq C_k(\mathcal{N}) - R > 0$.

Step 1: Capacity-achieving type and k-type class. Since the number of k-block types is finite, the supremum in $C_k(\mathcal{N})$ is attained; let $P^*_{X^k}$ be a maximizing type (Definition 5.3), so that for any $x^n \in \mathcal{T}_k(P^*_{X^k})$:

$$\bar{\rho}_k(x^n | \, q(x^n, v^n)) \leq \frac{H(P^*_{X^k})}{k} - C_k(\mathcal{N}) \qquad \forall \, v^n \in \mathcal{N}, \tag{76}$$

where we used Remark 5.2: since $\mathcal{N}$ is closed under k-block permutations, the inner minimum in $C_k(\mathcal{N})$ is the same for all sequences in $\mathcal{T}_k(P^*_{X^k})$. The k-type class $\mathcal{T}_k(P^*_{X^k})$ is defined in Definition 5.3. By the method of types applied to the alphabet $\mathcal{X}^k$ with block length n/k,

$$|\mathcal{T}_k(P^*_{X^k})| \;\geq\; 2^{(n/k)H(P^*_{X^k}) - o(n/k)} \;=\; 2^{nH(P^*_{X^k})/k - o(n)}. \tag{77}$$

Since $R < C_k(\mathcal{N}) \leq H(P^*_{X^k})/k$ (the inequality $C_k(\mathcal{N}) \leq H(P^*_{X^k})/k$ holds because $\bar{\rho}_k \geq 0$), we have

$$|\mathcal{T}_k(P^*_{X^k})| \geq 2^{nR} = M \tag{78}$$

for all sufficiently large n, so the ensemble is well-defined.

Step 2: Random codebook and decoding rule. Draw M codewords $X(1), \ldots, X(M)$ independently and uniformly at random from $\mathcal{T}_k(P^*_{X^k})$. The shared randomness ω is available to encoder and decoder. On receiving y^n, the k-block resetting conditional LZ decoder (Definition 5.2) outputs

$$\hat{i} \;=\; \arg\min_{i \in \mathcal{M}_n} \; \bar{\rho}_k(X(i) \mid y^n), \tag{79}$$

breaking ties arbitrarily.

Step 3: Correct codeword score. Fix $v^n \in \mathcal{N}$ and a transmitted codeword $x^n = X(i) \in \mathcal{T}_k(P^*_{X^k})$, with $y^n = q(x^n, v^n)$. We claim:

$$\bar{\rho}_k(x^n \mid y^n) \leq \frac{H(P^*_{X^k})}{k} - C_k(\mathcal{N}). \tag{80}$$

This follows directly from (76), which applies to x^n since $x^n \in \mathcal{T}_k(P^*_{X^k})$. Note that the threshold $H(P^*_{X^k})/k - C_k(\mathcal{N})$ is the same for *every* $x^n \in \mathcal{T}_k(P^*_{X^k})$ and every $v^n \in \mathcal{N}$.

Step 4: Confusable set. For fixed y^n, define the confusable set with threshold from (80):

$$\mathcal{G}(y^n) \triangleq \big\{\tilde{x}^n \in \mathcal{X}^n : \bar{\rho}_k(\tilde{x}^n \mid y^n) \leq \frac{H(P^*_{X^k})}{k} - C_k(\mathcal{N})\big\}. \tag{81}$$

By the Kraft counting Lemma 5.1:

$$|\mathcal{G}(y^n)| \leq 2^{n(H(P^*_{X^k})/k - C_k(\mathcal{N})) + o(n)}. \tag{82}$$

A wrong codeword $\widetilde{X}^n = X(j)$, $j \neq i$, drawn uniformly from $\mathcal{T}_k(P^*_{X^k})$ independently of y^n, causes an error only if $\widetilde{X}^n \in \mathcal{G}(y^n)$. Since $\widetilde{X}^n$ is independent of y^n:

$$\Pr_{\omega}\big[\widetilde{X}^n \in \mathcal{G}(y^n) \;\big|\; X^n = x^n\big] \leq \frac{|\mathcal{G}(y^n)|}{|\mathcal{T}_k(P^*_{X^k})|}. \tag{83}$$

Step 5: Union bound and average error. By the union bound over $M-1$ competing codewords, using (82), (83), and $|\mathcal{T}_k(P^*_{X^k})| \geq 2^{nH(P^*_{X^k})/k+o(n)}$ from (77):

$$\begin{aligned}\mathbb{E}_\omega\big[\hat{P}_e(i) \;\big|\; X(i)=x^n\big] &\leq 2^{nR}\cdot\frac{2^{n(H(P^*_{X^k})/k-C_k(\mathcal{N}))+o(n)}}{2^{nH(P^*_{X^k})/k+o(n)}}\\ &= 2^{n(R-C_k(\mathcal{N}))+o(n)} = 2^{-n\epsilon+o(n)}.\end{aligned} \tag{84}$$

Hence:

$$\mathbb{E}_\omega\big[\hat{P}_e(i) \;\big|\; X(i)=x^n\big] \leq 2^{-n\epsilon+o(n)} \qquad \forall\, x^n \in \mathcal{T}_k(P^*_{X^k}),\ \forall\, v^n \in \mathcal{N}. \tag{85}$$

Averaging over $X(i)$ and then over all messages:

$$\mathbb{E}_\omega\big[\bar{P}_e\big] \;\leq\; 2^{-n\epsilon+o(n)} \;\xrightarrow{n\to\infty}\; 0, \tag{86}$$

simultaneously for every $v^n \in \mathcal{N}$. By the probabilistic method, there exists a deterministic codebook $\mathcal{C}^{(0)}$ such that

$$\bar{P}_e\big(\mathcal{C}^{(0)}\big) \;\leq\; 2^{-n\epsilon/2} \tag{87}$$

simultaneously for every $v^n \in \mathcal{N}$.

Expurgation. Let $\mathcal{B} \triangleq \{i : \hat{P}_e(i,\mathcal{C}^{(0)}) > 2\cdot 2^{-n\epsilon/2}\}$ be the set of "bad" messages for this codebook. From (87), $|\mathcal{B}|\cdot 2\cdot 2^{-n\epsilon/2} \leq M\cdot 2^{-n\epsilon/2}$, so $|\mathcal{B}| \leq M/2$. Delete the at most $M/2$ bad codewords. The remaining sub-code $\mathcal{C}^* \subset \mathcal{C}^{(0)}$ has at least $M/2 = 2^{nR-1}$ messages (rate $R - n^{-1}$) and satisfies:

$$\hat{P}_e(i,\mathcal{C}^*) \;\leq\; 2\cdot 2^{-n\epsilon/2} \;=\; 2^{-n\epsilon/2+1} \;\to\; 0 \tag{88}$$

for every remaining message i and every $v^n \in \mathcal{N}$ simultaneously. Since $R - n^{-1} \to R$ and the bound vanishes, the theorem is proved. □

5.4 Application to Joint Source-Channel Coding

Suppose the encoder first compresses a source sequence $u^n \in \mathcal{U}^n$ into a channel input $x^n \in \mathcal{X}^n$ via a *block finite-state invertible* (block-FSI) encoder with state count s: a causal encoder with s states that maps m-blocks of u^n to m-blocks of x^n, and whose inverse (recovering u^n from x^n) is also a block-FSM with at most s states. The resulting (u^n, x^n, y^n) chain realizes joint source-channel coding (JSCC). The following lemma shows that the empirical entropy is nearly preserved by such an encoder.

Lemma 5.2 (Approximate empirical entropy equivalence)**.** *Let $u^n \mapsto x^n$ be a block-FSI encoder with encoder state count s and inverse-decoder state count s, and let $m|n$. Then for every $y^n \in \mathcal{Y}^n$:*

$$|\hat{H}_m(u^n) - \hat{H}_m(x^n)| \leq \frac{2\log s}{m}, \tag{89}$$

$$|\hat{H}_m(u^n|y^n) - \hat{H}_m(x^n|y^n)| \leq \frac{2\log s}{m}, \tag{90}$$

$$|\hat{I}_m(u^n;y^n) - \hat{I}_m(x^n;y^n)| \leq \frac{4\log s}{m}. \tag{91}$$

All bounds vanish as $m \gg \log s$.

Proof. **Proof of** (89)**.** Since the encoder state at the start of block i takes at most s values, $\hat{H}_m(x^n|u^n) \le (\log s)/m$ (given u^n, at most s choices for the encoder state determine x^n block by block). By the chain rule $\hat{H}_m(x^n) \le \hat{H}_m(u^n) + \hat{H}_m(x^n|u^n)$, so $\hat{H}_m(x^n) - \hat{H}_m(u^n) \le (\log s)/m$. By symmetry (the inverse decoder has s states, giving $\hat{H}_m(u^n|x^n) \le (\log s)/m$): $\hat{H}_m(u^n) - \hat{H}_m(x^n) \le (\log s)/m$. Combining gives (89).

Proof of (90)**.** Since the encoder is block-FSI, the encoder state at the start of each m-block is determined by u^n up to that point. Hence $\hat{H}_m(x^n|u^n, y^n) \le (\log s)/m$. Using the chain rule:

$$\hat{H}_m(x^n|y^n) \le \hat{H}_m(u^n|y^n) + \hat{H}_m(x^n|u^n, y^n) \le \hat{H}_m(u^n|y^n) + \frac{\log s}{m},$$

and by symmetry $\hat{H}_m(u^n|y^n) \le \hat{H}_m(x^n|y^n) + (\log s)/m$. Combining gives (90).

Proof of (91)**.** Follows from (89) and (90) by the triangle inequality applied to $\hat{I}_m = \hat{H}_m(\cdot) - \hat{H}_m(\cdot|\cdot)$. □

Combining Theorem 3.2 with Lemma 5.2 yields an individual-sequence Fano inequality stated directly in terms of the source.

Theorem 5.4 (Individual-sequence Fano inequality for JSCC)**.** *Under the conditions of Lemma 5.2 (block-FSI encoder $u^n \mapsto x^n$ with state count s, $m|n$), let y^n be the channel's response to x^n. Let $\hat{u}^n$ be the output of a block-FSM decoder with block length m and state count s applied directly to y^n, and define the source reconstruction error frequency*

$$\hat{P}_e^u \triangleq \frac{1}{n}\sum_{t=1}^{n} \mathbf{1}\{\hat{u}_t \ne u_t\}.$$

Then:

$$\begin{aligned}\phi\Big(\hat{P}_e^u\Big) &\ge \hat{H}_m(u^n|y^n) - \frac{\log s}{m} \\ &\ge \hat{H}_m(u^n) - \hat{I}_m(u^n; y^n) - \frac{\log s}{m}. \end{aligned} \tag{92}$$

In particular, if reliable source reconstruction is required for every $v^n \in \mathcal{N}$:

$$\hat{H}_m(u^n) \le \hat{C}_m(\mathcal{N}) + \phi\Big(\hat{P}_e^u\Big) + \frac{5\log s}{m}. \tag{93}$$

As $\hat{P}_e^u \to 0$ and $m \gg \log s$: $\hat{H}_m(u^n) \le \hat{C}_m(\mathcal{N})$.

Proof. Applying Theorem 3.2 to the pair (u^n, y^n) with the block-FSM decoder $y^n \mapsto \hat{u}^n$ (block length m, state count s) gives:

$$\phi\Big(\hat{P}_e^u\Big) \ge \hat{H}_m(u^n|y^n) - \frac{\log s}{m}.$$

Since $\hat{H}_m(u^n|y^n) = \hat{H}_m(u^n) - \hat{I}_m(u^n; y^n)$, this gives both inequalities in (92). Now, $\hat{I}_m(u^n; y^n) \le \hat{I}_m(x^n; y^n) + (4\log s)/m$ by Lemma 5.2 (91), and $\hat{I}_m(x^n; y^n) \le \hat{C}_m(\mathcal{N})$ by definition of capacity. Hence $\hat{I}_m(u^n; y^n) \le \hat{C}_m(\mathcal{N}) + (4\log s)/m$, and therefore:

$$\phi\Big(\hat{P}_e^u\Big) \ge \hat{H}_m(u^n) - \hat{C}_m(\mathcal{N}) - \frac{5\log s}{m}.$$

Rearranging gives (93). □

Remark 5.3 (LZ-level JSCC converse)**.** *An analogue of* (93) *at the LZ complexity level follows from the upper sandwiching bound (Theorem 5.1). Applying* (58) *to the source sequence* u^n *with order parameter* $m \ll k$*:*

$$\bar{\rho}_k(u^n) \le \hat{H}_m(u^n) + \delta_+(k, m, n, |\mathcal{U}|),$$

where $\delta_+ \to 0$ *as* $n \gg k \gg m \gg 1$*. Substituting the converse bound* (93)*:*

$$\bar{\rho}_k(u^n) \le \hat{C}_m(\mathcal{N}) + \phi\Big(\hat{P}_e^u\Big) + \frac{5 \log s}{m} + \delta_+.$$

The empirical capacity $\hat{C}_m(\mathcal{N})$ *itself satisfies* $\hat{C}_m(\mathcal{N}) \le C_k(\mathcal{N}) + \delta'_+$ *for a second (generally different) redundancy term* δ'_+*, obtained by applying the upper sandwich bound to each* $\hat{I}_m(x^n; y^n)$ *in the definition of* $\hat{C}_m$*. Combining:*

$$\bar{\rho}_k(u^n) \le C_k(\mathcal{N}) + \phi\Big(\hat{P}_e^u\Big) + \frac{5 \log s}{m} + \delta_+ + \delta'_+, \tag{94}$$

where both redundancy terms vanish in the regime $n \gg k \gg m \gg \log s$*. This is the individual-sequence LZ-level source-channel coding converse: reliable end-to-end transmission is impossible when* $\bar{\rho}_k(u^n) > C_k(\mathcal{N})$*.* ◊

Source-channel separation achievability. Theorem 5.4 (converse) and Theorem 5.3 (LZ achievability, established above) together yield an individual-sequence analogue of the source-channel separation theorem. The achievability direction by separation is as follows. Suppose $\bar{\rho}_k(u^n) < C_k(\mathcal{N}) - \varepsilon$ for some $\varepsilon > 0$. First, compress u^n losslessly using LZ78 [3] to obtain a message index $w \in \{1, \dots, M\}$ with $\log M \le n\bar{\rho}_k(u^n) + O(\log n) < n(C_k(\mathcal{N}) - \varepsilon/2)$ for large enough n. Then transmit w using the empirical-entropy random code of Theorem 4.2 at rate $R = (\log M)/n < C_k(\mathcal{N}) - \varepsilon/2$, which yields vanishing per-message error frequency for all $v^n \in \mathcal{N}$. Together with the LZ-level converse (94), this constitutes an individual-sequence source-channel separation theorem: reliable communication is possible if and only if the LZ complexity of the source does not exceed the LZ channel capacity.

Why separation holds, and when it is optimal. It is worth pausing to interpret the separation theorem correctly. In classical information theory, the separation principle is in a sense *bad news*: separate source and channel coding can *always* be carried out, but the separation theorem merely says that one cannot do *better* than separation — only in situations where the principle fails (e.g., when the source and channel noise are statistically dependent) can joint source-channel coding outperform separate coding. The individual-sequence setting makes this even more transparent: here the noise is adversarial, so the system must be designed against the worst-case $v^n \in \mathcal{N}$. Under worst-case noise, the decoupling between source and channel is complete — no joint design can exploit correlations between the source and the noise, since no such correlations are assumed. Consequently, one cannot expect performance better than that offered by separate source and channel coding, and the separation theorem is tight.

More precisely, $C_k(\mathcal{N})$ is a worst-case (minimax) quantity optimized against the hardest noise sequence in $\mathcal{N}$, and LZ78 produces a nearly uniform bit-string w from u^n, decoupling the source statistics from the channel; the channel code needs only to transmit w reliably, with no knowledge of the source structure.

For a fixed non-worst-case v^n, joint source-channel coding could in principle exploit the gap between the per-realization capacity and $C_k(\mathcal{N})$ — but only with knowledge of which v^n occurred. Since neither encoder nor decoder knows v^n, this improvement is unavailable in the adversarial setting, and separation incurs no cost.

Note that the absence of a Markov structure $U^n - X^n - Y^n$ in the individual-sequence setting does not by itself contradict separation. As an extreme example, consider a modulo-additive channel with $v^n = u^n$: transmitting $x^n = 0^n$ (rate zero) yields $y^n = u^n$, so the source is reconstructed perfectly despite $H = \log \alpha$ and $C = 0$. There is no contradiction because $v^n = u^n$ is presumably not the worst noise sequence in $\mathcal{N}$ (and may not even belong to $\mathcal{N}$); the separation theorem is a statement about the worst-case v^n, not about favourable individual instances.

6 Discussion

6.1 The Double Limit and Asymptotic Capacity

The capacity $C_k(\mathcal{N}_n)$ depends on both n and k. All results in this paper hold at fixed n and k; here we discuss how to pass to the double limit in two distinct settings: pure channel coding and joint source-channel coding. The two settings require different limit operations, for reasons explained below, but both yield the same threshold $C_{\mathrm{LZ}}(\mathcal{N})$ — which is consistent, since both are answers to the same underlying question: what is the maximum achievable rate?

Pure channel coding. In pure channel coding the rate R is a fixed number (not a sequence in n), and the converse (Theorem 5.2) gives, for every n and k:

$$R \leq C_k(\mathcal{N}_n) + \phi(\hat{P}_e) + \Gamma(n,k),$$

where $\Gamma(n,k) \to 0$ as $n \to \infty$ for fixed k. Since R does not depend on n, taking $\liminf_{n\to\infty}$ of the right-hand side is the natural (and most conservative) operation:

$$R \leq \liminf_{n\to\infty} C_k(\mathcal{N}_n) + \phi(\hat{P}_e).$$

There is no mismatch between subsequences here: R is fixed, and the bound holds at every n, so it holds in the $\liminf$. Taking $\liminf_{k\to\infty}$ then gives:

$$R \leq \liminf_{k\to\infty} \liminf_{n\to\infty} C_k(\mathcal{N}_n) + \phi(\hat{P}_e).$$

This motivates the pessimistic definition of asymptotic capacity:

$$C_{\mathrm{LZ}}(\mathcal{N}) \triangleq \liminf_{k\to\infty} \liminf_{n\to\infty} C_k(\mathcal{N}_n). \tag{95}$$

The asymptotic converse is then simply $R \leq C_{\mathrm{LZ}}(\mathcal{N}) + \phi(\hat{P}_e)$.

For the achievability: if $R < C_{\mathrm{LZ}}(\mathcal{N})$, then for all sufficiently large k and n, $R < C_k(\mathcal{N}_n) - \epsilon/2$ for some $\epsilon > 0$, and Theorem 5.3 gives $\mathbb{E}_\omega[\hat{P}_e] \to 0$. Together: reliable communication at rate R is possible if and only if $R \leq C_{\mathrm{LZ}}(\mathcal{N})$ (with the boundary case $R = C_{\mathrm{LZ}}(\mathcal{N})$ left open, as is standard in the LZ framework).

A necessary consistency check: the paper has two capacity expressions at finite n and k — the empirical-entropy capacity $\hat{C}_m(\mathcal{N}_n)$ (Definition 4.1) and the LZ capacity $C_k(\mathcal{N}_n)$ (Definition 5.1) — and they must agree in the double limit:

$$\liminf_{k\to\infty} \liminf_{n\to\infty} \hat{C}_m(\mathcal{N}_n) \;=\; \liminf_{k\to\infty} \liminf_{n\to\infty} C_k(\mathcal{N}_n) \;=\; C_{\mathrm{LZ}}(\mathcal{N}).$$

This is not a coincidence: both are operational thresholds for the same channel coding problem, so they *must* agree asymptotically. The sandwich relation between $\hat{C}_m(\mathcal{N}_n)$ and $C_k(\mathcal{N}_n)$ (via the upper and lower sandwiching bounds of Section 5) is what makes this precise at finite n and k, and guarantees agreement in the limit.

Joint source-channel coding. In JSCC, the source LZ complexity $\bar{\rho}_k(x^n)$ enters the picture alongside the channel capacity, and the two must be compared *at the same* n. Here the relevant object is the difference:

$$D_k(n) \stackrel{\Delta}{=} \bar{\rho}_k(x^n) - C_k(\mathcal{N}_n). \tag{96}$$

A natural but incorrect approach would be to compare $\limsup_n \bar{\rho}_k(x^n)$ with $\liminf_n C_k(\mathcal{N}_n)$ separately, paralleling the ZL78 definition of compressibility $\bar{\rho}(x) \stackrel{\Delta}{=} \limsup_{n\to\infty} \rho_{\mathrm{LZ}}(x^n)$. This is wrong because the two limits may be attained on *different* subsequences: at some n_0, $\bar{\rho}_k(x^{n_0})$ may be large while $C_k(\mathcal{N}_{n_0})$ is also large (a good channel length cancels a bad source length, giving a vacuous bound); at another n_1, $\bar{\rho}_k(x^{n_1})$ may be small while $C_k(\mathcal{N}_{n_1})$ is also small (a bad channel length cancels a good source length). Taking separate limits misses this coupling entirely.

The correct formulation limsups the *difference* $D_k(n)$ at the same n:

$$\limsup_{n\to\infty} D_k(n) \le \phi(\hat{P}_e), \tag{97}$$

obtained by taking $\limsup_n$ of the LZ-level JSCC converse (94) and noting that redundancy terms vanish. The achievability condition is $\limsup_n D_k(n) < 0$.

Taking $\limsup_{k\to\infty}$ (since whether $\lim_k \limsup_n D_k(n)$ exists is open), and denoting:

$$D \;\stackrel{\Delta}{=}\; \limsup_{k\to\infty} \limsup_{n\to\infty} D_k(n), \tag{98}$$

the asymptotic JSCC coding theorem is: $D \le \phi(\hat{P}_e)$ (converse) and $D < 0 \Rightarrow \mathbb{E}_\omega[\hat{P}_e] \to 0$ (achievability). For reliable communication ($\hat{P}_e \to 0$) this collapses to $D \le 0$, i.e.,

$$\limsup_{k\to\infty} \limsup_{n\to\infty} \big[\bar{\rho}_k(x^n) - C_k(\mathcal{N}_n)\big] \le 0,$$

which is the individual-sequence analogue of the source-channel separation condition $\bar{\rho}(x) \le C_{\mathrm{LZ}}(\mathcal{N})$.

Relation to finite-n results. All theorems in this paper hold for fixed n and k, giving per-sequence bounds without any limiting argument. The asymptotic formulations above are the natural summaries that emerge from the double limit.

6.2 Bridging the Gap Between the Converse and the Achievability

There is an apparent asymmetry between the converse and the achievability on the decoder side. The converse (Theorem 5.2) is stated for a block-FSM decoder with s states and block length m, which processes y^n sequentially in blocks of length m with bounded inter-block memory. The achievability decoder, by contrast, compares metrics computed over the *entire* sequence y^n: it selects the codeword minimizing $\bar{\rho}_k(\tilde{x}^n \mid y^n)$ globally, and in that sense requires observing all n output symbols before producing any decision. This global look at y^n is not a block-FSM operation at the n-symbol scale, regardless of the fact that the metric $\bar{\rho}_k$ is computed in k-blocks internally.

The gap between the two decoder classes is closed by passing to *super-blocks* of T independent transmissions of length n. Over a super-block of length nT, the achievability decoder operates independently on each of the T sub-blocks of length n (since the codebook and channel are both applied independently per block), so the global nT-symbol decoder decomposes into T independent n-symbol decisions. In the limit $T \to \infty$ (with n fixed), this is equivalent to an n-block FSM with $s = 1$ state (no inter-block memory), which falls within the converse's block-FSM class with block length $m = n$ and $s = 1$. The correction term then satisfies $(\log s)/m = 0$ identically, and $\Gamma_0 = \delta_+(k, m, n, \alpha\beta) \to 0$ as $n \gg k \gg m \gg 1$ (with $m = n$ being the "super-block" block length passed to the converse).

On the encoder side, the converse imposes no finite-state constraint; the achievability's lookup-table block encoder (which maps each message i to a full n-length codeword $X(i)$ drawn uniformly from the k-type class $\mathcal{T}_k(P^*_{X^k})$) is therefore compatible with the converse without any additional restriction.

In the double limit $n \to \infty$ followed by $k \to \infty$, the converse bound is tight: $\Gamma_0 \to 0$ and $C_k(\mathcal{N}_n) \to C_{\mathrm{LZ}}(\mathcal{N})$, the individual-sequence LZ capacity defined in (95) as

$$\liminf_{k\to\infty} \liminf_{n\to\infty} C_k(\mathcal{N}_n).$$

Whether the converse extends to arbitrary (unsynchronized) decoders that need not operate block-by-block is also an open problem.

6.3 Does Feedback Increase the Individual-Sequence LZ Capacity?

An important open problem is whether feedback from the decoder to the encoder can increase the individual-sequence LZ capacity $C_k(\mathcal{N})$. In the classical probabilistic setting, feedback does not increase capacity for memoryless channels, but it can for channels with memory [20]. In the individual-sequence setting, the noise sequence v^n introduces temporal correlations across channel uses, so feedback may strictly increase the individual-sequence LZ capacity. If so, the achievability schemes of Shayevitz, Feder, and Lomnitz [11, 12, 14], which use feedback, may be achieving a genuinely larger capacity than our converse (Theorem 5.2) bounds. Characterizing the individual-sequence LZ capacity with feedback is an open problem.

References

[1] R. M. Fano, *Transmission of Information*, MIT Press and Wiley, New York, 1961.

[2] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Trans. Inform. Theory*, vol. IT-23, pp. 337–343, May 1977.

[3] J. Ziv, "Coding theorems for individual sequences," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 405–412, July 1978.

[4] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 530–536, September 1978.

[5] J. Ziv, "Universal decoding for finite-state channels," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 453–460, July 1985.

[6] N. Merhav, "Universal detection of messages via finite-state channels," *IEEE Trans. Inform. Theory*, vol. 46, pp. 2242–2246, September 2000.

[7] T. Uyematsu and S. Kuzuoka, "Conditional Lempel-Ziv complexity and its application to source coding theorem with side information," *IEICE Trans. Fundamentals*, Vol. E86-A, no. 10, pp. 2615–2617, October 2003.

[8] J. Ziv, "Distortion-rate theory for individual sequences," *IEEE Trans. Inform. Theory*, vol. IT-26, no. 2, pp. 137–143, March 1980.

[9] N. Merhav, "On the data processing theorem in the semi-deterministic setting," *IEEE Trans. Inform. Theory*, vol. 60, no. 10, pp. 6032–6040, October 2014.

[10] N. Merhav, "Finite-state source-channel coding for individual source sequences with source side information at the decoder," *IEEE Trans. Inform. Theory*, vol. 68, no. 3, pp. 1532–1544, March 2022.

[11] O. Shayevitz and M. Feder, "Communication with feedback via the point-to-point channel," *IEEE Trans. Inform. Theory*, vol. 57, no. 10, pp. 6546–6565, October 2011.

[12] Y. Lomnitz and M. Feder, "Universal communication over modulo-additive channels with an individual noise sequence," *IEEE Trans. Inform. Theory*, vol. 60, no. 12, pp. 7473–7489, December 2014.

[13] Y. Lomnitz and M. Feder, "Communication over individual channels," *IEEE Trans. Inform. Theory*, vol. 57, no. 11, pp. 7333–7358, November 2011.

[14] Y. Lomnitz and M. Feder, "Universal communication over arbitrarily varying channels," *IEEE Trans. Inform. Theory*, vol. 59, no. 6, pp. 3720–3752, June 2013.

[15] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inform. Theory*, vol. 40, no. 4, pp. 1147–1157, July 1994.

[16] K. Eswaran, A. D. Sarwate, A. Sahai, and M. Gastpar, "Zero-rate feedback can achieve the empirical capacity," *IEEE Trans. Inform. Theory*, vol. 56, no. 1, pp. 25–39, January 2010.

[17] N. Merhav, "Lempel–Ziv complexity, empirical entropies, and chain rules," *IEEE Trans. Inform. Theory*, vol. 72, no. 5, pp. 3486–3496, May 2026.

[18] I. Csiszár, "The method of types," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2505–2523, October 1998.

[19] D. Blackwell, L. Breiman, and A. J. Thomasian, "The capacities of certain channel classes under random coding," *Ann. Math. Statist.*, vol. 31, no. 3, pp. 558–567, September 1960.

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

[21] R. Ahlswede, "Elimination of correlation in random codes for arbitrarily varying channels," *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, vol. 44, pp. 159–175, 1978.

[22] I. Csiszár and P. Narayan, "The capacity of the arbitrarily varying channel revisited: positivity, constraints," *IEEE Trans. Inform. Theory*, vol. 34, no. 2, pp. 181–193, March 1988.

[23] I. Csiszár and P. Narayan, "Capacity of the Gaussian arbitrarily varying channel," *IEEE Trans. Inform. Theory*, vol. 37, no. 1, pp. 18–26, January 1991.

[24] B. Hughes and P. Thomas, "The capacity of a vector Gaussian arbitrarily varying channel," *IEEE Trans. Inform. Theory*, vol. 42, no. 2, pp. 461–474, March 1996.

[25] M. Langberg, "Private codes or succinct random codes that are (almost) perfect," in *Proc. 45th Annu. IEEE Symp. Foundations of Computer Science (FOCS)*, Rome, Italy, Oct. 2004, pp. 325–334.