Universal Random Coding for Successive Refinement of Individual Sequences Based on Lempel–Ziv Complexity

Neri Merhav

The Viterbi Faculty of Electrical and Computer Engineering
Technion - Israel Institute of Technology
Technion City, Haifa 3200003, ISRAEL
E–mail: merhav@technion.ac.il

Abstract

This work extends an earlier proposed universal random-coding ensemble for sample-wise lossy compression of individual sequences to the two-stage (successive-refinement) setting — an extension that is not quite straightforward, as explained in the sequel. The construction has three layers. First, a complete, self-contained direct/converse match in the language of type classes and typical sequences, a two-stage analogue of Rimoldi's classical characterization (applied in the superalphabet of blocks). Second, a realization of the same rates via a random-coding scheme built from the Lempel–Ziv (LZ78) complexity rather than the type-class uniform measure the first layer relies on, since an LZ-based ensemble is universal across source statistics and distortion measures, carries no block-length parameter in the codebook itself, and is directly implementable. Third, we give a fully type-free achievability construction (no joint types of ℓ-vectors appear anywhere in its search criterion) and show it matches the same converse exactly, reaching every point of the achievable region: the same converse from the first layer already binds any code, type-free schemes included, so no separate argument is needed to certify the match. We also describe (without proof) how to extend the type-free construction to any fixed number $r > 2$ of stages. Finally, we show this paper's achievable region is a superset of the one obtained by a companion paper, which is based on a finite-state modeling approach to a similar two-stage problem, extending a known single-stage domination result to two stages.

1 Introduction

This paper is about compressing an arbitrary individual (deterministic) sequence of data progressively, in two stages: a first, coarse description that satisfies a prescribed distortion level D_1, followed by a second description that adds refinement, held to its own prescribed level D_2 — possibly under a different distortion measure than the first stage's, as detailed in Section 2. This kind of progressive compression is natural whenever a receiver may need to stop listening relatively early (imagine a

picture that is first shown in low resolution and then progressively sharpened) and the question this paper addresses is how to do it well, and how to know when a given scheme is already as good as possible.

A recent paper [1] addressed the single-stage version of this problem: how to compress a single fixed sequence, in one shot, as well as possible, without knowing anything about its statistics in advance. The scheme it proposed is based on universal random coding: each candidate reproduction codeword $\hat{\boldsymbol{x}}$ is drawn independently at random, using a universal probability distribution proportional to $2^{-LZ(\hat{\boldsymbol{x}})}$, where $LZ(\cdot)$ is the Lempel–Ziv (LZ78) codelength [2] and the index of the first candidate accurate enough is transmitted; it was shown that no coding scheme can do substantially better than this, for the given sequence. The aim of the present paper is to extend the same random-coding approach to progressive compression with two stages, and to show that the resulting scheme is, again, essentially as good as any two-stage scheme could be; this two-stage case is this paper's actual content, developed in three stages throughout: first entirely in the language of type classes (Section 3), second — via a random-coding scheme built from LZ78 complexity (Section 4), and third — separately, via a fully type-free construction matching the same converse (Sections 4.2–4.3). Finally, Section 5 gives a description (without a full proof) of how the type-free mechanism extends to a general fixed number $r > 2$ of stages. A companion paper [3] studies a similar two-stage compression problem of individual sequences, but via a different paradigm: a finite-state lossless encoder fed by a vector quantizer. The present paper stays within the random-coding approach of [1] instead, and we show (Section 1.3) that this paper's achievable region is a superset of [3]'s, extending [1]'s own single-stage domination of any finite-state scheme to two stages.

Extending the lossy compression scheme of [1] into two stages might sound routine once the single-stage version is settled, but it is not — not because successive refinement itself harbors any new subtlety (it does not: both difficulties below already appear in Rimoldi's own, classical characterization), but because transplanting them to the individual-sequence, universal setting [1]'s own framework operates in takes real work — work that a further ambition, avoiding any commitment to a type in the search criterion, makes harder still.

The first is on the achievability side: the coarse description $\hat{\boldsymbol{x}}_1$ must also serve as good side information for the second stage, not merely satisfy the D_1 constraint. Rimoldi's own fix — commit

Stage 1's search to the needed joint behavior, not mere distortion-feasibility — transplants with new vocabulary: type in place of law, the universal LZ measure in place of a type-class-uniform one (Section 4.1). A fully type-free construction, though, cannot commit to anything in advance, and needs a different device: a criterion built from the actual rate a candidate induces, not from any type it belongs to (Section 4.2).

The second, deeper and less obvious, is on the converse side. At a single stage, there is one achievable rate and one converse binding every code at once; at two stages, the region is a union over many pieces, one per type of joint behavior — already Rimoldi's own structure. There, concentration ties a successful code to a single auxiliary law, so a converse proved at that law already binds it; with an individual sequence and an arbitrary code, no such concentration exists, so a converse proved at just one type leaves a real gap. Section 3 closes it by checking every type at once — needed only once type-freedom is on the table, since a code committed to one type (Theorems 3 and 4) is already bound by that single check.

1.1 Related work

This work lies at the intersection of three lines of research, and draws throughout on the classical theory of rate-distortion coding [4, 5, 6].

Universal lossy compression (probabilistic sources). A long line of work bounds the redundancy achievable when compressing a random source of unknown statistics, from the $\Theta(\log n/n)$-order results of [7, 8] and the ergodic-source treatment of [9], to [10]'s CLT for the redundancy and, with [11], a characterization of optimal compression via the negative log-probability of a distortion sphere — the probabilistic-setting analogue of this paper's own sphere-probability quantity, and the one connection from this line drawn on directly below. Mahmood and Wagner [12, 13] treat codes universal in both source and distortion measure.

The individual-sequence approach. Pioneered by Ziv, assuming no source statistics at all and imposing finite-state limitations on the encoder/decoder instead [14, 15, 2], later extended to side information [16]. Article [17] gives the implementation this paper relies on (Section 4): realizing a random-coding ensemble via an LZ decompressor fed by random bits. Article [18] studies the same kind of device for a different purpose — optimally guessing a secret sequence via random bits

fed into an LZ78 decoder — and shows the guessing effort needed is governed by the finite-state compressibility of the given sequence. Article [1], which serves as the single-stage basis for this paper, combines this philosophy with [11]'s characterization, proposing the random coding distribution $U(\hat{\boldsymbol{x}}) \propto 2^{-LZ(\hat{\boldsymbol{x}})}$ specifically, with a matching converse (for the vast majority of codewords within a type) and pointwise achievability.

Successive refinement. Equitz and Cover [20] proved, for memoryless sources, that refining D_1 into $D_2 \leq D_1$ achieves both rate-distortion optima simultaneously iff the optimal reproductions satisfy $X - \hat{X}_2 - \hat{X}_1$ — necessary *and* sufficient for the specific pair, via a chain rule built on El Gamal and Cover's [21] own multiple-description achievable region; Koshelev [22] and Kanlis and Narayan [24] obtained related divisibility and error-exponent results. Rimoldi [23] characterized the full region as a union over auxiliary distributions, the template Section 3 follows. [3], this paper's direct predecessor, is discussed in detail in Section 1.3.

1.2 Why Lempel–Ziv complexity

As mentioned earlier, article [1] already establishes, for the single-stage problem, that an LZ78-based random-coding ensemble has several properties no type-restricted construction can match. Section 3 builds a purely type-based achievability/converse pair for the two-stage problem — a direct translation of [23]'s own classical achievable-region characterization to individual sequences, included only to supply a matching converse, not as a contribution in its own right; this paper calls it the type-covering skeleton throughout. This paper's question is whether an LZ78-based construction can match that skeleton's rates exactly while also carrying the same several properties into the two-stage setting.

1. *Universality across the entire problem specification.* A scheme built on the random-coding distribution U_{Q^ℓ}, which is the uniform distribution across a certain type class Q^ℓ — an empirical distribution over blocks of length ℓ — requires commitment to a fixed Q^ℓ *before* the codebook is built, and Q^ℓ in turn must depend on everything the problem specifies: the type of the source sequence, both distortion measures, and both budgets. Also, Stage 2 needs its own, analogous conditional-type object. All this is completely avoided by the use of the universal LZ random-coding distribution and its conditional counterpart for the second stage, which are

universally asymptotically optimal across the type of the source sequences, the two distortion measures and the two distortion levels.

2. *No commitment to a block length in the codebook itself.* The choice of U_{Q^ℓ} requires also commitment to a certain choice of ℓ, and so, changing ℓ amounts to regenerating the codebook. By contrast, the LZ universal measure references no ℓ at all, so the same codebook serves every choice, and adjusting ℓ costs nothing beyond changing what is searched for.

3. *Actual implementability.* The type-class uniform distribution U_{Q^ℓ} is a proof device: covering-codebook existence is established probabilistically, with no suggestion of how to efficiently draw from it. On the other hand, the LZ78 algorithm is real and widely deployed; its ensemble is realized concretely by feeding shared random bits into an LZ78 decompressor [17] (Section 4's explicit protocol).

Section 4 shows all three properties survive the extension to two stages, with the same rate Section 3's skeleton achieves (Theorem 3) matched exactly — none of them cost a weaker guarantee. This buys one further thing beyond the three items above. Section 3's own construction needs a different codebook for every point of Section 3.4's frontier; Section 4.2's type-free construction, built from the same pair $(U, V(\cdot|\hat{\boldsymbol{x}}_1))$ but referencing no target type at all, lets a single, fixed ensemble realize every point of the achievable region simultaneously, with which point obtained selected purely by the search — though this search criterion pays a real price for the convenience, checked against the full remaining problem for each candidate tried, unlike the type-committed route's simpler, single-candidate check (Section 4.3's own closing discussion).

None of this is automatic. Progressive decodability — the first $L_1(\boldsymbol{x})$ bits alone (Stage 1's own codelength) must already give a valid D_1-reconstruction, before the next $L_2(\boldsymbol{x})$ bits (Stage 2's incremental codelength) exist — means $\hat{\boldsymbol{x}}_1$ must also serve as good side information for Stage 2, not merely satisfy D_1 on its own. Section 3 explains why, in a classical argument tracing back to Rimoldi's own characterization; Section 4.1 confirms the same requirement survives, unchanged, once Stage 1 draws from the unrestricted U instead.

1.3 Relation to the finite-state-encoder approach of [3]

Article [3] solves the identical two-stage problem, but differs from this paper in both modeling and achievability. Its encoder model comprises a general vector quantizer (VQ) followed by a finite-state lossless encoder operating on the reconstruction, with convergence only after a double limit ($n \to \infty$, then $q \to \infty$, q being the number of states); once that limit is taken, however, the resulting guarantee holds for each individual sequence, not merely on average over some ensemble. This paper's encoder, on the other hand, carries no finite-state restriction at all. This trade is already known to be favorable at the single-stage level: Article [1] already shows, even there, that the unrestricted route is never worse — and in general, strictly better — than the VQ-plus-finite-state-encoder paradigm, which in turn yields the minimum LZ compression ratio across the D-ball around the given source sequence. The intuitive reason is that the finite-state compressor and the LZ78 compressor alike act on the VQ's own output, yet neither is built to exploit the fact that this output — the actual input to the lossless encoder — is confined to a sparse subset of $\hat{\mathcal{X}}^n$, since not every sequence there can arise as VQ output. This sparsity goes unexploited, creating considerable room for improvement, exploited in [1]. Likewise, the achievable rate region of [3] is subsumed by that of the present paper, as the converse of [3] is also framed in the encoder model of a VQ followed by q-state encoders.

1.4 Outline

Section 2 establishes the notation and the definitions needed. Section 3 builds a complete direct/converse match in the language of types. Section 4 provides the LZ-based random coding scheme. Section 4.2 then gives a fully type-free construction. Section 4.3 shows it matches a converse exactly. Section 4.4 then uses this same type-free construction to show this paper's achievable region is a superset of [3]. Section 5 extends the type-free construction to a general fixed number of stages. Finally, Section 6 summarizes.

2 Formulation and Notation Conventions

Let n be a given positive integer and ℓ divide n. Define $m = \frac{n}{\ell}$. Let $\boldsymbol{z} \in \mathcal{A}^n$ denote a generic sequence over an alphabet $\mathcal{A}$. Partition $\boldsymbol{z}$ into m non-overlapping ℓ-blocks $z_1, \ldots, z_m \in \mathcal{A}^\ell$. Its

block type is defined as the empirical distribution of these m blocks over $\mathcal{A}^\ell$, denoted $\hat{P}_{\boldsymbol{z}}$, i.e.,

$$\hat{P}_{\boldsymbol{z}}(a) \stackrel{\Delta}{=} \frac{1}{m}|\{i \in \{1,\ldots,m\} : z_i = a\}|, \quad a \in \mathcal{A}^\ell. \tag{1}$$

A *block permutation* π acts on $\boldsymbol{z}$ by permuting its m blocks as whole units; obviously, block permutations preserve the block type, i.e. $\hat{P}_{\pi(\boldsymbol{z})} = \hat{P}_{\boldsymbol{z}}$. Likewise, for a fixed number k of length-n sequences $\boldsymbol{z}^{(1)},\ldots,\boldsymbol{z}^{(k)}$, over alphabets $\mathcal{A}_1,\ldots,\mathcal{A}_k$ respectively, write $\boldsymbol{z}^{(1)} \odot \cdots \odot \boldsymbol{z}^{(k)}$ for their *product sequence*, namely, the length-n sequence over the product alphabet $\mathcal{A}_1 \times \cdots \times \mathcal{A}_k$ whose i-th symbol is $(z_i^{(1)},\ldots,z_i^{(k)})$. Their *joint (block) type* is denoted

$$\hat{P}_{\boldsymbol{z}^{(1)}\ldots\boldsymbol{z}^{(k)}} \stackrel{\Delta}{=} \hat{P}_{\boldsymbol{z}^{(1)}\odot\cdots\odot\boldsymbol{z}^{(k)}}. \tag{2}$$

Clearly, a common block permutation applied to all $\boldsymbol{z}^{(i)}$, $i = 1,\ldots k$, simultaneously, preserves this joint type. Denote the fixed source sequence by $\boldsymbol{x}$, with block type $P^\ell \stackrel{\Delta}{=} \hat{P}_{\boldsymbol{x}}$ and finite alphabet $\mathcal{X}$ of cardinality $K = |\mathcal{X}|$. The reproduction alphabets of the two reconstruction stages are denoted by $\hat{\mathcal{X}}_1$ and $\hat{\mathcal{X}}_2$ with $\hat{K}_1 = |\hat{\mathcal{X}}_1|$, $\hat{K}_2 = |\hat{\mathcal{X}}_2|$; where a single, generic $\hat{\mathcal{X}}$ or $\hat{K}$ appears below (Lemma 2, stated once for either stage), it stands for whichever of $\hat{\mathcal{X}}_1, \hat{\mathcal{X}}_2$ (and $\hat{K}_1, \hat{K}_2$) is being instantiated, exactly as with d, D below.

Let $\hat{\boldsymbol{x}}_1 \in \hat{\mathcal{X}}_1^n$ and $\hat{\boldsymbol{x}}_2 \in \hat{\mathcal{X}}_2^n$ denote the reconstruction vectors of the first stage and second stage, respectively. The distortion functions of the first and the second stage, d_1 and d_2, both depend on $\boldsymbol{x}$ and the corresponding reconstruction ($\hat{\boldsymbol{x}}_1$ and $\hat{\boldsymbol{x}}_2$, respectively) only via their first-order (single-symbol) joint empirical distribution, and, for fixed such distribution, they grow linearly in n, which is naturally the case with additive distortion functions.

Define also $\mathcal{S}(\boldsymbol{x}, D_1) = \{\hat{\boldsymbol{x}}_1 : d_1(\boldsymbol{x}, \hat{\boldsymbol{x}}_1) \le nD_1\}$, $\hat{\mathcal{S}}(\hat{\boldsymbol{x}}_1, D_1) = \{\boldsymbol{x} : d_1(\boldsymbol{x}, \hat{\boldsymbol{x}}_1) \le nD_1\}$, and analogously $\mathcal{S}(\boldsymbol{x}, D_2), \hat{\mathcal{S}}(\hat{\boldsymbol{x}}_2, D_2)$ with d_2 in place of d_1; $\mathcal{S}_2(\boldsymbol{x}, D_1, D_2) = \mathcal{S}(\boldsymbol{x}, D_1) \times \mathcal{S}(\boldsymbol{x}, D_2)$, $\hat{\mathcal{S}}_2(\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2, D_1, D_2) = \{\boldsymbol{x} : (\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2) \in \mathcal{S}_2(\boldsymbol{x}, D_1, D_2)\}$. For a joint type F, define $\mathcal{T}(F|\boldsymbol{x}) \stackrel{\Delta}{=} \{\hat{\boldsymbol{x}} : \hat{P}_{\boldsymbol{x}\hat{\boldsymbol{x}}} = F\}$, along with an extension in the same way to more than one conditioning sequence, $\mathcal{T}(F|\boldsymbol{x}, \hat{\boldsymbol{x}}_1) \stackrel{\Delta}{=} \{\hat{\boldsymbol{x}}_2 : \hat{P}_{\boldsymbol{x}\hat{\boldsymbol{x}}_1\hat{\boldsymbol{x}}_2} = F\}$ and so on. Wherever a single, generic d or D appears below (Lemma 2, stated once for either stage), it stands for whichever of d_1, d_2 (and D_1, D_2) is being instantiated.

A *two-stage code* $\Phi = (\phi_1, \phi_2, \psi_1, \psi_2)$ consists of two encoders $\phi_1, \phi_2 : \mathcal{X}^n \to \{0,1\}^*$, with images $\mathcal{B}_1 \stackrel{\Delta}{=} \phi_1(\mathcal{X}^n)$ and $\mathcal{B}_2 \stackrel{\Delta}{=} \phi_2(\mathcal{X}^n)$ — in general distinct subsets of $\{0,1\}^*$, the set of finite

variable-length binary strings — and two decoders $\psi_1 : \mathcal{B}_1 \to \hat{\mathcal{X}}_1^n$, $\psi_2 : \mathcal{B}_1 \times \mathcal{B}_2 \to \hat{\mathcal{X}}_2^n$; the encoder output on $\boldsymbol{x}$ is $(\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2) \triangleq (\psi_1(\phi_1(\boldsymbol{x})), \psi_2(\phi_1(\boldsymbol{x}), \phi_2(\boldsymbol{x})))$, with codeword lengths $L_1(\boldsymbol{x}) \triangleq |\phi_1(\boldsymbol{x})|$, $L_2(\boldsymbol{x}) \triangleq |\phi_2(\boldsymbol{x})|$ and data-dependent rates $\rho_1(\boldsymbol{x}) \triangleq L_1(\boldsymbol{x})/n$, $\rho_2(\boldsymbol{x}) \triangleq L_2(\boldsymbol{x})/n$. The notations R_1 and R_2 are reserved for the coordinates of the rate-pair plane. The code Φ *meets both distortion constraints* if $d_1(\boldsymbol{x}, \hat{\boldsymbol{x}}_1) \le nD_1$ and $d_2(\boldsymbol{x}, \hat{\boldsymbol{x}}_2) \le nD_2$ for every $\boldsymbol{x}$; it gives a *one-to-one correspondence between codewords and binary strings at each stage* if ψ_1 is injective on $\mathcal{B}_1$ and $(u, v) \mapsto (\psi_1(u), \psi_2(u, v))$ is injective on $\{(\phi_1(\boldsymbol{x}), \phi_2(\boldsymbol{x})) : \boldsymbol{x} \in \mathcal{X}^n\}$ — distinct reconstructions get distinct binary representations, at each stage and jointly, with neither prefix-freedom nor unique decodability required (all Theorem 1's own counting argument needs). Both conditions are assumed throughout.

For a type Q^ℓ on $\hat{\mathcal{X}}^\ell$, $\mathcal{T}_n(Q^\ell) \triangleq \{\hat{\boldsymbol{x}} \in \hat{\mathcal{X}}^n : \hat{P}_{\hat{\boldsymbol{x}}} = Q^\ell\} \subset \hat{\mathcal{X}}^n$ is its type class. Let $U_{Q^\ell}(\cdot)$ denote the uniform distribution on it,

$$U_{Q^\ell}(\hat{\boldsymbol{x}}) = \begin{cases} \frac{1}{|\mathcal{T}_n(Q^\ell)|} & \hat{\boldsymbol{x}} \in \mathcal{T}_n(Q^\ell) \\ 0 & \text{elsewhere,} \end{cases} \tag{3}$$

Notation of marginals and conditionals of W^ℓ. Let X, $\hat{X}_1$ and $\hat{X}_2$ denote auxiliary random variables, taking values on $\mathcal{X}^\ell$, $\hat{\mathcal{X}}_1^\ell$, and $\hat{\mathcal{X}}_2^\ell$, respectively, jointly distributed according to a given joint type of ℓ-vectors $\hat{P}_{\boldsymbol{x}\hat{\boldsymbol{x}}_1\hat{\boldsymbol{x}}_2} = W$. For any subset of $\{X, \hat{X}_1, \hat{X}_2\}$, possibly with a further subset as conditioning, W^ℓ subscripted by that subset — with a bar for conditioning, following the customary notation rules of probability theory, denotes the corresponding marginal and/or conditional type induced by W^ℓ, e.g. W_X^ℓ is W^ℓ's own X-marginal, $W_{X\hat{X}_1}^\ell$ its own $(X, \hat{X}_1)$ sub-type, $W_{\hat{X}_2|\hat{X}_1}^\ell$ its own conditional type of $\hat{X}_2$ given $\hat{X}_1$, etc. For the sake of notational simplicity, the few such objects used repeatedly throughout also carry a short alias, defined here once and used interchangeably with the subscripted form thereafter:

$$P^\ell \triangleq W_X^\ell, \qquad Q_1^\ell \triangleq W_{\hat{X}_1}^\ell, \qquad Q_2^\ell \triangleq W_{\hat{X}_2}^\ell, \qquad W_1^\ell \triangleq W_{X,\hat{X}_1}^\ell, \qquad W_2^\ell \triangleq W_{X,\hat{X}_2}^\ell. \tag{4}$$

Objects used only rarely are left in the fully subscripted form.

For the type Q_1^ℓ, treated as a probability distribution on its own alphabet, $\hat{H}_{Q_1^\ell}(\hat{X}_1) \triangleq -\sum_a Q_1^\ell(a) \log Q_1^\ell(a)$ is its (empirical) entropy; for a joint type W_1^ℓ, $\hat{H}_{W_1^\ell}(\hat{X}_1|X)$, $\hat{I}_{W_1^\ell}(X; \hat{X}_1) = \hat{H}_{Q_1^\ell}(\hat{X}_1) - \hat{H}_{W_1^\ell}(\hat{X}_1|X)$ denote the corresponding conditional entropy and mutual information, and

so on. We will also use alternative notations like $\hat{H}(Q_1^\ell)$ and $\hat{H}(Q_2^\ell)$ for $\hat{H}_{Q_1^\ell}(\hat{X}_1)$ and $\hat{H}_{Q_2^\ell}(\hat{X}_2)$ whenever convenient and safe from any risk of ambiguity.

2.1 A general random-coding tool

One achievability mechanism is used throughout this paper: search a large randomly and independently selected codebook, drawn from some fixed random coding distribution, for the first candidate meeting a certain target set — and the resulting codelength is governed entirely by the target's probability under that measure. Both pieces vary from one use to the next, but the underlying principle remains the same. The following lemma is proved in Appendix A very similarly to the proof of Theorem 2 of [1].

Lemma 1 (Universal random-coding achievability) *Let μ be a probability distribution on $\hat{\mathcal{X}}^n$ with $\mu(\hat{\boldsymbol{x}}) \geq B^{-n}$ for every $\hat{\boldsymbol{x}}$ in its support, for some $B \geq 1$ (possibly depending on n). For every $A > B$ and $\epsilon > 0$ there is $N = N(\epsilon, A, B)$ such that, for every $n > N$, there is a codebook of A^n sequences, drawn independently at random from μ, such that for every $\boldsymbol{x}$ and every target $\mathcal{E}(\boldsymbol{x}) \subset \hat{\mathcal{X}}^n$, the first codeword landing in $\mathcal{E}(\boldsymbol{x})$ gives codelength*

$$L(\boldsymbol{x}) \leq -\log \mu[\mathcal{E}(\boldsymbol{x})] + (2+\epsilon)\log n + c, \tag{5}$$

$c > 0$ depends only on A. The restriction to $n > N(\epsilon)$ cannot be dropped: $N(\epsilon) \to \infty$ as $\epsilon \to 0^+$, and the argument fails outright at $\epsilon = 0$ — see the proof.

Intuitively, each candidate lands in $\mathcal{E}(\boldsymbol{x})$ independently with probability $\mu[\mathcal{E}(\boldsymbol{x})]$, so the first success typically takes about $1/\mu[\mathcal{E}(\boldsymbol{x})]$ draws, and encoding this number costs roughly about $\log\left(1/\mu[\mathcal{E}(\boldsymbol{x})]\right) = -\log \mu[\mathcal{E}(\boldsymbol{x})]$ bits, which is the bound's leading term.

3 The type-covering skeleton

Even in Rimoldi's own, classical (probabilistic, non-universal) setting — with $X, \hat{X}_1, \hat{X}_2$ single-symbol random variables whose joint law is the classical auxiliary distribution, and the actual scheme operating on the length-n sequences $X^n, \hat{X}_1^n, \hat{X}_2^n$, each drawn i.i.d. according to it — the two-stage rate region is not achieved by letting Stage 1 choose $\hat{X}_1^n$ to merely satisfy the first distortion

constraint, then designing an optimal Stage-2 code for whatever $\hat{X}_1^n$ results. The standard covering-lemma proof of Stage 1's own achievability draws candidate sequences i.i.d., each coordinate from the auxiliary's own marginal law, and searches for the first one *jointly typical* with X^n under the specific joint law — not merely distortion-compliant. Bounding Stage 2's rate, given only that $\hat{X}_1^n$ meets D_1, requires an expectation of the form $E\{-\log(\text{Stage-2 success probability} \mid \hat{X}_1^n)\}$ over that uncontrolled, random $\hat{X}_1^n$; since $-\log(\cdot)$ is convex, Jensen's inequality bounds this only from below, the wrong direction for achievability. The gap is between two same-size Stage-2 codebooks: one with codewords drawn fully independently from the compound law, achieving Jensen's own lower bound; the real scheme's own, whose codewords form a single *cloud* around whichever $\hat{X}_1^n$ results, conditionally — not marginally — independent, achieving the true, larger rate instead. This is exactly why Rimoldi's characterization is stated as a union over auxiliary *joint* distributions of $(X, \hat{X}_1, \hat{X}_2)$, not over distortion-feasible marginals: fixing $\hat{X}_1$'s relationship to X at the single-letter level pins down $\hat{X}_1^n$'s relationship to X^n throughout, removing the expectation, hence the Jensen gap, entirely. This paper's own joint-type matching, throughout what follows, is the individual-sequence transplant of that same cure.

3.1 A single double-counting lemma, instantiated twice

The following lemma and its proof already appear in [1, Theorem 1]. We restate it here, in the single-reconstruction case first, for the sake of completeness, and then note separately how it extends to the two-stage reconstruction case.

Lemma 2 (Double counting) *Given two types P^ℓ (on $\mathcal{X}^\ell$) and Q^ℓ (on $\hat{\mathcal{X}}^\ell$) the following holds. For any two representatives, $\boldsymbol{x} \in \mathcal{T}_n(P^\ell)$ and $\hat{\boldsymbol{x}} \in \mathcal{T}_n(Q^\ell)$,*

$$\frac{|\mathcal{T}_n(Q^\ell) \cap \mathcal{S}(\boldsymbol{x}, D)|}{|\mathcal{T}_n(Q^\ell)|} = \frac{|\mathcal{T}_n(P^\ell) \cap \hat{\mathcal{S}}(\hat{\boldsymbol{x}}, D)|}{|\mathcal{T}_n(P^\ell)|}, \tag{6}$$

and both sides are independent of the particular representatives chosen.

The significance of this lemma is as follows. The left-hand side is readily interpreted as $U_{Q^\ell}[\mathcal{S}(\boldsymbol{x}, D)]$, which is the probability that a randomly selected codeword under U_{Q^ℓ} would fall within distance nD from $\boldsymbol{x}$. The right-hand side is the reciprocal of the sphere-covering ratio in the input space. While the former is intimately related to random-coding achievability, the latter is a

central building block of the converse. Intuitively, Lemma 2 can be considered as a combinatorial analogue (considering the method of types) to the simple fact that the rate-distortion function can be either represented as the minimum of $H(X) - H(X|\hat{X}) = I(X;\hat{X})$ or as the minimum of $H(\hat{X}) - H(\hat{X}|X) = I(X;\hat{X})$, both taken subject to the distortion constraint.

Proof:[Proof sketch] Let $N(D) \triangleq \sum_{\boldsymbol{x},\hat{\boldsymbol{x}}} \mathbf{1}\{\boldsymbol{x} \in \mathcal{T}_n(P^\ell),\ \hat{\boldsymbol{x}} \in \mathcal{T}_n(Q^\ell),\ d(\boldsymbol{x},\hat{\boldsymbol{x}}) \le nD\}$. Counting it by grouping over $\boldsymbol{x}$ first, or over $\hat{\boldsymbol{x}}$ first, gives $N(D) = |\mathcal{T}_n(P^\ell)| \cdot |\mathcal{T}_n(Q^\ell) \cap \mathcal{S}(\boldsymbol{x}, D)|$ (any representative $\boldsymbol{x}$) and $N(D) = |\mathcal{T}_n(Q^\ell)| \cdot |\mathcal{T}_n(P^\ell) \cap \hat{\mathcal{S}}(\hat{\boldsymbol{x}}, D)|$ (any representative $\hat{\boldsymbol{x}}$): each equality uses that distortion depends only on the first-order joint empirical distribution (Section 2), so any blockwise permutation carrying one representative to another carries the corresponding intersection onto itself bijectively, leaving its size unchanged — hence $|\mathcal{T}_n(Q^\ell) \cap \mathcal{S}(\boldsymbol{x}, D)|$ does not depend on which $\boldsymbol{x}$ is chosen, nor $|\mathcal{T}_n(P^\ell) \cap \hat{\mathcal{S}}(\hat{\boldsymbol{x}}, D)|$ on which $\hat{\boldsymbol{x}}$ is chosen. Equating the two expressions for $N(D)$ and dividing by $|\mathcal{T}_n(P^\ell)| \cdot |\mathcal{T}_n(Q^\ell)|$ gives the claim. ■

Extension to two reconstructions, and to type-match targets. This extends Lemma 2 to the pair $(\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2)$, treated as a single reconstruction over the product alphabet $\hat{\mathcal{X}}_1 \times \hat{\mathcal{X}}_2$, exactly as a single $\hat{\boldsymbol{x}}$ was treated over $\hat{\mathcal{X}}$ above. The lemma applies verbatim: $\mathcal{S}_2(\boldsymbol{x}, D_1, D_2)$, and likewise $\mathcal{T}(W^\ell|\boldsymbol{x}) \triangleq \{\hat{\boldsymbol{x}} : \hat{P}_{\boldsymbol{x},\hat{\boldsymbol{x}}} = W^\ell\}$ — now with $\hat{\boldsymbol{x}} = (\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2)$ ranging over the product alphabet — are each still invariant under a common blockwise permutation of $\boldsymbol{x}$ and the reconstruction, for exactly the same reason $\mathcal{S}(\boldsymbol{x}, D)$ is (Section 2) — two distortion constraints, or a full type match, are no different from one in this respect. So nothing changes beyond relabeling: $U_{W^\ell}[\mathcal{S}_2(\boldsymbol{x}, D_1, D_2)] = |\mathcal{T}_n(P^\ell) \cap \hat{S}_2|/|\mathcal{T}_n(P^\ell)|$ and $U_{Q^\ell}[\mathcal{T}(W^\ell|\boldsymbol{x})] = |\mathcal{T}_n(P^\ell) \cap \mathcal{T}(W^\ell|\hat{\boldsymbol{x}})|/|\mathcal{T}_n(P^\ell)|$ — the form the rest of this paper uses throughout, for both one and two reconstructions. This concludes the extension.

The single-reconstruction case — target $\mathcal{T}(W_1^\ell|\boldsymbol{x})$ over the alphabet $\hat{\mathcal{X}}_1^\ell$ — is what Section 3 uses throughout; the two-stage case — target $\mathcal{T}(W^\ell|\boldsymbol{x})$ (the full-triple type match) over the doubled alphabet $\hat{\mathcal{X}}_1^\ell \times \hat{\mathcal{X}}_2^\ell$ — is what the proofs below use. Both are used without re-deriving the lemma.

3.2 Converse

Lemma 2 gives the bound this subsection is built from: a rate lower bound for any code, holding for the vast majority of source sequences of a given type, but conditional on the code actually realizing that type on the sequence in question. Theorem 1 below states this bound formally, its proof

deferred to Appendix B; it serves as a preparatory, supporting step toward the actual converse, Theorem 2, established next.

Theorem 1 (Type-conditional rate bound) *Let $\Phi = (\phi_1, \phi_2, \psi_1, \psi_2)$ be any two-stage code that satisfies the two distortion constraints. Fix any joint type W^ℓ with X-marginal $W_X = P^\ell$ and let ϵ be an arbitrarily small positive real. For at least a $(1 - 2n^{-\epsilon})$ fraction of the source sequences $\{\boldsymbol{x}\}$ in $\mathcal{T}_n(P^\ell)$ the following holds: If $(\boldsymbol{x}, \psi_1(\phi_1(\boldsymbol{x})), \psi_2(\phi_1(\boldsymbol{x}), \phi_2(\boldsymbol{x})))$ falls in joint type W^ℓ, then*

$$L_1(\boldsymbol{x}) \geq nR_1^*(W^\ell) - \epsilon \log n, \qquad L_1(\boldsymbol{x}) + L_2(\boldsymbol{x}) \geq nR^*(W^\ell) - \epsilon \log n, \tag{7}$$

where

$$R_1^*(W^\ell) \triangleq -\frac{1}{n} \log \left(U_{Q_1^\ell}[\mathcal{T}(W_1^\ell|\boldsymbol{x})]\right), \qquad R^*(W^\ell) \triangleq -\frac{1}{n} \log \left(U_{W^\ell}[\mathcal{T}(W^\ell|\boldsymbol{x})]\right). \tag{8}$$

The single-description bound is the special case of the marginal bound with the second stage absent.

Why this is not yet a converse. Theorem 1's guarantee is conditional on the event described in the "if" clause above — but Φ and W^ℓ are chosen independently of each other, and nothing forces Φ to realize that particular type on more than a negligible subset of $\mathcal{T}_n(P^\ell)$. Concretely: for a fixed Φ, let $G_\Phi(W^\ell) \triangleq \{\boldsymbol{x} \in \mathcal{T}_n(P^\ell) : (\boldsymbol{x}, \psi_1(\phi_1(\boldsymbol{x})), \psi_2(\phi_1(\boldsymbol{x}), \phi_2(\boldsymbol{x}))) \text{ has joint type } W^\ell\}$. Theorem 1 only bounds how much of $G_\Phi(W^\ell)$ can violate the rate bound, as a fraction of $|\mathcal{T}_n(P^\ell)|$ — not as a fraction of $|G_\Phi(W^\ell)|$ itself. A code that realizes type W^ℓ only on a small, handpicked subset of $\mathcal{T}_n(P^\ell)$ (spreading its remaining output thinly across other types) can make $G_\Phi(W^\ell)$ itself smaller than Theorem 1's own exceptional set, and so can violate the rate bound on *all* of $G_\Phi(W^\ell)$ without contradicting the theorem as stated. Turning Theorem 1 into an actual converse takes tying the exceptional-set size to the number of types available, not to a single, pre-chosen W^ℓ in isolation — done in Theorem 2 next, by union-bounding over every type in $\mathcal{W}$ simultaneously.

Let $\mathcal{W}$ be the family of types with X-marginal P^ℓ meeting both distortion budgets, with $\mathcal{T}_n(W^\ell) \neq \emptyset$; since $\ell = \ell(n) = \Theta(\log n)$ is fixed at each n and the alphabets are fixed, $|\mathcal{W}| \leq n^c$ for a constant $c = c(|\mathcal{X}|, \hat{K}_1, \hat{K}_2)$. Write

$$\mathcal{R}(\boldsymbol{x}) \triangleq \bigcup_{W^\ell \in \mathcal{W}} \{(R_1, R_2) : R_1 \geq R_1^*(W^\ell),\ R_1 + R_2 \geq R^*(W^\ell)\} \tag{9}$$

for the union of every type's own quadrant — Section 3.4 below shows this is exactly the achievable region, once Theorem 3's own achievability is in place.

Theorem 2 (Type converse) *For every type P^ℓ (on $\mathcal{X}^\ell$), every two-stage code Φ that satisfies both distortion constraints, and every $\epsilon > 0$, there is at most a $2n^{-\epsilon}$ fraction of $\boldsymbol{x} \in \mathcal{T}_n(P^\ell)$ for which Φ's own rate pair $(\rho_1(\boldsymbol{x}), \rho_1(\boldsymbol{x}) + \rho_2(\boldsymbol{x}))$ falls outside $\mathcal{R}(\boldsymbol{x})$, up to $O(\log n/n)$ slack in each coordinate.*

Proof: *Step 1: a single exceptional set, uniform over types.* For each $W^\ell \in \mathcal{W}$, apply Theorem 1 to Φ at this W^ℓ, with ϵ there set to $\epsilon + c$: this excludes a set $E_{W^\ell} \subseteq \mathcal{T}_n(P^\ell)$ of size at most $2n^{-\epsilon-c}|\mathcal{T}_n(P^\ell)|$, outside of which Φ's output having type W^ℓ implies both

$$L_1(\boldsymbol{x}) \geq nR_1^*(W^\ell) - (\epsilon + c)\log n \qquad \text{and} \qquad L_1(\boldsymbol{x}) + L_2(\boldsymbol{x}) \geq nR^*(W^\ell) - (\epsilon + c)\log n$$

— a single exceptional set for both bounds, since Theorem 1 already gives them together, not as two separately-excluded events. Let $E \triangleq \bigcup_{W^\ell \in \mathcal{W}} E_{W^\ell}$; since $|\mathcal{W}| \leq n^c$, the union bound gives

$$|E| \;\leq\; n^c \cdot 2n^{-\epsilon-c}|\mathcal{T}_n(P^\ell)| \;=\; 2n^{-\epsilon}|\mathcal{T}_n(P^\ell)|.$$

Step 2: apply it at whichever type is actually realized. Fix any $\boldsymbol{x} \notin E$. Since Φ meets both distortion constraints, its output on $\boldsymbol{x}$ has some joint type $W^{\ell,*} \in \mathcal{W}$ (distortion depends on $\boldsymbol{x}$ and the reconstruction only through their joint type). As $\boldsymbol{x} \notin E_{W^{\ell,*}}$, the bounds above hold at $W^\ell = W^{\ell,*}$, so $(\rho_1(\boldsymbol{x}), \rho_1(\boldsymbol{x}) + \rho_2(\boldsymbol{x}))$ lies, up to $O(\log n/n)$ slack, in the quadrant

$$\{(R_1, R_2) : R_1 \geq R_1^*(W^{\ell,*}),\ R_1 + R_2 \geq R^*(W^{\ell,*})\},$$

one of the quadrants unioned to form $\mathcal{R}(\boldsymbol{x})$. Hence the rate pair lies in $\mathcal{R}(\boldsymbol{x})$, up to the same slack, for every $\boldsymbol{x} \notin E$ — and $|E| \leq 2n^{-\epsilon}|\mathcal{T}_n(P^\ell)|$, as claimed. ∎

This holds for every code — type-committed, type-free, or otherwise: each such code has its own exceptional set of at most $2n^{-\epsilon}$ of the $\boldsymbol{x}$'s, outside which it cannot achieve a rate pair outside $\mathcal{R}(\boldsymbol{x})$, whichever type its own output happens to realize. It is matched, on the nose, by both achievability routes ahead, though for different reasons. Theorem 3 (and, via the type-committed LZ route, Theorem 4) realizes a single, chosen type W^ℓ on *every* $\boldsymbol{x}$ it is run on, so Theorem 1 applies to it

directly, at that one type, with no need for the union-over-types argument just given. Theorem 5's type-free construction, by contrast, never commits to a specific type — exactly why it needs this result, region-level and unconditional in $\boldsymbol{x}$, rather than Theorem 1 applied at a single, pre-chosen type.

Relative to [1, Theorem 1]'s own single-stage converse, this result is the stronger one in the one respect the proof of Theorem 1 addresses: [1] bounds the fraction of *codewords* (distinct reconstructions) satisfying the length bound, while the counting argument there, via Lemma 2's type-match extension, converts this directly into a bound on the fraction of *source sequences* instead.

3.3 Achievability

The bound just proved is matched exactly: for every joint type meeting both distortion budgets, a scheme exists reaching that type's own rate pair.

Theorem 3 (Type-class achievability) *Fix any joint type W^ℓ with X-marginal P^ℓ, meeting both distortion constraints, with $\mathcal{T}_n(W^\ell) \neq \emptyset$. For every $\epsilon > 0$ and all sufficiently large n, there is a two-stage scheme achieving:*

$$L_1(\boldsymbol{x}) \leq nR_1^*(W^\ell) + (2+\epsilon)\log n + c, \qquad L_1(\boldsymbol{x}) + L_2(\boldsymbol{x}) \leq nR^*(W^\ell) + (4+\epsilon)\log n + c, \tag{10}$$

with $R_1^*(W^\ell)$ and $R^*(W^\ell)$ as defined in Theorem 1, and $c = c(\epsilon)$.

Dividing by n gives $\rho_1(\boldsymbol{x}) \to R_1^*(W^\ell)$, $\rho_1(\boldsymbol{x}) + \rho_2(\boldsymbol{x}) \to R^*(W^\ell)$, matching Theorem 1 exactly, for *every* such W^ℓ: no further restriction on W^ℓ is needed, since achievability and converse are stated in terms of the same type-probability quantity throughout.

Proof: *Stage 1.* Since $\mathcal{T}_n(W^\ell) \neq \emptyset$, some $\hat{\boldsymbol{x}}_1$ achieves $\hat{P}_{\boldsymbol{x}\hat{\boldsymbol{x}}_1} = W_1^\ell$ for $\boldsymbol{x} \in \mathcal{T}_n(P^\ell)$, so $u_1 \triangleq U_{Q_1^\ell}[\mathcal{T}(W_1^\ell|\boldsymbol{x})] > 0$ and $R_1^*(W^\ell) < \infty$. Q_1^ℓ, being uniform on a type class of size at most $\hat{K}_1^n$, satisfies $Q_1^\ell(\hat{\boldsymbol{x}}) = 1/|\mathcal{T}_n(Q_1^\ell)| \geq \hat{K}_1^{-n}$ for every $\hat{\boldsymbol{x}}$ in its support: Lemma 1's hypothesis, with $B = \hat{K}_1$. Draw independent candidates $\boldsymbol{z}$ i.i.d. $\sim Q_1^\ell$ and search for the first one achieving exactly joint type W_1^ℓ with $\boldsymbol{x}$; this is Lemma 1 with $\mu = Q_1^\ell$ (any $A > \hat{K}_1$) and $\mathcal{T}(W_1^\ell|\boldsymbol{x})$ in place of the target, giving $L_1(\boldsymbol{x}) \leq -\log(u_1) + (2+\epsilon)\log n + c$, once u_1 itself is computed.

Treating the $m = n/\ell$ blocks of $(\boldsymbol{x}, \boldsymbol{z})$ as symbols over the super-alphabets $\mathcal{X}^\ell, \hat{\mathcal{X}}_1^\ell$: by the method-of-types, the probability of a random i.i.d. sequence falling into a given type decays exponentially in the divergence between that type and the underlying distribution (e.g. [25]) — applied here to the pair-valued sequence $(\boldsymbol{x}_i, \boldsymbol{z}_i)$, $i = 1, \ldots, m$, with $\boldsymbol{x}$ fixed (so its own randomness is degenerate) and $\boldsymbol{z}$'s blocks i.i.d. $\sim Q_1^\ell$, giving underlying distribution $P^\ell \times Q_1^\ell$ and target type W_1^ℓ —

$$u_1 = \Pr\left[(\boldsymbol{x}, \boldsymbol{z}) \text{ has type } W_1^\ell\right] = 2^{-mD(W_1^\ell \| P^\ell \times Q_1^\ell) + O(\log m)}. \tag{11}$$

Since $D(W_1^\ell \| P^\ell \times Q_1^\ell) = \hat{I}_{W_1^\ell}(X; \hat{X}_1)$ whenever W_1^ℓ's own X-marginal is P^ℓ (the standard identity relating a joint type's divergence from the product of its own marginals to its mutual information), and $\hat{I}_{W_1^\ell}(X; \hat{X}_1)/\ell = R_1^*(W^\ell)$ by definition (Theorem 1),

$$u_1 = \exp_2\{-m\hat{I}_{W_1^\ell}(X; \hat{X}_1) + O(\log m)\} = 2^{-nR_1^*(W^\ell) + O(\log n)}, \tag{12}$$

giving $L_1(\boldsymbol{x}) \le nR_1^*(W^\ell) + (2+\epsilon)\log n + c$ directly.

Stage 2. Given $\hat{\boldsymbol{x}}_1$, draw $\hat{\boldsymbol{x}}_2$'s blocks i.i.d., the i-th block sampled $\sim W^\ell_{\hat{X}_2|\hat{X}_1}(\cdot|\hat{x}_{1,i})$ (W^ℓ's own $(\hat{X}_2|\hat{X}_1)$ conditional law), and search for the first one completing the *full* triple to joint type W^ℓ with $(\boldsymbol{x}, \hat{\boldsymbol{x}}_1)$; write p for this event's per-candidate probability.

The identical estimate applies conditionally, treating the type W_1^ℓ of $(\boldsymbol{x}, \hat{\boldsymbol{x}}_1)$ as fixed and $\hat{\boldsymbol{x}}_2$'s blocks as drawn i.i.d. from the conditional law $W^\ell_{\hat{X}_2|\hat{X}_1}(\cdot|a_1)$ given each block's own $\hat{\boldsymbol{x}}_1$-value a_1: standard conditional type-counting (again [25]) gives

$$p = \Pr[\text{full type } W^\ell] = \exp_2\{-mD(W^\ell \| W_1^\ell \times W^\ell_{\hat{X}_2|\hat{X}_1} | W_1^\ell) + O(\log m)\}, \tag{13}$$

where $D(W^\ell \| W_1^\ell \times W^\ell_{\hat{X}_2|\hat{X}_1} | W_1^\ell) \triangleq \sum_{a,a_1} W_1^\ell(a, a_1) D(W^\ell(\cdot|a, a_1) \| W^\ell_{\hat{X}_2|\hat{X}_1}(\cdot|a_1))$ is exactly the conditional mutual information $\hat{I}_{W^\ell}(X; \hat{X}_2|\hat{X}_1)$ (the same identity as Stage 1's, now applied conditionally on $\hat{X}_1$), so $p = \exp_2\{-m\hat{I}_{W^\ell}(X; \hat{X}_2|\hat{X}_1) + O(\log m)\}$. By the exact entropy chain rule, $\hat{I}_{W^\ell}(X; \hat{X}_1, \hat{X}_2) = \hat{I}_{W_1^\ell}(X; \hat{X}_1) + \hat{I}_{W^\ell}(X; \hat{X}_2|\hat{X}_1)$, and the identical type-probability estimate applied to the full triple directly (uniform sampling of $(\hat{X}_1, \hat{X}_2)$ pairs within their own joint type class, in place of $\hat{X}_1$ alone) gives $R^*(W^\ell) = \hat{I}_{W^\ell}(X; \hat{X}_1, \hat{X}_2)/\ell$; so $p = 2^{-n[R^*(W^\ell) - R_1^*(W^\ell)] + O(\log n)}$ exactly, with no assumption that $X - \hat{X}_1 - \hat{X}_2$ forms a Markov chain: the sampling is conditional on $\hat{X}_1$ alone (legitimate, since $\hat{X}_1$ is known to both encoder and decoder after Stage 1), while the success event checks the full joint type with $(\boldsymbol{x}, \hat{\boldsymbol{x}}_1)$ together, not merely the $(\hat{X}_1, \hat{X}_2)$ marginal relationship.

Applying Lemma 1 again, conditionally, gives $L_2(\boldsymbol{x}) \le n[R^*(W^\ell) - R_1^*(W^\ell)] + (2+\epsilon)\log n + c$, so $L_1(\boldsymbol{x}) + L_2(\boldsymbol{x}) \le nR^*(W^\ell) + 2(2+\epsilon)\log n + 2c$, which is the claimed bound after relabeling $2\epsilon, 2c$ as ϵ, c (i.e. for a target ϵ, apply both stages with $\epsilon/2$). ∎

3.4 The achievable region and its Pareto frontier

Ranging the matching converse-achievability pair over every type gives the full region, not just one corner of it: combining Theorem 2's own converse with Theorem 3's own achievability, $\mathcal{R}(\boldsymbol{x})$ — defined above — is exactly the achievable region.

The Pareto-optimal points of $\mathcal{R}(\boldsymbol{x})$ — those undominated in both coordinates — form the lower envelope $R_2^{\min}(R_1) = \min_{W^\ell \in \mathcal{W}: R_1^*(W^\ell) \le R_1} \max(0, R^*(W^\ell) - R_1)$; not every $W^\ell \in \mathcal{W}$ need itself be Pareto-optimal, since the union automatically discards any that are dominated. A specific frontier point is reached by fixing one coordinate as a budget and minimizing the other over $\mathcal{W}$ (ϵ-constraint method): for a target R_1-budget b, $\min_{W^\ell: R_1^*(W^\ell) \le b} R^*(W^\ell) - b$; symmetrically for a total-budget c, $\min_{W^\ell: R^*(W^\ell) \le c} R_1^*(W^\ell)$. Since $\mathcal{W}$ is a finite set (finitely many joint types on a finite alphabet, for fixed n, ℓ), the minimum defining $R_2^{\min}(R_1)$ is a minimum over finitely many values: the frontier is always achieved by some $W^\ell \in \mathcal{W}$, not merely approached.

Which $W^\ell \in \mathcal{W}$ to target is not resolved by either theorem — it is a free design choice, exactly as the choice of auxiliary distribution is free in Rimoldi's classical characterization; the ϵ-constraint recipe above describes how to make it for a desired operating point.

4 From types to Lempel–Ziv complexity

4.1 Why this is not automatic

Section 1.2 already located the difficulty: $\hat{\boldsymbol{x}}_1$ must also serve as good side information for Stage 2, not merely satisfy D_1 on its own. Section 3's own joint-type matching exists precisely to avoid the Jensen obstruction explained there. The same requirement carries over unchanged when Stage 1 draws from U — the full, unrestricted, LZ-weighted measure — instead of a type-restricted one: nothing about switching measures removes the need to fix $\hat{x}_1$'s *type*, not merely its distortion.

The underlying idea is as follows. Build joint typicality between $\hat{\boldsymbol{x}}_1$ and $\boldsymbol{x}$ directly into Stage 1's search criterion, so Stage 2 conditions on a candidate is *guaranteed* to keep its own rate low (The-

orem 4). This is the same criterion Section 3 already uses for its own, type-restricted Stage 1; checking it directly against U — with no shortcut around it — is what the rest of this section works out. For a full description of the random coding scheme, we first pause to define certain ingredients associated with the LZ78 algorithm and its conditional version.

For $\hat{\boldsymbol{x}}_1 \in \hat{\mathcal{X}}_1^n$: the *incremental parsing procedure* of the LZ78 algorithm [2] sequentially parses $\hat{\boldsymbol{x}}_1$ into phrases, each new phrase being the shortest string that has not been encountered before as a parsed phrase, with the possible exception of the last phrase, which may be incomplete. For example, the incremental parsing of $\hat{x}_1 = 011010011000100$ is $0, 1, 10, 100, 11, 00, 01, 00$. Let $c(\hat{\boldsymbol{x}}_1)$ denote the resulting number of phrases (eight, in this example), and let $LZ(\hat{\boldsymbol{x}}_1)$ denote the length, in bits, of the LZ78 binary compressed code for $\hat{\boldsymbol{x}}_1$ — approximately $c(\hat{\boldsymbol{x}}_1) \log c(\hat{\boldsymbol{x}}_1)$, a standard mechanical fact about the algorithm's own construction. Let $U(\hat{\boldsymbol{x}}_1) \propto 2^{-LZ(\hat{\boldsymbol{x}}_1)}$ on $\hat{\mathcal{X}}_1^n$ — normalizable since $\hat{\mathcal{X}}_1^n$ is finite, and in fact $\sum_{\hat{\boldsymbol{x}}_1} 2^{-LZ(\hat{\boldsymbol{x}}_1)} \leq 1$ by Kraft's inequality (LZ78 being uniquely decodable), a fact used later — the same, fixed, source-independent measure throughout; no restriction of its support is used anywhere below.

Similarly, for a pair of sequences $(\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2)$ of the same length n: apply LZ78's incremental parsing to the sequence of *pairs* $((\hat{x}_{2,1}, \hat{x}_{1,1}), \ldots, (\hat{x}_{2,n}, \hat{x}_{1,n}))$, treating each pair as one product-alphabet symbol [26]; let $c(\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2)$ denote the resulting number of distinct phrases, $c'(\hat{\boldsymbol{x}}_1)$ — the number of distinct $\hat{\boldsymbol{x}}_1$-phrases this same joint parse induces (generally different from $\hat{\boldsymbol{x}}_1$'s own, unconditional phrase count), $\hat{\boldsymbol{x}}_1(l)$ the l-th such phrase, and $c_l(\hat{\boldsymbol{x}}_2|\hat{\boldsymbol{x}}_1)$ the number of its occurrences — equivalently, the number of distinct $\hat{\boldsymbol{x}}_2$-phrases jointly appearing with it — for $l = 1, \ldots, c'(\hat{\boldsymbol{x}}_1)$. The *conditional LZ complexity* of $\hat{\boldsymbol{x}}_2$ given $\hat{\boldsymbol{x}}_1$ is

$$\rho_{LZ}(\hat{\boldsymbol{x}}_2|\hat{\boldsymbol{x}}_1) \triangleq \frac{1}{n} \sum_{l=1}^{c'(\hat{\boldsymbol{x}}_1)} c_l(\hat{\boldsymbol{x}}_2|\hat{\boldsymbol{x}}_1) \log c_l(\hat{\boldsymbol{x}}_2|\hat{\boldsymbol{x}}_1) \tag{14}$$

[27] — the individual-sequence analogue of conditional entropy, built to exploit empirical correlation between $\hat{\boldsymbol{x}}_2$ and $\hat{\boldsymbol{x}}_1$, not merely $\hat{\boldsymbol{x}}_2$'s own statistics: $\hat{\boldsymbol{x}}_1$ enters throughout as side information available to both encoder and decoder, never as a static resource $\hat{\boldsymbol{x}}_2$ draws phrases from. Let $LZ(\hat{\boldsymbol{x}}_2|\hat{\boldsymbol{x}}_1)$ denote the codelength of the conditional LZ78 coding scheme built on this side information [26]: $LZ(\hat{\boldsymbol{x}}_2|\hat{\boldsymbol{x}}_1) \leq n\rho_{LZ}(\hat{\boldsymbol{x}}_2|\hat{\boldsymbol{x}}_1) + n\hat{\epsilon}(n)$, $\hat{\epsilon}(n) \to 0$ — the conditional analogue of the unconditional mechanical bound above. Let $V(\hat{\boldsymbol{x}}_2|\hat{\boldsymbol{x}}_1) \propto 2^{-LZ(\hat{\boldsymbol{x}}_2|\hat{\boldsymbol{x}}_1)}$ on $\hat{\mathcal{X}}_2^n$ be the resulting conditional measure — Stage 2's satellite codebook is drawn from $V(\cdot|\hat{\boldsymbol{x}}_1)$ once $\hat{\boldsymbol{x}}_1$ is fixed. The following lemma will be

useful throughout — see, e.g., [1], [18], as well as [19] and [5, Lemma 13.5.5] for somewhat different, but intimately related family of inequalities, collectively referred to as *Ziv's inequality*.

Lemma 3 (Ziv's inequality) *For any $\hat{\boldsymbol{x}} \in \mathcal{T}_n(Q^\ell)$,*

$$LZ(\hat{\boldsymbol{x}}) \le \frac{n\hat{H}(Q^\ell)}{\ell} + n\Delta_n(\ell), \qquad \Delta_n(\ell) \to 1/\ell. \tag{15}$$

This is a standard fact about the LZ78 algorithm, well known from [2] (see also [1]), and not proved here.

Consequently, for any $\mathcal{S} \subseteq \mathcal{T}_n(Q^\ell)$,

$$U(\mathcal{S}) \ge U_{Q^\ell}(\mathcal{S}) \cdot 2^{-n\Delta_n(\ell)}. \tag{16}$$

To see why (**??**) holds true, use $U(\hat{\boldsymbol{x}}) \ge 2^{-LZ(\hat{\boldsymbol{x}})}$ (which holds thanks to Kraft's inequality, since the normalizing constant of $U(\cdot)$ is at most 1, owing to the unique decodability of the LZ78 algorithm), combined with the pointwise bound $2^{-LZ(\hat{\boldsymbol{x}})} \ge 2^{-n\hat{H}(Q^\ell)/\ell - n\Delta_n(\ell)}$ from Lemma 3, and sum over $\hat{\boldsymbol{x}} \in \mathcal{S}$; then use the standard type-size formula $|\mathcal{T}_n(Q^\ell)| = 2^{n\hat{H}(Q^\ell)/\ell + O(\log n)}$ (its own $O(\log n)$ error already folded into $\Delta_n(\ell)$'s own $O(\log n/n)$ term):

$$\begin{aligned} U(\mathcal{S}) = \sum_{\hat{\boldsymbol{x}} \in \mathcal{S}} U(\hat{\boldsymbol{x}}) \;&\ge\; \sum_{\hat{\boldsymbol{x}} \in \mathcal{S}} 2^{-LZ(\hat{\boldsymbol{x}})} \;\ge\; |\mathcal{S}| \cdot 2^{-n\hat{H}(Q^\ell)/\ell - n\Delta_n(\ell)} \\ &= \; |\mathcal{S}| \cdot |\mathcal{T}_n(Q^\ell)|^{-1} \cdot 2^{-n\Delta_n(\ell)} \;=\; U_{Q^\ell}(\mathcal{S}) \cdot 2^{-n\Delta_n(\ell)}, \end{aligned} \tag{17}$$

using $U_{Q^\ell}(\mathcal{S}) \stackrel{\Delta}{=} |\mathcal{S}|/|\mathcal{T}_n(Q^\ell)|$ by definition (no restriction or renormalization of U needed, since $\mathcal{S}$ is already contained in a single type).

This second, set-level bound is the bridge Theorem 4 below uses to identify what its search achieves with Section 3's type-based targets.

Lemma 4 (Ziv's inequality, conditional version) *For $\hat{\boldsymbol{x}}_2$ of the conditional type consistent with W^ℓ given $\hat{\boldsymbol{x}}_1$:*

$$LZ(\hat{\boldsymbol{x}}_2|\hat{\boldsymbol{x}}_1) \le n\hat{H}_{W^\ell}(\hat{X}_2|\hat{X}_1)/\ell + n\Delta'_n(\ell), \qquad \Delta'_n(\ell) \to 0. \tag{18}$$

This is the direct conditional analogue of Lemma 3, for Ziv's conditional LZ construction, first established in [28]. This bound holds by the following consideration: realize a conditional block

Shannon code on the ℓ-blocks of $\hat{\boldsymbol{x}}_2$, using the blocks of $\hat{\boldsymbol{x}}_1$ as side information. Such an encoder can be implemented as a finite-state encoder. Ziv and Lempel's own finite-state converse, extended to side information available at both encoder and decoder [26], lower-bounds any such encoder's codelength by (essentially) the conditional LZ complexity, for every individual sequence pair.

Now, observe that by Lemma 3, $LZ(\hat{\boldsymbol{x}}) \leq n(1+\epsilon_n)\log \hat{K}$ for every $\hat{\boldsymbol{x}} \in \hat{\mathcal{X}}^n$ (the worst case, exactly as in [1]), so $U(\hat{\boldsymbol{x}}) \geq 2^{-LZ(\hat{\boldsymbol{x}})} \geq 2^{-n(1+\epsilon_n)\log \hat{K}}$ uniformly: $\mu = U$ satisfies Lemma 1's hypothesis with $B = \hat{K}^{1+\epsilon_n} \to \hat{K}$. Lemma 1 is used below, and throughout this section, at $\mu = U$.

For $\boldsymbol{x}, \hat{\boldsymbol{x}}_1$ of joint type W_1^ℓ, define $\mathcal{T}(W^\ell|\boldsymbol{x}, \hat{\boldsymbol{x}}_1) \stackrel{\Delta}{=} \{\hat{\boldsymbol{x}}_2 : \ \hat{P}_{\boldsymbol{x}\hat{\boldsymbol{x}}_1\hat{\boldsymbol{x}}_2} = W^\ell\}$. By the method of types [25],

$$|\mathcal{T}(W^\ell|\boldsymbol{x}, \hat{\boldsymbol{x}}_1)| = 2^{m\hat{H}_{W^\ell}(\hat{X}_2|X,\hat{X}_1)+O(\log m)}. \tag{19}$$

Lemma 5 (Stage-2 coverage, for every candidate of the right type) *Fix a D_1-compatible type W_1^ℓ. For every $\hat{\boldsymbol{x}}_1 \in \mathcal{T}(W_1^\ell|\boldsymbol{x})$*

$$V[\mathcal{T}(W^\ell|\boldsymbol{x}, \hat{\boldsymbol{x}}_1)|\hat{\boldsymbol{x}}_1] \geq 2^{-n[R^*(W^\ell)-R_1^*(W^\ell)]-n\Delta_n'(\ell)+o(n)}. \tag{20}$$

Proof: By Lemma 4, every $\hat{\boldsymbol{x}}_2$ with $(\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2)$ of type $W^\ell_{\hat{X}_2|\hat{X}_1}$ satisfies $LZ(\hat{\boldsymbol{x}}_2|\hat{\boldsymbol{x}}_1) \leq n\hat{H}_{W^\ell}(\hat{X}_2|\hat{X}_1)/\ell + n\Delta_n'(\ell)$. Summing $2^{-LZ(\hat{\boldsymbol{x}}_2|\hat{\boldsymbol{x}}_1)} \geq 2^{-n\hat{H}_{W^\ell}(\hat{X}_2|\hat{X}_1)/\ell - n\Delta_n'(\ell)}$ over $\mathcal{T}(W^\ell|\boldsymbol{x}, \hat{\boldsymbol{x}}_1)$ and using the count (19) $|\mathcal{T}(W^\ell|\boldsymbol{x}, \hat{\boldsymbol{x}}_1)| = 2^{n\hat{H}_{W^\ell}(\hat{X}_2|X,\hat{X}_1)/\ell+o(n)}$

$$\begin{aligned} V[\mathcal{T}(W^\ell|\boldsymbol{x}, \hat{\boldsymbol{x}}_1)|\hat{\boldsymbol{x}}_1] &\geq |\mathcal{T}(W^\ell|\boldsymbol{x}, \hat{\boldsymbol{x}}_1)| \cdot 2^{-n\hat{H}_{W^\ell}(\hat{X}_2|\hat{X}_1)/\ell - n\Delta_n'(\ell)} \\ &= 2^{-n[\hat{H}_{W^\ell}(\hat{X}_2|\hat{X}_1) - \hat{H}_{W^\ell}(\hat{X}_2|X,\hat{X}_1)]/\ell - n\Delta_n'(\ell)+o(n)}. \end{aligned} \tag{21}$$

The bracketed difference is exactly the conditional mutual information $\hat{I}_{W^\ell}(X; \hat{X}_2|\hat{X}_1) = R^*(W^\ell) - R_1^*(W^\ell)$ (as established just above Theorem 3), giving the claim. ∎

Theorem 4 (LZ achievability matches the type-based converse) *Fix any $W^\ell \in \mathcal{W}$. There is a two-stage LZ78-based random-coding scheme achieving, for all sufficiently large n and every $\boldsymbol{x} \in \mathcal{T}_n(P^\ell)$,*

$$\begin{aligned} L_1(\boldsymbol{x}) &\leq nR_1^*(W^\ell) + n\Delta_n(\ell) + O(\log n), \\ L_1(\boldsymbol{x}) + L_2(\boldsymbol{x}) &\leq nR^*(W^\ell) + n[\Delta_n(\ell) + \Delta_n'(\ell)] + o(n) + O(\log n). \end{aligned} \tag{22}$$

Proof: *Stage 1.* Draw $\hat{\boldsymbol{x}}_1^{(1)}, \ldots, \hat{\boldsymbol{x}}_1^{(A^n)}$ i.i.d. from U — the full, unrestricted, $\boldsymbol{x}$-independent measure, $A > \hat{K}_1$. Search for the first index with $\hat{\boldsymbol{x}}_1^{(\cdot)} \in \mathcal{T}(W_1^\ell|\boldsymbol{x})$. By Theorem 1's own definition of $R_1^*(W^\ell)$, $U_{Q_1^\ell}[\mathcal{T}(W_1^\ell|\boldsymbol{x})] = 2^{-nR_1^*(W^\ell)}$ directly. By Lemma 3, $U[\mathcal{T}(W_1^\ell|\boldsymbol{x})] \geq U_{Q_1^\ell}[\mathcal{T}(W_1^\ell|\boldsymbol{x})] \cdot 2^{-n\Delta_n(\ell)} = 2^{-nR_1^*(W^\ell)-n\Delta_n(\ell)}$. By Lemma 1 with $T = \mathcal{T}(W_1^\ell|\boldsymbol{x})$, $L_1(\boldsymbol{x}) \leq R_1^*(W^\ell)n + n\Delta_n(\ell) + O(\log n)$.

Stage 2. Given this $\hat{\boldsymbol{x}}_1 \in \mathcal{T}(W_1^\ell|\boldsymbol{x})$ — guaranteed, not random — draw a satellite codebook i.i.d. from Ziv's conditional construction $V(\cdot|\hat{\boldsymbol{x}}_1)$. Search for the first index landing in $\mathcal{T}(W^\ell|\boldsymbol{x}, \hat{\boldsymbol{x}}_1)$ — a valid target for the D_2 constraint, since $\mathcal{T}(W^\ell|\boldsymbol{x}, \hat{\boldsymbol{x}}_1) \subseteq \mathcal{S}(\boldsymbol{x}, D_2)$ (the same type-determined-distortion property, applied to W^ℓ's own $(X, \hat{X}_2)$ sub-marginal). By Lemma 5 — applied here at just this one, Stage-1-selected $\hat{\boldsymbol{x}}_1$; Section 4.3 below exploits its full generality — $V[\mathcal{T}(W^\ell|\boldsymbol{x}, \hat{\boldsymbol{x}}_1)|\hat{\boldsymbol{x}}_1] \geq 2^{-n[R^*(W^\ell)-R_1^*(W^\ell)]-n\Delta'_n(\ell)+o(n)}$. By Lemma 1 with $\mathcal{E}(\boldsymbol{x}) = \mathcal{T}(W^\ell|\boldsymbol{x}, \hat{\boldsymbol{x}}_1)$ and the conditional measure $V(\cdot|\hat{\boldsymbol{x}}_1)$ in place of U, $L_2(\boldsymbol{x}) \leq (R^*(W^\ell) - R_1^*(W^\ell))n + n\Delta'_n(\ell) + o(n) + O(\log n)$.

Combining the two stages gives the claim. ∎

Dividing by n and applying Proposition 1 (so $\Delta_n(\ell), \Delta'_n(\ell) \to 0$ as $n \to \infty$, for $\ell = \ell(n)$ as specified there) gives $\rho_1(\boldsymbol{x}) \to R_1^*(W^\ell)$, $\rho_1(\boldsymbol{x}) + \rho_2(\boldsymbol{x}) \to R^*(W^\ell)$, matching Theorem 1 exactly at this W^ℓ — for every $\boldsymbol{x} \in \mathcal{T}_n(P^\ell)$, not merely most, since neither ingredient the proof combines (Lemma 1, Lemma 2's exact type symmetry, Lemma 5's coverage of every candidate) admits an exception.

Searching $\mathcal{T}(W^\ell|\boldsymbol{x}, \hat{\boldsymbol{x}}_1)$ rather than $\mathcal{S}(\boldsymbol{x}, D_2)$ directly is a real restriction in general: direct search achieves at least as good a rate (Section 1.2's fourth point), and $\mathcal{S}(\boldsymbol{x}, D_2)$, unlike $\mathcal{T}(W^\ell|\boldsymbol{x}, \hat{\boldsymbol{x}}_1)$, is not tied to any type, so Lemma 4 has nothing to say about it and no matching lower bound is provable this way. But the gap does not actually manifest here, since Stage 2 is the last stage: waiting for the first $\hat{\boldsymbol{x}}_2$ landing directly in $\mathcal{S}(\boldsymbol{x}, D_2)$ is exponentially equivalent to waiting for the first one jointly typical, given $\boldsymbol{x}, \hat{\boldsymbol{x}}_1$, with the dominant type within that ball — the type that also minimizes the incremental rate $\hat{I}_{W^\ell}(X; \hat{X}_2|\hat{X}_1)$, manifesting part of the Pareto-optimization of the rate pair. Direct search is a legitimate, rate-equivalent alternative here; the type-match route is used regardless, being the one a matching converse can be proved for. With a further stage to follow, this equivalence would not hold: $\hat{\boldsymbol{x}}_2$ would also need to keep that stage's own rate low, which the dominant type within $\mathcal{S}(\boldsymbol{x}, D_2)$ need not do — the same Jensen obstruction Section 3 identifies for Stage 1.

Proposition 1 (Parameter ordering) *Let* $K_* = \max(|\mathcal{X}|, \hat{K}_1, \hat{K}_2)$, $\ell(n) = \lfloor \frac{1}{4} \log n / \log K_* \rfloor$. *Then, as* $n \to \infty$, *the error terms of Theorems 1–3, Lemma 3, and eq. (19) (and their doubled-alphabet forms) all* $\to 0$, *given* $\Delta'_n(\ell)$ *has the same qualitative two-term shape as* $\Delta_n(\ell)$ *(Lemma 4's status).*

Proof: $1/\ell(n) = \Theta(1/\log n) \to 0$; $\hat{K}_1^{\ell(n)}, \hat{K}_2^{\ell(n)}, \hat{K}_1^{\ell(n)} \hat{K}_2^{\ell(n)} \le K_*^{2\ell(n)} = n^{1/2+o(1)}$, so each alphabet-size error term is $O(n^{-1/2+o(1)} \log n) \to 0$. ∎

4.2 A type-free alternative for Stage 1

Theorem 4's Stage 1 search criterion, joint typicality with a chosen type W^ℓ, is defined via that type directly. This is not very satisfying: a construction whose entire point is to move past types (Section 1.2) still routes its lookahead through one. This subsection asks whether Stage 1 can look ahead to Stage 2 without types at all, and shows that the answer is yes.

For any $\hat{\boldsymbol{x}}_1$, define — the sequence-argument analogue of the incremental Stage-2 rate $R^*(W^\ell) - R_1^*(W^\ell)$ a type induces —

$$\rho_2^*(\hat{\boldsymbol{x}}_1) \triangleq -\frac{1}{n} \log V[\mathcal{S}(\boldsymbol{x}, D_2) | \hat{\boldsymbol{x}}_1], \tag{23}$$

the actual conditional-LZ rate $\hat{\boldsymbol{x}}_1$ induces for Stage 2 — built only from U, V, with no type anywhere. For any $\tau \ge 0$, let $\mathcal{G}_\tau \triangleq \mathcal{S}(\boldsymbol{x}, D_1) \cap \{\hat{\boldsymbol{x}}_1 : \rho_2^*(\hat{\boldsymbol{x}}_1) \le \tau\}$.

Type-free schemes. Call a two-stage scheme *type-free* if its search phases include no joint typicality criterion — unlike Theorem 4's own Stage-1 search, which checks membership in $\mathcal{T}(W_1^\ell | \boldsymbol{x})$ directly.

Proposition 2 (Type-free achievability) *For every* $\tau \ge 0$ *and* $\epsilon > 0$, *and all sufficiently large* n, *there is a type-free, two-stage LZ78-based scheme achieving, for every* $\boldsymbol{x}$,

$$L_1(\boldsymbol{x}) \le -\log U[\mathcal{G}_\tau] + (2+\epsilon)\log n + c, \qquad L_1(\boldsymbol{x}) + L_2(\boldsymbol{x}) \le -\log U[\mathcal{G}_\tau] + n\tau + (4+\epsilon)\log n + c, \tag{24}$$

$c = c(\epsilon)$.

Proof: Stage 1 draws from the full, unrestricted U and searches for $\mathcal{G}_\tau$-membership; Lemma 1 (stated for any target set) applies directly, giving $L_1(\boldsymbol{x}) \le -\log U[\mathcal{G}_\tau] + (2+\epsilon)\log n + c$. The

resulting $\hat{\boldsymbol{x}}_1 \in \mathcal{G}_\tau$ satisfies $\rho_2^*(\hat{\boldsymbol{x}}_1) \le \tau$ by construction — guaranteed, not probabilistically. Stage 2 draws from the full $V(\cdot|\hat{\boldsymbol{x}}_1)$ and searches for $\mathcal{S}(\boldsymbol{x}, D_2)$ directly; Lemma 1 applies again, conditionally, giving $L_2(\boldsymbol{x}) \le n\rho_2^*(\hat{\boldsymbol{x}}_1) + (2+\epsilon)\log n + c \le n\tau + (2+\epsilon)\log n + c$, so $L_1(\boldsymbol{x}) + L_2(\boldsymbol{x}) \le -\log U[\mathcal{G}_\tau] + n\tau + 2(2+\epsilon)\log n + 2c$, which is the claimed bound after relabeling $2\epsilon, 2c$ as ϵ, c. ■

Dividing by n gives

$$\rho_1(\boldsymbol{x}) \le -\frac{1}{n}\log U[\mathcal{G}_\tau] + o(1), \qquad \rho_1(\boldsymbol{x}) + \rho_2(\boldsymbol{x}) \le -\frac{1}{n}\log U[\mathcal{G}_\tau] + \tau + o(1), \tag{25}$$

the form used below.

Ranging τ traces a family of achievable points, exactly paralleling the type-indexed family $\mathcal{W}$ of Section 3, but continuously and without reference to any type.

4.3 The type-free scheme matches the converse

Theorem 2 (Section 3) already covers this scheme, like any other: it holds for every code, whatever type its output happens to realize. What remains is only the achievability side: a demonstration that Proposition 2's type-free scheme's own achieved rate actually reaches, rather than merely respects, that converse. The next theorem supplies this, by reusing Lemma 5 (Section 4) — used inside Theorem 4 at only the one, Stage-1-selected $\hat{\boldsymbol{x}}_1$, it already holds for *every* $\hat{\boldsymbol{x}}_1 \in \mathcal{T}(W_1^\ell|\boldsymbol{x})$ simultaneously; that full generality, unused until now, is exactly what is needed here.

Theorem 5 (The type-free scheme reaches every point of $\mathcal{R}(\boldsymbol{x})$) *Recall Proposition 2's construction: Stage 1 draws from U and searches for $\mathcal{G}_\tau$-membership — jointly, D_1-compatible* and *with conditional-LZ rate at most τ — and Stage 2 draws from $V(\cdot|\hat{\boldsymbol{x}}_1)$ and searches directly for D_2. For every $W^\ell \in \mathcal{W}$, running this construction at*

$$\tau(W^\ell) \triangleq R^*(W^\ell) - R_1^*(W^\ell) + \Delta_n'(\ell) \tag{26}$$

achieves

$$\rho_1(\boldsymbol{x}) \;\le\; R_1^*(W^\ell) + \Delta_n(\ell) + o(1), \qquad \rho_1(\boldsymbol{x}) + \rho_2(\boldsymbol{x}) \;\le\; R^*(W^\ell) + \Delta_n(\ell) + \Delta_n'(\ell) + o(1), \tag{27}$$

converging to $(R_1^*(W^\ell), R^*(W^\ell))$ as $n \to \infty$. Ranging W^ℓ over $\mathcal{W}$, the type-free scheme's own achievable curve $\{(\rho_1(\tau), \rho_1(\tau) + \rho_2(\tau)) : \tau \ge 0\}$ therefore contains every corner of $\mathcal{R}(\boldsymbol{x})$ — in particular every point of its Pareto frontier.

Proof: *Step 1: the entire type-match set lies in $\mathcal{G}_{\tau(W^\ell)}$.* Unpacking $\mathcal{G}_\tau$'s own two-part definition, this means showing, for every $\hat{\boldsymbol{x}}_1 \in \mathcal{T}(W_1^\ell|\boldsymbol{x})$: (i) $\hat{\boldsymbol{x}}_1 \in \mathcal{S}(\boldsymbol{x}, D_1)$, and (ii) $\rho_2^*(\hat{\boldsymbol{x}}_1) \le \tau(W^\ell)$.

(i) holds directly, by the type-determined-distortion property: $\mathcal{T}(W_1^\ell|\boldsymbol{x}) \subseteq \mathcal{S}(\boldsymbol{x}, D_1)$.

For (ii): by Lemma 5, every such $\hat{\boldsymbol{x}}_1$ satisfies

$$V[\mathcal{T}(W^\ell|\boldsymbol{x}, \hat{\boldsymbol{x}}_1)|\hat{\boldsymbol{x}}_1] \;\ge\; 2^{-n[R^*(W^\ell)-R_1^*(W^\ell)]-n\Delta_n'(\ell)+o(n)}. \tag{28}$$

Since $\mathcal{T}(W^\ell|\boldsymbol{x}, \hat{\boldsymbol{x}}_1) \subseteq \mathcal{S}(\boldsymbol{x}, D_2)$ (the same type-determined-distortion property, now applied to W^ℓ's own $(X, \hat{X}_2)$-marginal), $V[\mathcal{S}(\boldsymbol{x}, D_2)|\hat{\boldsymbol{x}}_1]$ is at least as large:

$$V[\mathcal{S}(\boldsymbol{x}, D_2)|\hat{\boldsymbol{x}}_1] \ge V[\mathcal{T}(W^\ell|\boldsymbol{x}, \hat{\boldsymbol{x}}_1)|\hat{\boldsymbol{x}}_1] \ge 2^{-n[R^*(W^\ell)-R_1^*(W^\ell)]-n\Delta_n'(\ell)+o(n)}. \tag{29}$$

Taking $-\frac{1}{n}\log$ of both sides (reversing the inequality) gives exactly $\rho_2^*(\hat{\boldsymbol{x}}_1) \le R^*(W^\ell) - R_1^*(W^\ell) + \Delta_n'(\ell) + o(1) = \tau(W^\ell) + o(1)$.

Both (i) and (ii) hold for *every* $\hat{\boldsymbol{x}}_1 \in \mathcal{T}(W_1^\ell|\boldsymbol{x})$ simultaneously — a uniform bound, not a majority statement, since Lemma 5 itself already holds uniformly. So $\mathcal{T}(W_1^\ell|\boldsymbol{x}) \subseteq \mathcal{G}_{\tau(W^\ell)+o(1)}$; absorbing this vanishing term into $\tau(W^\ell)$'s own definition is harmless (monotonicity of $\mathcal{G}_\tau$ in τ), so $\mathcal{T}(W_1^\ell|\boldsymbol{x}) \subseteq \mathcal{G}_{\tau(W^\ell)}$ exactly, as used below.

Step 2: a lower bound on $U[\mathcal{G}_{\tau(W^\ell)}]$. By Lemma 3 applied to $\mathcal{T}(W_1^\ell|\boldsymbol{x}) \subseteq \mathcal{T}_n(Q_1^\ell)$, $U[\mathcal{T}(W_1^\ell|\boldsymbol{x})] \ge U_{Q_1^\ell}[\mathcal{T}(W_1^\ell|\boldsymbol{x})] \cdot 2^{-n\Delta_n(\ell)} = 2^{-nR_1^*(W^\ell)-n\Delta_n(\ell)}$, using $R_1^*(W^\ell)$'s own definition (Theorem 1). By Step 1 and monotonicity of $U[\cdot]$, $U[\mathcal{G}_{\tau(W^\ell)}] \ge 2^{-nR_1^*(W^\ell)-n\Delta_n(\ell)}$.

Step 3: convert to a rate bound. Proposition 2, run at $\tau = \tau(W^\ell)$, gives $\rho_1(\boldsymbol{x}) \le -\frac{1}{n}\log U[\mathcal{G}_{\tau(W^\ell)}] + o(1) \le R_1^*(W^\ell) + \Delta_n(\ell) + o(1)$, and $\rho_1(\boldsymbol{x}) + \rho_2(\boldsymbol{x}) \le -\frac{1}{n}\log U[\mathcal{G}_{\tau(W^\ell)}] + \tau(W^\ell) + o(1) \le R_1^*(W^\ell) + \Delta_n(\ell) + [R^*(W^\ell) - R_1^*(W^\ell) + \Delta_n'(\ell)] + o(1) = R^*(W^\ell) + \Delta_n(\ell) + \Delta_n'(\ell) + o(1)$, as claimed. ∎

Theorem 5, combined with Theorem 2 (Section 3, applying here exactly as it does to any code), gives exact equality: the type-free scheme's own achievable region — the union, over $\tau \ge 0$, of the quadrant above each achieved $(\rho_1(\tau), \rho_1(\tau) + \rho_2(\tau))$ — coincides with $\mathcal{R}(\boldsymbol{x})$, not merely touches its corners from inside. Neither direction needs to know, or control, which type $\mathcal{G}_\tau$'s own random search actually realizes.

Search complexity: a real trade-off, not a wash. This equality is about achieved rates, not about what it costs to search for them, and the two constructions are not equivalent on that second axis. Theorem 4's own, type-committed Stage-1 criterion — joint typicality of $(\boldsymbol{x}, \hat{\boldsymbol{x}}_1)$ with a type W_1^ℓ fixed in advance — is checkable from $\boldsymbol{x}$ and a single candidate $\hat{\boldsymbol{x}}_1$ alone, with no reference to Stage 2's own search space at all; Stage 2 begins its own, separate search only once Stage 1 has already succeeded, and only once. Proposition 2's type-free criterion, $\mathcal{G}_\tau$-membership, is not so cheap to check: verifying it for a single candidate $\hat{\boldsymbol{x}}_1$ requires evaluating $\rho_2^*(\hat{\boldsymbol{x}}_1)$, itself the V-measure of the entire set $\mathcal{S}(\boldsymbol{x}, D_2)$ conditioned on that one $\hat{\boldsymbol{x}}_1$ — a computation reaching into Stage 2's own candidate space, repeated for every $\hat{\boldsymbol{x}}_1$ Stage 1 tries along the way, not merely the one it eventually keeps. What the type-free construction buys for this added cost is real, not merely conceptual: it needs no target type W^ℓ fixed in advance, and so no prior solution of the type-side optimization Section 3.4's own ϵ-constraint recipe requires just to decide which W^ℓ to aim for — reaching every point of $\mathcal{R}(\boldsymbol{x})$ off a single scalar dial τ instead; and, as the abstract states at the outset, no joint type of ℓ-vectors is ever constructed or checked by its own search criterion, only individual-sequence quantities throughout.

4.4 Domination over the finite-state-encoder route of [3]

Section 1.3 promised the two-stage extension of [1]'s own single-stage domination inequality (its eq. (51)); this subsection proves it, now that $\mathcal{G}_\tau$, U, V, and Proposition 2 are in place. Recall $\mathcal{S}_2(\boldsymbol{x}, D_1, D_2) = \mathcal{S}(\boldsymbol{x}, D_1) \times \mathcal{S}(\boldsymbol{x}, D_2)$ (Section 2) — the admissible set exactly as [3] defines it, since d_1, d_2 each depend on only their own reconstruction, so the two constraints are independent.

The finite-state-encoder route's own achievable region. For a genuine q-state finite-state encoder — q bounded, not growing with n — [3]'s own scheme partitions each reconstruction into non-overlapping blocks and restarts the LZ78 mechanism independently within each block, trading some rate for a bounded state count (unrestricted LZ78, run continuously over the whole sequence, needs a number of states not small compared to n). Write $\mathcal{R}_q(\boldsymbol{x})$ for the resulting true q-state achievable region. Theorem 1 of [3] shows this is contained in the region obtained from unrestricted LZ78 applied once, continuously, over the whole sequence:

$$\mathcal{R}_q(\boldsymbol{x}) \;\subseteq\; \mathcal{R}^o(\boldsymbol{x}) \triangleq \bigcup_{(\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2) \in \mathcal{S}_2(\boldsymbol{x}, D_1, D_2)} \Big\{ (R_1, R_2) : R_1 \geq LZ(\hat{\boldsymbol{x}}_1)/n - \Delta_1(q, n),$$

$$R_1 + R_2 \geq [LZ(\hat{\boldsymbol{x}}_1) + LZ(\hat{\boldsymbol{x}}_2|\hat{\boldsymbol{x}}_1)]/n - \Delta_2(q,n)\Big\}, \tag{30}$$

with $\Delta_1(q,n), \Delta_2(q,n) \to 0$ as $q, n \to \infty$ (its eqs. (21)–(24)) — the block-restart mechanism only shrinks the achievable region further, it never grows it beyond $\mathcal{R}^o(\boldsymbol{x})$.

In plain terms, this is what the next theorem shows: any rate pair [3]'s own scheme achieves, via some particular reconstruction $(\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2)$, is achieved by this paper's type-free construction too, at least as well — simply by adjusting τ to match whichever $\hat{\boldsymbol{x}}_1$ [3] would have used. [3]'s own reconstruction is never reproduced, only the threshold it implies.

Theorem 6 (Two-stage domination over the finite-state-encoder route) *For any admissible reconstruction pair $(\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2) \in \mathcal{S}_2(\boldsymbol{x}, D_1, D_2)$ — an arbitrary comparison target, entering only through the threshold below, never as a search criterion — setting $\tau = \rho_2^*(\hat{\boldsymbol{x}}_1)$ in Proposition 2's construction achieves*

$$\rho_1(\boldsymbol{x}) \;\leq\; \frac{1}{n} LZ(\hat{\boldsymbol{x}}_1) + o(1), \qquad \rho_1(\boldsymbol{x}) + \rho_2(\boldsymbol{x}) \;\leq\; \frac{1}{n}[LZ(\hat{\boldsymbol{x}}_1) + LZ(\hat{\boldsymbol{x}}_2|\hat{\boldsymbol{x}}_1)] + o(1). \tag{31}$$

Proof: Proposition 2's own construction searches for $\mathcal{G}_\tau$-membership without ever examining $\hat{\boldsymbol{x}}_1$ itself — only the threshold τ depends on it. Since $\hat{\boldsymbol{x}}_1 \in \mathcal{S}(\boldsymbol{x}, D_1)$ and $\rho_2^*(\hat{\boldsymbol{x}}_1) \leq \tau$ (equality, by choice of τ), $\hat{\boldsymbol{x}}_1 \in \mathcal{G}_\tau$, so $U[\mathcal{G}_\tau] \geq U(\hat{\boldsymbol{x}}_1) \geq 2^{-LZ(\hat{\boldsymbol{x}}_1)}$ (Kraft's inequality gives U's own normalizing sum $Z \leq 1$, and $U(\hat{\boldsymbol{x}}_1) = 2^{-LZ(\hat{\boldsymbol{x}}_1)}/Z \geq 2^{-LZ(\hat{\boldsymbol{x}}_1)}$). By Proposition 2, $\rho_1(\boldsymbol{x}) \leq -\frac{1}{n}\log U[\mathcal{G}_\tau] + o(1) \leq \frac{1}{n} LZ(\hat{\boldsymbol{x}}_1) + o(1)$. For the sum rate, the same Kraft argument applied to the conditional measure $V(\cdot|\hat{\boldsymbol{x}}_1)$ — itself a uniquely-decodable code given the known side information $\hat{\boldsymbol{x}}_1$ — gives $V[\mathcal{S}(\boldsymbol{x}, D_2)|\hat{\boldsymbol{x}}_1] \geq V(\hat{\boldsymbol{x}}_2|\hat{\boldsymbol{x}}_1) \geq 2^{-LZ(\hat{\boldsymbol{x}}_2|\hat{\boldsymbol{x}}_1)}$, so $\rho_2^*(\hat{\boldsymbol{x}}_1) \leq \frac{1}{n} LZ(\hat{\boldsymbol{x}}_2|\hat{\boldsymbol{x}}_1)$; combined with Proposition 2's $\rho_1(\boldsymbol{x}) + \rho_2(\boldsymbol{x}) \leq -\frac{1}{n}\log U[\mathcal{G}_\tau] + \tau + o(1) \leq \frac{1}{n} LZ(\hat{\boldsymbol{x}}_1) + \rho_2^*(\hat{\boldsymbol{x}}_1) + o(1)$, this gives the claim. ■

This is exactly [1]'s own single-stage inequality (its eq. (51), $\min_{\hat{\boldsymbol{x}} \in \mathcal{S}(\boldsymbol{x}, D)} LZ(\hat{\boldsymbol{x}}) \geq -\log U[\mathcal{S}(\boldsymbol{x}, D)]$), applied once unconditionally and once more conditionally on whichever $\hat{\boldsymbol{x}}_1$ is in play — no chain rule for LZ complexity is needed, since each inequality is proved separately, mechanically, from Kraft alone. Since Theorem 6 holds for *every* admissible pair, not merely the best one, this paper's achievable region contains, coordinate by coordinate, every rate pair in $\mathcal{R}^o(\boldsymbol{x})$ — and since $\mathcal{R}_q(\boldsymbol{x}) \subseteq \mathcal{R}^o(\boldsymbol{x})$, it contains [3]'s own true achievable region

too, up to [3]'s own vanishing ($q, n \to \infty$) slack, without needing any separate analysis of the block-restart mechanism: this paper's region is a superset of [3]'s, generally a strict one. This proof leans on the fact that Ziv's conditional LZ78 construction, built on $\hat{\boldsymbol{x}}_1$ as side information available to both encoder and decoder, is itself a uniquely-decodable code satisfying its own Kraft inequality for each fixed $\hat{\boldsymbol{x}}_1$ — confirmed directly by Ziv's own construction [26], which builds an explicit, uniquely-decodable conditional LZ78 coding scheme achieving exactly this length function; standard, and implicit throughout [3]'s own use of the same construction.

5 Extension to $r > 2$ stages

The two-stage construction of Sections 3–4 generalizes to any fixed number $r \geq 2$ of stages along lines that are natural given the two-stage case. What follows is an informal description of this extension, not a formal proof: as emphasized throughout, the two-stage case is where this paper's actual difficulty lives, and the picture below is offered as the natural next step, with no claim to the same rigor.

The type-covering skeleton. Fix an r-way joint type W^ℓ of $(X, \hat{X}_1, \ldots, \hat{X}_r)$ meeting all r distortion budgets $D_1, \ldots, D_r$. Exactly as in the two-stage case (itself a direct translation of [23]'s classical characterization to ℓ-blocks), the type-conditional bound (Theorem 1's own Kraft-counting argument, run once per prefix $k = 1, \ldots, r$, over the k-tupled alphabet $\hat{\mathcal{X}}_1 \times \cdots \times \hat{\mathcal{X}}_k$, each stage's own, not assumed equal) and the achievability construction (Theorem 3's own Stage-1/Stage-2 mechanism, run once per stage, each conditioning on the exact history realized so far) generalize with no new idea: at stage k, search for the first candidate $\hat{\boldsymbol{x}}_k$ completing the full joint type W^ℓ restricted to $(\boldsymbol{x}, \hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_k)$, given the exact $(\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_{k-1})$ already fixed by the previous $k-1$ stages. This gives $\rho_1(\boldsymbol{x}) + \ldots + \rho_k(\boldsymbol{x}) \to R_k^*(W^\ell)$ for every k simultaneously, $R_k^*(W^\ell)$ the natural generalization of R_1^*, R^* above to the cumulative rate through stage k.

The LZ-based realization. At stage k, candidates for $\hat{\boldsymbol{x}}_k$ would be drawn i.i.d. from the measure proportional to $2^{-LZ(\hat{\boldsymbol{x}}_k | \hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_{k-1})}$ — the conditional LZ78 codelength of $\hat{\boldsymbol{x}}_k$ given all $k-1$ previously realized reconstructions as side information, generalizing U (stage 1, $k = 1$, no prior history) and $V(\cdot|\hat{\boldsymbol{x}}_1)$ (stage 2, conditional on the single prior reconstruction) to conditioning on the full history. The same principle that resolves the two-stage Jensen obstruction (Section 3) would

need to apply at every stage: $\hat{\boldsymbol{x}}_k$'s search criterion must guarantee, not merely make likely, that $\hat{\boldsymbol{x}}_k$ keeps stage $k+1$'s own rate low — generalizing Stage 1's own joint-typicality criterion from looking one stage ahead to looking ahead across however many stages remain, by requiring joint typicality of the full history $(\boldsymbol{x}, \hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_k)$ rather than of $(\boldsymbol{x}, \hat{\boldsymbol{x}}_1)$ alone.

The type-free realization: full lookahead by backward recursion. Sections 4.2–4.3's type-free scheme also generalizes, and settles the lookahead-depth question just raised: one-step lookahead (stage k guaranteeing only stage $k+1$'s affordability) does *not* suffice, because stage $k+1$ itself must guarantee stage $k+2$'s affordability, so the quantity stage k needs to bound is stage $k+1$'s own *restricted* rate, not its naive, unrestricted one — a dependency running the wrong way for one-step-at-a-time chaining. Full lookahead across the entire remaining chain is genuinely needed, but it is a single *backward* recursion, computed once as design-time bookkeeping, not a search the running encoder performs. Write $\hat{\boldsymbol{x}}^k \triangleq (\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_k)$ for a history of length k ($\hat{\boldsymbol{x}}^0$ empty), and $U_k(\cdot|\hat{\boldsymbol{x}}^{k-1})$ for the measure proportional to $2^{-LZ(\cdot|\hat{\boldsymbol{x}}_1,\ldots,\hat{\boldsymbol{x}}_{k-1})}$ just described ($U_1 \triangleq U$, no history; $U_2(\cdot|\hat{\boldsymbol{x}}_1) \triangleq V(\cdot|\hat{\boldsymbol{x}}_1)$, recovering Stage 2's own measure). Concretely: define $\rho_r^*(\hat{\boldsymbol{x}}^{r-1}) \triangleq -\frac{1}{n}\log U_r[\mathcal{S}(\boldsymbol{x}, D_r)|\hat{\boldsymbol{x}}^{r-1}]$ (unrestricted — nothing follows stage r), and recursively, for $k = r-1, \ldots, 1$, given $\tau_{k+1}, \ldots, \tau_{r-1}$ already fixed,

$$\mathcal{G}_{\tau_k}^{(k)}(\hat{\boldsymbol{x}}^{k-1}) \triangleq \mathcal{S}(\boldsymbol{x}, D_k) \cap \{\hat{\boldsymbol{x}}_k : \rho_{k+1}^{*\rightarrow}(\hat{\boldsymbol{x}}^{k-1}, \hat{\boldsymbol{x}}_k) \le \tau_k\}, \qquad \rho_{k+1}^{*\rightarrow}(\hat{\boldsymbol{x}}^k) \triangleq -\frac{1}{n}\log U_{k+1}[\mathcal{G}_{\tau_{k+1}}^{(k+1)}(\hat{\boldsymbol{x}}^k)|\hat{\boldsymbol{x}}^k], \tag{32}$$

generalizing $\mathcal{G}_\tau$ ($k = 1$ case, once τ_1 is fixed as below) to every stage. Since $\hat{\boldsymbol{x}}_1 \in \mathcal{G}_{\tau_1}^{(1)}$ or $\hat{\boldsymbol{x}}_k \in \mathcal{G}_{\tau_k}^{(k)}(\hat{\boldsymbol{x}}^{k-1})$ are, by this definition, statements about the specific, already-realized $\hat{\boldsymbol{x}}_k$ — checked before acceptance, not averaged over how it arose — the guarantee never has to survive the randomness of how the previous stage's search turned out; this is the same mechanism that resolves the $r = 2$ Jensen obstruction, applied once per stage rather than once. A downward induction (from $k = r-1$ to $k = 1$) shows the entire type-match set $\mathcal{T}(W_{\le k}^\ell|\boldsymbol{x})$ is absorbed into $\mathcal{G}_{\tau_k(W^\ell)}^{(k)}$ once $\tau_k(W^\ell) \triangleq R_{k+1}^*(W^\ell) - R_k^*(W^\ell) + \Delta_n^{(k)}(\ell)$ — $\Delta_n^{(k)}(\ell) \to 0$ the natural, conditioning-depth-k generalization of $\Delta_n'(\ell)$ (itself the depth-1 case) — mirroring Lemma 5 at every conditioning depth; the resulting rates telescope exactly, $\rho_1(\boldsymbol{x}) + \ldots + \rho_k(\boldsymbol{x}) \to R_k^*(W^\ell)$ for every k simultaneously, matching Theorem 5's $r = 2$ result at every stage, not merely at the last one.

What remains. The lookahead-depth question raised above is settled by the construction itself —

full lookahead, via backward recursion, with no obstruction of a new character, the same mechanism as $r = 2$ repeated $r - 1$ times rather than found anew. One place this construction might look, at first, like it leans on something beyond what the two-stage case already uses: Stage k's search needs Lemma 4's codelength bound with the side information taken jointly from $k-1$ prior reconstructions, not the single $\hat{\boldsymbol{x}}_1$ the lemma is stated for. But conditioning on a fixed-length tuple $(\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_{k-1})$ is conditioning on *one* sequence over the product alphabet $\hat{X}_1 \times \ldots \times \hat{X}_{k-1}$ — finite, since each factor is finite and $k - 1 \leq r - 1$ is bounded independent of n — and nothing in Lemma 4's own derivation uses any structural property of the conditioning alphabet beyond its being finite. So this is the same lemma, applied once with the conditioning alphabet relabeled, not a separate instance needed at each depth — no additional assumption, at any k, beyond the one the $r = 2$ case already carries.

6 Summary and Conclusion

This paper extended the single-stage, LZ-based random-coding approach of [1] to two-stage successive refinement of individual sequences. Section 3 established a complete direct/converse match in the language of type classes — a two-stage analogue of Rimoldi's classical characterization — culminating in a converse (Theorem 2) that binds every code, not merely type-committed ones. Section 4 realized the same rates via an LZ78-based random-coding scheme, universal across source statistics and distortion measures, with no block-length parameter built into the codebook itself. Sections 4.2–4.3 then gave a fully type-free construction, reaching every point of the achievable region without ever committing to a type — matching the same converse exactly, though at a real search-cost premium over the type-committed route.

This type-free flexibility turns out not to enlarge the achievable rate region itself: the type-free and type-committed constructions reach exactly the same points. The gain is elsewhere. Section 4.4 showed the type-free construction dominates the finite-state-encoder approach of [3]: any rate pair that scheme achieves is achieved by this paper's own construction too, simply by adjusting the search threshold accordingly — extending [1]'s own single-stage domination result to two stages. Section 5 extended the type-free construction to a general, fixed number $r > 2$ of stages, via a single backward recursion, with no obstruction of a new character beyond what the two-stage case already

resolves.

A Proof of the universal random-coding achievability lemma

This proof is closely modeled on [1]'s own achievability construction (its Theorem 2).

Proof:[Proof of Lemma 1 (Universal random-coding achievability)] This follows the same construction as [1]'s own achievability theorem (its Theorem 2), generalized in two respects: the target, from [1]'s own specific family of distortion balls $\mathcal{S}(\boldsymbol{x}, D)$ to an arbitrary target family $\mathcal{E}(\boldsymbol{x})$; and the measure, from the specific U to an arbitrary μ meeting the stated hypothesis ($\mu(\hat{\boldsymbol{x}}) \geq B^{-n}$ for every $\hat{\boldsymbol{x}}$ with $\mu(\hat{\boldsymbol{x}}) > 0$). Two further differences below are purely mechanical: the expectation bound uses the untruncated geometric mean directly, where [1] carries the A^n truncation through an explicit sum to the same conclusion; and the worst-case LZ bound needed for search failure is drawn from Lemma 3, already established earlier in this paper, rather than re-derived from the counting-sequence construction [1] uses for the same purpose.

Index encoding. Given the codebook $\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_{A^n}$ (drawn once, i.i.d. $\sim \mu$, then revealed to encoder and decoder alike) and a source $\boldsymbol{x}$ with target $\mathcal{E}(\boldsymbol{x})$, let $I(\boldsymbol{x})$ be the index of the first codeword landing in $\mathcal{E}(\boldsymbol{x})$, or $I(\boldsymbol{x}) = A^n$ if none does. Encode $I(\boldsymbol{x})$ using the length function $L(\boldsymbol{x}) = -\log u[I(\boldsymbol{x})]$ for $u[i] = (1/i)/\sum_{k=1}^{A^n}(1/k)$, $i = 1, \ldots, A^n$ — a genuine probability distribution on $\{1, \ldots, A^n\}$, so L is realizable by a valid prefix-free code. Since $\sum_{k=1}^{A^n} 1/k \leq \ln(A^n) + 1 = n \ln A + 1$,

$$L(\boldsymbol{x}) \leq \log I(\boldsymbol{x}) + \log(n \ln A + 1) \leq \log I(\boldsymbol{x}) + \log n + c, \qquad c \triangleq \log(\ln A + 1). \tag{A.1}$$

A single-target expectation bound. For fixed $\boldsymbol{x}$, if all codewords are drawn i.i.d. $\sim U$, then for any positive integer N, $\Pr[I(\boldsymbol{x}) > N] = (1 - U[\mathcal{E}(\boldsymbol{x})])^N \leq e^{-N \cdot U[\mathcal{E}(\boldsymbol{x})]}$. By Jensen's inequality (once), writing $p \triangleq U[\mathcal{E}(\boldsymbol{x})]$,

$$E[\log I(\boldsymbol{x})] \leq \log E[I(\boldsymbol{x})] = \log(1/p) = -\log(U[\mathcal{E}(\boldsymbol{x})]), \tag{A.2}$$

using the standard geometric-distribution fact $E[I] = 1/p$ (treating $I(\boldsymbol{x})$, before the truncation at A^n, as the index of the first success among i.i.d. Bernoulli(p) trials). Combining with the encoding bound above,

$$E[L(\boldsymbol{x})] \leq -\log(U[\mathcal{E}(\boldsymbol{x})]) + \log n + c \triangleq L^+(\boldsymbol{x}). \tag{A.3}$$

From one $\boldsymbol{x}$ in expectation to every $\boldsymbol{x}$ simultaneously. Define, exactly as in [1]'s own eq. (42) — with the excess threshold $(1+\epsilon)\log n$, not merely $\epsilon \log n$; this distinction matters and is explained below —

$$E_n \triangleq E\Big\{ \max\Big(\max_{\boldsymbol{x}\in\mathcal{X}^n} \mathbf{1}\{I(\boldsymbol{x}) = A^n \text{ fails}\},\ [\max_{\boldsymbol{x}\in\mathcal{X}^n} (L(\boldsymbol{x}) - L^+(\boldsymbol{x}) - (1+\epsilon)\log n)]_+ \Big)\Big\}, \tag{A.4}$$

the expectation over the random codebook. If $E_n \to 0$, then for some $N = N(\epsilon, A, B)$ and every $n > N$, some codebook realization makes both terms inside the max simultaneously below 1 for every $\boldsymbol{x}$ — the first term is 0 or 1-valued, so below 1 means the search succeeds for every $\boldsymbol{x}$, and the second gives $L(\boldsymbol{x}) \le L^+(\boldsymbol{x}) + (1+\epsilon)\log n = -\log(U[\mathcal{E}(\boldsymbol{x})]) + (2+\epsilon)\log n + c$ for every $\boldsymbol{x}$, which is the claimed bound. It remains to show $E_n \to 0$; since the max of two nonnegative terms is at most their sum, bound each separately.

Search failure, uniformly. By the union bound over $\boldsymbol{x}$ and $\Pr[I(\boldsymbol{x}) = A^n \text{ fails}] = (1-p)^{A^n} \le e^{-A^n p}$,

$$E\Big\{ \max_{\boldsymbol{x}} \mathbf{1}\{\text{failure}\}\Big\} \le \sum_{\boldsymbol{x}\in\mathcal{X}^n} e^{-A^n U[\mathcal{E}(\boldsymbol{x})]}. \tag{A.5}$$

Since $U[\mathcal{E}(\boldsymbol{x})] \ge U(\hat{\boldsymbol{x}}_0)$ for any single $\hat{\boldsymbol{x}}_0 \in \mathcal{E}(\boldsymbol{x})$ (assuming $\mathcal{E}(\boldsymbol{x}) \neq \emptyset$, else the achievability claim is vacuous), and U's normalization together with the LZ78 codelength bound (Lemma 3, $LZ(\hat{\boldsymbol{x}}) \le n(1+\epsilon_n)\log \hat{K}$ for the worst case, exactly as in [1]) gives $U[\mathcal{E}(\boldsymbol{x})] \ge 2^{-n(1+\epsilon_n)\log \hat{K}}$ uniformly, this sum is at most $|\mathcal{X}|^n \exp\{-2^{n[\log A - (1+\epsilon_n)\log \hat{K}]}\} \to 0$ doubly exponentially fast whenever $A > \hat{K}$, exactly as in [1]'s own Theorem 2.

Excess codelength, uniformly — why the threshold must be $(1+\epsilon)\log n$, not $\epsilon \log n$. As in [1],

$$E\Big\{[\max_{\boldsymbol{x}}(L(\boldsymbol{x}) - L^+(\boldsymbol{x}) - (1+\epsilon)\log n)]_+\Big\} \le \sum_{\boldsymbol{x}\in\mathcal{X}^n} \int_0^{n\log A} \Pr\Big[I(\boldsymbol{x}) \ge \frac{2^{(1+\epsilon)\log n + s}}{U[\mathcal{E}(\boldsymbol{x})]}\Big]\, \mathrm{d}s \tag{A.6}$$

$$\le \sum_{\boldsymbol{x}\in\mathcal{X}^n} \int_0^{n\log A} \exp\{-2^s n^{1+\epsilon}\}\, \mathrm{d}s \le |\mathcal{X}|^n (n\log A) e^{-n^{1+\epsilon}} \to 0, \tag{A.7}$$

using the geometric tail bound $\Pr[I(\boldsymbol{x}) \ge N] \le e^{-Np} \le \exp\{-2^{(1+\epsilon)\log n + s} \cdot U[\mathcal{E}(\boldsymbol{x})]\}$ with $N = 2^{(1+\epsilon)\log n + s}/U[\mathcal{E}(\boldsymbol{x})]$, evaluated at $s = 0$ for the final step. This is precisely why the threshold cannot merely be $\epsilon \log n$: that would give $|\mathcal{X}|^n \exp\{-n^\epsilon\}$ in place of $|\mathcal{X}|^n \exp\{-n^{1+\epsilon}\}$, and n^ϵ grows *slower* than $n\log|\mathcal{X}|$ for $\epsilon < 1$, so this weaker bound *diverges* rather than vanishing — the polynomial threshold $n^{1+\epsilon}$, not just n^ϵ, is what is needed to dominate the $|\mathcal{X}|^n$ union-bound factor. Both terms $\to 0$, so $E_n \to 0$, completing the proof — but only for $n > N(\epsilon)$: it is exactly this

term, $|\mathcal{X}|^n(n\log A)e^{-n^{1+\epsilon}}$, whose rate of decay sets $N(\epsilon)$, since smaller ϵ needs larger n before $n^{1+\epsilon}$ overtakes the union-bound factor $|\mathcal{X}|^n$, so $N(\epsilon)\to\infty$ as $\epsilon\to 0^+$; at $\epsilon=0$ exactly this same term becomes $|\mathcal{X}|^n(n\log A)e^{-n}$, which diverges rather than vanishes whenever $|\mathcal{X}|\geq 3$, and the argument breaks down outright. ■

B Proof of the Type-Conditional Rate Bound

The proof uses that any single reconstruction can be responsible for at most $|\mathcal{T}_n(P^\ell)\cap\mathcal{T}(W_1^\ell|\boldsymbol{z})|$ sources of type W_1^ℓ, itself already given by Lemma 2's type-match extension, to convert a bound on the fraction of *codewords* into one on the fraction of *source sequences* directly. This observation is used throughout the proof that follows.

Proof:[Proof of Theorem 1 (Type-conditional rate bound)] Fix the type W_1^ℓ; the argument for W^ℓ (total rate) is identical with W_1^ℓ replaced by W^ℓ, Q_1^ℓ by W^ℓ's own $(\hat{X}_1,\hat{X}_2)$-marginal, and the reconstruction alphabet $\hat{\mathcal{X}}_1$ by the doubled alphabet $\hat{\mathcal{X}}_1\times\hat{\mathcal{X}}_2$ (treating the pair $(\hat{\boldsymbol{x}}_1,\hat{\boldsymbol{x}}_2)$ as a single reconstruction), so we give the single argument and combine at the end.

A codeword-counting bound. Throughout, $\boldsymbol{z}$ denotes a generic candidate ranging over $\mathcal{T}_n(Q_1^\ell)$, kept distinct from $\hat{\boldsymbol{x}}_1$, reserved below for Φ's own, specific output on a given $\boldsymbol{x}$. By Lemma 2's type-match extension, writing $\mathcal{T}(W_1^\ell|\boldsymbol{z})\triangleq\{\boldsymbol{x}:\hat{P}_{\boldsymbol{x}\boldsymbol{z}}=W_1^\ell\}$ for the reverse-direction type-match set, $|\mathcal{T}_n(P^\ell)\cap\mathcal{T}(W_1^\ell|\boldsymbol{z})|=|\mathcal{T}_n(P^\ell)|\cdot U_{Q_1^\ell}[\mathcal{T}(W_1^\ell|\boldsymbol{x})]$ for every $\boldsymbol{z}\in\mathcal{T}_n(Q_1^\ell)$ and every $\boldsymbol{x}\in\mathcal{T}_n(P^\ell)$; write $u_1\triangleq U_{Q_1^\ell}[\mathcal{T}(W_1^\ell|\boldsymbol{x})]$ for this common value (a single number, the same for every choice of $\boldsymbol{z}$ and $\boldsymbol{x}$ in their respective type classes, by Lemma 2's own symmetry). Fix $\ell_0>0$, and let $\mathcal{Z}_{\ell_0}\triangleq\{\boldsymbol{z}:\boldsymbol{z}=\psi_1(\phi_1(\boldsymbol{x}))$ for some $\boldsymbol{x}$ with $\hat{P}_{\boldsymbol{x}\boldsymbol{z}}=W_1^\ell$ and $|\phi_1(\boldsymbol{x})|<\ell_0\}$ — $\boldsymbol{x}\in\mathcal{T}_n(P^\ell)$ and $\boldsymbol{z}\in\mathcal{T}_n(Q_1^\ell)$ are automatic, already implied by $\hat{P}_{\boldsymbol{x},\boldsymbol{z}}=W_1^\ell$ — the set of type-Q_1^ℓ reconstructions occurring as outputs of type-W_1^ℓ sources with codeword length below ℓ_0. The claim follows from four statements.

(i) $|\mathcal{Z}_{\ell_0}|<2^{\ell_0}$. The first-stage decoder ϕ_1 assigns each distinct reconstruction its own binary string, so $\mathcal{Z}_{\ell_0}$ injects into the binary strings of length $<\ell_0$, of which there are $\sum_{k<\ell_0}2^k=2^{\ell_0}-1$.

(ii) For every $\boldsymbol{z}\in\mathcal{T}_n(Q_1^\ell)$, $|\{\boldsymbol{x}\in\mathcal{T}_n(P^\ell):\hat{P}_{\boldsymbol{x}\boldsymbol{z}}=W_1^\ell,\ \psi_1(\phi_1(\boldsymbol{x}))=\boldsymbol{z}\}|\leq|\mathcal{T}_n(P^\ell)|\cdot u_1$ *($u_1\triangleq U_{Q_1^\ell}[\mathcal{T}(W_1^\ell|\boldsymbol{x})]$, as above).* Every such $\boldsymbol{x}$ satisfies $\hat{P}_{\boldsymbol{x}\boldsymbol{z}}=W_1^\ell$ by definition of the set being

counted, so this set is contained in $\mathcal{T}_n(P^\ell) \cap \mathcal{T}(W_1^\ell|\boldsymbol{z})$, of size $|\mathcal{T}_n(P^\ell)| \cdot u_1$ by the opening display.

(iii) The sets $\{\boldsymbol{x} \in \mathcal{T}_n(P^\ell) : \psi_1(\phi_1(\boldsymbol{x})) = \boldsymbol{z}\}$, $\boldsymbol{z} \in \mathcal{Z}_{\ell_0}$, are pairwise disjoint. $\psi_1 \circ \phi_1$ is a function of $\boldsymbol{x}$, so each $\boldsymbol{x}$ belongs to exactly one such set.

(iv) Combining (i)–(iii):

$$\begin{aligned} |\{\boldsymbol{x} \in \mathcal{T}_n(P^\ell) : (\boldsymbol{x}, \hat{\boldsymbol{x}}_1) \text{ has type } W_1^\ell,\ L_1(\boldsymbol{x}) < \ell_0\}| &= \sum_{\boldsymbol{z} \in \mathcal{Z}_{\ell_0}} |\{\boldsymbol{x} : \hat{P}_{\boldsymbol{x}\boldsymbol{z}} = W_1^\ell,\ \psi_1(\phi_1(\boldsymbol{x})) = \boldsymbol{z}\}| \\ &\le |\mathcal{Z}_{\ell_0}| \cdot |\mathcal{T}_n(P^\ell)| \cdot u_1 \ < \ 2^{\ell_0} |\mathcal{T}_n(P^\ell)| u_1 \end{aligned} \tag{B.1}$$

the first equality since a union of pairwise disjoint sets (iii) has size equal to the sum of their sizes, the first inequality by (ii) applied to each term, and the last by (i).

Choosing $\ell_0 = \log_2(1/u_1) - \epsilon \log_2 n$, so $2^{\ell_0} = n^{-\epsilon}/u_1$, this bound becomes exactly $n^{-\epsilon}|\mathcal{T}_n(P^\ell)|$: at most an $n^{-\epsilon}$ fraction of $\boldsymbol{x} \in \mathcal{T}_n(P^\ell)$ can have output type W_1^ℓ and $L_1(\boldsymbol{x}) < -\log_2(u_1) - \epsilon \log_2 n = nR_1^*(W^\ell) - \epsilon \log n$. Equivalently, for all but at most an $n^{-\epsilon}$ fraction of $\boldsymbol{x} \in \mathcal{T}_n(P^\ell)$, every code whose output on this $\boldsymbol{x}$ has type W_1^ℓ satisfies the first bound.

The identical argument on the doubled alphabet (reconstruction $(\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2)$, type W^ℓ, target $\mathcal{T}(W^\ell|\boldsymbol{x})$, total codeword length $L_1 + L_2$, since (ϕ_1, ϕ_2) jointly determine one binary string of this length from which $(\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2)$ is recovered) excludes a second set of at most an $n^{-\epsilon}$ fraction of $\boldsymbol{x} \in \mathcal{T}_n(P^\ell)$. By the union bound, all but at most a $2n^{-\epsilon}$ fraction of $\boldsymbol{x} \in \mathcal{T}_n(P^\ell)$ avoid both exceptional sets, giving both bounds simultaneously whenever the code's output on such an $\boldsymbol{x}$ has joint type W^ℓ (which forces its own $(X, \hat{X}_1)$ sub-type to be W_1^ℓ, so the first bound is exactly the marginal case of the second). ∎

References

[1] N. Merhav, "A Universal Random Coding Ensemble for Sample-Wise Lossy Compression," *Entropy*, vol. 25, no. 8, art. 1199, 2023.

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

[3] N. Merhav, "Successive Refinement for Lossy Compression of Individual Sequences," *Entropy*, vol. 27, no. 4, art. 370, 2025.

[4] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*, Prentice-Hall, Englewood Cliffs, NJ, 1971.

[5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed., Wiley, Hoboken, NJ, 2006.

[6] R. G. Gallager, *Information Theory and Reliable Communication*, Wiley, New York, 1968.

[7] Z. Zhang, E.-h. Yang, and V. Wei, "The redundancy of source coding with a fidelity criterion. I. known statistics," *IEEE Trans. Inform. Theory*, vol. 43, no. 1, pp. 71–91, Jan. 1997.

[8] B. Yu and T. Speed, "A rate of convergence result for a universal D-semifaithful code," *IEEE Trans. Inform. Theory*, vol. 39, no. 3, pp. 813–820, May 1993.

[9] D. S. Ornstein and P. C. Shields, "Universal almost sure data compression," *Ann. Probab.*, vol. 18, no. 2, pp. 441–452, 1990.

[10] I. Kontoyiannis, "Pointwise redundancy in lossy data compression and universal lossy data compression," *IEEE Trans. Inform. Theory*, vol. 46, no. 1, pp. 136–152, Jan. 2000.

[11] I. Kontoyiannis and J. Zhang, "Arbitrary source models and Bayesian codebooks in rate-distortion theory," *IEEE Trans. Inform. Theory*, vol. 48, no. 8, pp. 2276–2290, Aug. 2002.

[12] A. Mahmood and A. B. Wagner, "Lossy compression with universal distortion," arXiv:2110.07022, 2022.

[13] A. Mahmood and A. B. Wagner, "Minimax rate-distortion," arXiv:2202.04481, 2022.

[14] J. Ziv, "Coding theorems for individual sequences," *IEEE Trans. Inform. Theory*, vol. IT-24, no. 4, pp. 405–412, Jul. 1978.

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

[16] J. Ziv, "Fixed-rate encoding of individual sequences with side information," *IEEE Trans. Inform. Theory*, vol. IT-30, no. 2, pp. 348–352, Mar. 1984.

[17] N. Merhav and A. Cohen, "Universal randomized guessing with application to asynchronous decentralized brute-force attacks," *IEEE Trans. Inform. Theory*, vol. 66, no. 1, pp. 114–129, Jan. 2020.

[18] N. Merhav, "Guessing individual sequences: generating randomized guesses using finite-state machines," *IEEE Trans. Inform. Theory*, vol. 66, no. 5, pp. 2912–2920, May 2020.

[19] E. Plotnik, M. J. Weinberger, and J. Ziv, "Upper bounds on the probability of sequences emitted by finite-state sources and on the redundancy of the Lempel–Ziv algorithm," *IEEE Trans. Inform. Theory*, vol. 38, no. 1, pp. 66–72, Jan. 1992.

[20] W. H. R. Equitz and T. M. Cover, "Successive refinement of information," *IEEE Trans. Inform. Theory*, vol. 37, no. 2, pp. 269–275, Mar. 1991.

[21] A. El Gamal and T. M. Cover, "Achievable rates for multiple descriptions," *IEEE Trans. Inform. Theory*, vol. IT-28, no. 6, pp. 851–857, Nov. 1982.

[22] V. N. Koshelev, "On the divisibility of discrete sources with an additive single-letter distortion measure," *Problems of Information Transmission*, vol. 30, no. 1, pp. 27–43, 1994.

[23] B. Rimoldi, "Successive refinement of information: characterization of the achievable rates," *IEEE Trans. Inform. Theory*, vol. 40, no. 1, pp. 253–259, Jan. 1994.

[24] A. Kanlis and P. Narayan, "Error exponents for successive refinement by partitioning," *IEEE Trans. Inform. Theory*, vol. 42, no. 1, pp. 275–282, Jan. 1996.

[25] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed., Cambridge University Press, 2011.

[26] J. Ziv, "Universal decoding for finite-state channels," *IEEE Trans. Inform. Theory*, vol. IT-31, no. 4, pp. 453–460, Jul. 1985.

[27] 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, Oct. 2003.

[28] N. Merhav, "Universal detection of messages via finite-state channels," *IEEE Trans. Inform. Theory*, vol. 46, no. 6, pp. 2242–2246, Sep. 2000.