Subset–Universal Lossy Compression

Or Ordentlich
Tel Aviv University
ordent@eng.tau.ac.il

Ofer Shayevitz
Tel Aviv University
ofersha@eng.tau.ac.il

Abstract—A lossy source code $\mathcal{C}$ with rate R for a discrete memoryless source S is called subset–universal if for every $0 < R' < R$, almost every subset of $2^{nR'}$ of its codewords achieves average distortion close to the source's distortion-rate function $D(R')$. In this paper we prove the asymptotic existence of such codes. Moreover, we show the asymptotic existence of a code that is subset–universal with respect to all sources with the same alphabet.

I. INTRODUCTION

To motivate the topic studied in this work, let us consider the following scenario. Let S be a discrete source with a probability mass function (pmf) P_S over an alphabet $\mathcal{S}$. Assume one wants to convey n i.i.d. instances of S over a deterministic channel $f : \mathcal{X}^n \mapsto \mathcal{Y}^n$, with the smallest possible distortion. If the channel law $f(\cdot)$ is known at both ends, the channel corresponds to a bit pipe whose capacity is the normalized logarithm of the function's image size, i.e., $C = \frac{1}{n} \log |f(\mathcal{X}^n)|$. In this case a separation approach, where the source is first compressed with rate C and then the compression index is transmitted over the channel, is optimal and achieves the best possible average distortion $D_{P_S}(C)$, where $D_{P_S}(\cdot)$ is the distortion-rate function of the source P_S.

When only the encoder knows the function $f(\cdot)$ but the decoder does not, it may be possible to "learn" the channel law if the class of possible functions is sufficiently small. For instance, if $f(\cdot)$ tensorizes, i.e., if $f(\mathbf{x}^n) = h(x_1), \ldots, h(x_n)$, it is possible to learn $f(\cdot)$ with negligible cost for n large enough. In this case, the separation based approach continues to be asymptotically optimal. Moreover, sometimes even partial knowledge of $f(\cdot)$ at the decoder is sufficient for separation to be optimal. For example, if the "channel" is a binary memory device with $\approx pn$ cells stuck at either 0 or 1, and whose locations are known only to the encoder, a reliable communication rate approaching $\frac{1}{n} \log |f(\mathcal{X}^n)|$ can be achieved via Gelfand-Pinsker coding [1], [2], and subsequently separation can asymptotically attain the optimal distortion.

A separation approach, however, is completely useless when $f(\cdot)$ can be any *arbitrary* function known only to the encoder. This follows from the fact that the capacity of the compound channel $\mathcal{F}$ that consists of all $|\mathcal{Y}|^{n|\mathcal{X}|^n}$ possible functions from

$\mathcal{X}^n$ to $\mathcal{Y}^n$ is zero. It is therefore tempting to jump to the conclusion that the smallest average distortion that can be attained in this setting is $D_{P_S}(0)$. But is this indeed the case? We will show that the answer is negative, and that in fact there exists a coding scheme that achieves average distortion $D_{P_S}(\frac{1}{n} \log |f(\mathcal{X}^n)|)$ for almost every deterministic channel.

Note that the encoder, which knows the function $f(\cdot)$, can deterministically impose any output from the set $f(\mathcal{X}^n) \subseteq \mathcal{Y}^n$. However, due to its ignorance, the decoder does not know how to translate the possible outputs to a vector of $\log |f(\mathcal{X}^n)|$ bits. Instead, the encoder and decoder can agree in advance on a mapping $g : \mathcal{Y}^n \mapsto \hat{\mathcal{S}}^n$ from each possible channel output to a source reconstruction sequence. The set $g(\mathcal{Y}^n)$ of all possible reconstructions can be thought of as a source code $\mathcal{C}$ of rate $R = \log |\mathcal{Y}|$ bits per source symbol. The channel's effect is in diluting $\mathcal{C}$ to the source code $\mathcal{C}' = g(f(\mathcal{X}^n)) \subseteq \mathcal{C}$ whose rate is $R' = \frac{1}{n} \log |f(\mathcal{X}^n)|$ bits per symbol. In other words, a deterministic channel $f(\cdot)$ with capacity R' chooses a subset of $2^{nR'}$ codewords from the 2^{nR} codewords in $\mathcal{C}$.

This motivates the study of *subset–universal* source codes. A subset–universal source code of rate R has the property that for any $0 < R' < R$, almost every subset of $2^{nR'}$ of its codewords is close to being optimal, in the sense that its induced average distortion is close to the distortion-rate function $D_{P_S}(R')$. Our main result is the asymptotic existence of such codes. In fact, we show the asymptotic existence of a code that is subset–universal with respect to any i.i.d. source distribution over the alphabet $\mathcal{S}$.

Returning to our discussion on joint-source channel coding over a deterministic channel known only at the encoder, the existence of a subset-universal code implies that for almost every choice of $f(\cdot)$, an average distortion of $D_{P_S}(\frac{1}{n} \log |f(\mathcal{X}^n)|)$ can be asymptotically achieved.

II. MAIN RESULT

The entropy of a random variable Y with probability mass function (pmf) P_Y over the alphabet $\mathcal{Y}$, is defined as

$$H(Y) \triangleq - \sum_{y \in \mathcal{Y}} P_Y(y) \log P_Y(y).$$

For a pair of random variables (Y, Z) with a joint pmf $P_{YZ} = P_Y P_{Z|Y}$ over the product alphabet $\mathcal{Y} \times \mathcal{Z}$, the conditional entropy is defined as

$$H(Z|Y) \triangleq - \sum_{(y,z) \in (\mathcal{Y} \times \mathcal{Z})} P_{YZ}(y,z) \log P_{Z|Y}(z|y).$$

and the mutual information is defined as

$$I(Y;Z) \triangleq H(Y) - H(Y|Z) = H(Z) - H(Z|Y).$$

For two distributions P and Q defined on the same alphabet $\mathcal{Z}$, the KL-divergence is defined as

$$D(P||Q) \triangleq \sum_{z \in \mathcal{Z}} P(z) \log \frac{P(z)}{Q(z)}.$$

Let S be a discrete memoryless source (DMS) over the alphabet $\mathcal{S}$ with pmf P_S, $\hat{\mathcal{S}}$ a reconstruction alphabet, and $d : \mathcal{S} \times \hat{\mathcal{S}} \mapsto \mathbb{R}^+$ a bounded single-letter distortion measure. A $(2^{nR}, n)$ lossy source code consists of an encoder that assigns an index $m(\mathbf{s}^n) \in \{1, 2, \ldots, 2^{nR}\}$ to each n-dimensional vector $\mathbf{s}^n \in \mathcal{S}^n$, and a decoder that assigns an estimate $\hat{\mathbf{s}}^n(m) \in \hat{\mathcal{S}}^n$ to each index $m \in \{1, 2, \ldots, 2^{nR}\}$. The set $\mathcal{C} = \{\hat{\mathbf{s}}^n(1), \ldots, \hat{\mathbf{s}}^n(2^{nR})\}$ constitutes the codebook. We will assume throughout that the encoder assigns to each source sequence the index $m^*(\mathbf{s}^n)$ according to the optimal rule

$$m^*(\mathbf{s}^n) = \operatorname*{argmin}_{m \in \{1, 2, \ldots, 2^{nR}\}} \frac{1}{n} \sum_{i=1}^{n} d\left(s_i, (\hat{\mathbf{s}}^n(m))_i\right),$$

such that the lossy source code is completely specified by the codebook $\mathcal{C}$ (and the distortion measure d).

The expected distortion associated with a lossy source code $\mathcal{C}$ is defined as

$$\mathbb{E}_{\mathbf{S}^n}\left(d(\mathbf{S}^n, \hat{\mathbf{S}}^n)\right) \triangleq \mathbb{E}_{\mathbf{S}^n}\left(\frac{1}{n}\sum_{i=1}^{n} d(S_i, \hat{S}_i)\right).$$

A distortion-rate pair (D, R) is said to be achievable if there exists a sequence of $(2^{nR}, n)$ codes with $\limsup_{n \to \infty} \mathbb{E}_{\mathbf{S}^n}\left(d(\mathbf{S}^n, \hat{\mathbf{S}}^n)\right) \leq D$. The distortion-rate function $D_{P_S}(R)$ of the source P_S is the infimum of distortions D such that (D, R) is achievable, and is given by [3], [4]

$$D_{P_S}(R) \triangleq \min_{P_{\hat{S}|S} : I(S; \hat{S}) \leq R} \sum_{s \in \mathcal{S}, \hat{s} \in \hat{\mathcal{S}}} P_S(s) P_{\hat{S}|S}(\hat{s}|s) d(s, \hat{s}). \tag{1}$$

For a sequence of $(2^{nR}, n)$ codes and $0 < R' < R$, we say that almost every subset with cardinality $2^{nR'}$ satisfies a certain property if the fraction of subsets with cardinality $2^{nR'}$ that do not satisfy the property vanishes with n.

In [5], Ziv proved that there exists a codebook with rate R that asymptotically achieves the distortion-rate function $D(R)$, regardless of the underlying source distribution. His result holds for any stationary source and even for a certain class of nonstationary sources. Our main result, stated below, only deals with i.i.d. sources with unknown distribution, and is hence less general than [5] in terms of the assumptions made on the source's distribution. However, it extends [5] in the sense that the distortion attained by subsets, and not just the full codebook, is shown to be universally optimal.

Theorem 1: Let S be a DMS over the alphabet $\mathcal{S}$, $\hat{\mathcal{S}}$ a reconstruction alphabet, and $d : \mathcal{S} \times \hat{\mathcal{S}} \mapsto \mathbb{R}^+$ a bounded

distortion measure. For any $R > 0$, $\delta > 0$, there exists a sequence of $(2^{nR}, n)$ codebooks $\mathcal{C} \subset \hat{\mathcal{S}}^n$ with the property that for any source pmf P_S on $\mathcal{S}$ and any $0 < R' < R$, almost every subset of $2^{nR'}$ codewords from $\mathcal{C}$ achieves an expected distortion no greater than $D_{P_S}(R' - \delta)$.

Remark 1: Ziv's result [5] is based on first splitting the source sequence into ℓ blocks of length k ($\ell \gg k$). Then, the best $(2^{kR}, k)$ lossy source code, with respect to the average distortion among the ℓ blocks, is found, and conveyed to the receiver. Afterwards, each block is encoded using this source code and its index is sent to the receiver. If the number of blocks of length k is large enough, the overhead for transmitting the codebook becomes negligible, which allows the scheme to perform well for a large class of distributions. This construction, however, is not subset universal. To see this, note that the empirical distribution of most codewords in the obtained codebook is that of the largest type. Thus, a random subset of $2^{nR'}$ codewords will contain a negligible fraction of codewords from other types, and in general cannot attain the distortion-rate function unless the optimal reconstruction distribution at rate R' is equal to that of the largest type as well.

The proof of Theorem 1 relies on independently drawing the codewords of $\mathcal{C}$ from a mixture distribution, and will be given in Section IV. In the next section we lay the ground by proving certain properties of mixture distributions.

III. PROPERTIES OF MIXTURE DISTRIBUTIONS

We follow the notation of [6], and define the empirical pmf of an n-dimensional sequence $\mathbf{y}^n$ with elements from $\mathcal{Y}$ as

$$\pi(y|\mathbf{y}^n) \triangleq \tfrac{1}{n} |i \ : \ y_i = y| \text{ for } y \in \mathcal{Y}.$$

Similarly, the empirical pmf of a pair of n-dimensional sequences $(\mathbf{y}^n, \mathbf{z}^n)$ with elements from $\mathcal{Y} \times \mathcal{Z}$ is defined as

$$\pi(y, z|\mathbf{y}^n, \mathbf{z}^n) \triangleq \tfrac{1}{n} |i : (y_i, z_i) = (y, z)| \text{ for } (y, z) \in \mathcal{Y} \times \mathcal{Z}.$$

Let $P_{YZ} = P_Y P_{Z|Y}$ be a joint pmf on $\mathcal{Y} \times \mathcal{Z}$. The set of ε-typical n-dimensional sequences w.r.t. P_Y is defined as

$$\mathcal{T}_\varepsilon^{(n)}(P_Y) \triangleq \left\{\mathbf{y}^n \ : \ |\pi(y|\mathbf{y}^n) - P_Y(y)| \leq \varepsilon P_Y(y) \ \forall y \in \mathcal{Y}\right\},$$

and the set of jointly ε-typical n-dimensional sequences w.r.t. P_{YZ} is defined as

$$\mathcal{T}_\varepsilon^{(n)}(P_{YZ}) \triangleq \Big\{(\mathbf{y}^n, \mathbf{z}^n) \ :$$

$$|\pi(y, z|\mathbf{y}^n, \mathbf{z}^n) - P_{YZ}(y, z)| \leq \varepsilon P_{YZ}(y, z) \ \forall (y, z) \in \mathcal{Y} \times \mathcal{Z}\Big\}.$$

We also define the set of conditionally ε-typical n-dimensional sequences w.r.t. P_{YZ} as

$$\mathcal{T}_\varepsilon^{(n)}(P_{YZ}|\mathbf{y}^n) \triangleq \left\{\mathbf{z}^n \ : \ (\mathbf{y}^n, \mathbf{z}^n) \in \mathcal{T}_\varepsilon^{(n)}(P_{YZ})\right\}.$$

The next statement follows from the definitions above [6].

Proposition 1: Let $\mathbf{y}^n \in \mathcal{Y}^n$. For every $\mathbf{z}^n \in \mathcal{T}_\varepsilon^{(n)}(P_{YZ}|\mathbf{y}^n)$ we have $\mathbf{z}^n \in \mathcal{T}_\varepsilon^{(n)}(P_Z)$. If in addition, $\mathbf{y}^n \in \mathcal{T}_{\varepsilon'}^{(n)}(P_Y)$, for some $\varepsilon' < \varepsilon$, then for n large enough

$$|\mathcal{T}_\varepsilon^{(n)}(P_{YZ}|\mathbf{y}^n)| \geq (1-\varepsilon)2^{n(1-\varepsilon)H(Z|Y)}.$$

Let $\mathcal{P}^{|\mathcal{Z}|}$ denote the simplex containing all probability mass functions on $\mathcal{Z}$. For every $\theta \in \mathcal{P}^{|\mathcal{Z}|}$, let $P_\theta(z)$ be the corresponding pmf evaluated at z. Let $w(\theta)$ be some probability density function on $\mathcal{P}^{|\mathcal{Z}|}$. We may now define the *mixture* distribution Q as [7]

$$Q(\mathbf{z}^n) = \int_{\theta \in \mathcal{P}^{|\mathcal{Z}|}} w(\theta) \prod_{i=1}^{n} P_\theta(z_i)d\theta. \qquad (2)$$

The following propositions are proved in the appendix.

Proposition 2: Let $\mathbf{Z}^n$ be a random n-dimensional sequence drawn according to $Q(\mathbf{z}^n)$ defined in (2). Let P_{YZ} be some pmf on $\mathcal{Y} \times \mathcal{Z}$, and let $\mathbf{y}^n \in \mathcal{T}_{\varepsilon'}^{(n)}(P_Y)$, for some $\varepsilon' < \varepsilon$. For n large enough, we have

$$\Pr\left(\mathbf{Z}^n \in \mathcal{T}_\varepsilon^{(n)}(P_{YZ}|\mathbf{y}^n)\right)$$
$$\geq (1-\varepsilon)\int_{\theta \in \mathcal{P}^{|\mathcal{Z}|}} w(\theta)2^{-n(1+\varepsilon)(I(Y;Z)+D(P_Z||P_\theta)+2\varepsilon H(Z|Y))}d\theta.$$

Proposition 3: Let P_Z be some distribution in $\mathcal{P}^{|\mathcal{Z}|}$. For any $0 < \xi < 1/|\mathcal{Z}|^2$, there exists a subset $\mathcal{V} \subset \mathcal{P}^{|\mathcal{Z}|}$ with Lebesgue measure $\xi^{|\mathcal{Z}|-1}$ on $\mathcal{P}^{|\mathcal{Z}|}$, such that for all $P_\theta \in \mathcal{V}$

$$D(P_Z||P_\theta) \leq \log \frac{1}{1-\xi|\mathcal{Z}|^2},$$

Combining Proposition 2 and 3 yields the following lemma.

Lemma 1: Let $Q(\mathbf{z}^n)$ be as defined in (2), with $w(\theta)$ taken as the uniform distribution on $\mathcal{P}^{|\mathcal{Z}|}$, and let $\mathbf{Z}^n$ be a random n-dimensional sequence drawn according to $Q(\mathbf{z}^n)$. Let P_{YZ} be some pmf on $\mathcal{Y} \times \mathcal{Z}$, and let $\mathbf{y}^n \in \mathcal{T}_{\varepsilon'}^{(n)}(P_Y)$, for some $\varepsilon' < \varepsilon$. For n large enough, we have

$$\Pr\left(\mathbf{Z}^n \in \mathcal{T}_\varepsilon^{(n)}(P_{YZ}|\mathbf{y}^n)\right) \geq 2^{-n(I(Y;Z)+\delta(\epsilon))},$$

where $\delta(\epsilon) \to 0$ for $\epsilon \to 0$.
Proof. Let $c_{|\mathcal{Z}|}$ be the Lebesgue measure of the simplex $\mathcal{P}^{|\mathcal{Z}|}$. Clearly, we have that $w(\theta) = c_{|\mathcal{Z}|}^{-1}$ for $\theta \in \mathcal{P}^{|\mathcal{Z}|}$ and $w(\theta) = 0$ otherwise. Setting $\xi = (1-2^{-\varepsilon})/|\mathcal{Z}|^2$ in Proposition 3 implies that there exists a set $\mathcal{V} \subset \mathcal{P}^{|\mathcal{Z}|}$ with volume $((1-2^{-\varepsilon})/|\mathcal{Z}|^2)^{|\mathcal{Z}|-1}$ such that $D(P_Z||P_\theta) < \varepsilon$ for any $P_\theta \in \mathcal{V}$. Combining this with Proposition 2 gives

$$\Pr\left(\mathbf{Z}^n \in \mathcal{T}_\varepsilon^{(n)}(P_{YZ}|\mathbf{y}^n)\right)$$
$$\geq (1-\varepsilon)\int_{\theta \in \mathcal{V}} w(\theta)2^{-n(1+\varepsilon)(I(Y;Z)+\varepsilon+2\varepsilon H(Z|Y))}d\theta$$
$$> \frac{1-\varepsilon}{c_{|\mathcal{Z}|}}\left(\frac{1-2^{-\varepsilon}}{|\mathcal{Z}|^2}\right)^{|\mathcal{Z}|-1} 2^{-n(1+\varepsilon)(I(Y;Z)+\varepsilon+2\varepsilon H(Z|Y))}$$

$$= 2^{-n(I(Y;Z)+\delta(\epsilon))},$$

where $\delta(\epsilon) \to 0$ for $\epsilon \to 0$. ∎

IV. PROOF OF THEOREM 1

Random codebook generation: Let

$$Q(\hat{\mathbf{s}}^n) = \int_{\theta \in \mathcal{P}^{|\hat{\mathcal{S}}|}} w(\theta) \prod_{i=1}^{n} P_\theta(\hat{s}_i)d\theta,$$

where $\mathcal{P}^{|\hat{\mathcal{S}}|}$ is the simplex containing all probability mass functions on $\hat{\mathcal{S}}$ and $w(\theta)$ is the uniform distribution on $\mathcal{P}^{|\hat{\mathcal{S}}|}$. Randomly and independently generate 2^{nR} sequences $\hat{\mathbf{s}}^n(m)$, $m \in \{1, 2, \ldots, 2^{nR}\}$, each according to $Q(\hat{\mathbf{s}}^n)$. These sequence constitutes the full codebook $\mathcal{C}$.

A subset of the codebook, indexed by $(\mathcal{I}, R')$, consists of an index set $\mathcal{I} \subset \{1, 2, \ldots, 2^{nR}\}$ with cardinality $|\mathcal{I}| = 2^{nR'}$, $0 < R' < R$, and the corresponding sequences $\hat{\mathbf{s}}^n(m)$, $m \in \mathcal{I}$. An arbitrary $(\mathcal{I}, R')$ subset of $\mathcal{C}$ is revealed to the encoder and the decoder.

Encoding: Given the source sequence $\mathbf{s}^n$ and a subset $(\mathcal{I}, R')$ of $\mathcal{C}$, the optimal encoder sends the index of the codeword that achieves the minimal distortion, i.e., it sends

$$m^* = \underset{m \in \mathcal{I}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} d\left(s_i, (\hat{\mathbf{s}}^n(m))_i\right). \qquad (3)$$

Note that the optimal encoder (3) is universal, i.e., it does not have to know the true underlying source pmf, and that such knowledge can in no way improve its performance.

Decoding: Upon receiving the index m, the decoder simply sets the reconstruction sequence as $\hat{\mathbf{s}}^n(m)$.

Analysis: Let us define a quantized grid of rates in $[0, R)$ whose resolution is $\Delta > 0$

$$\mathcal{R}_\Delta \triangleq \{R_j : R_j = j \cdot \Delta, j = 1, \ldots, \lfloor R/\Delta \rfloor\}.$$

We further define a set of quantized distributions as follows. For $0 < \Delta < 1$ and $0 < p_{\min} < 1$ define the set of points

$$\mathcal{G}_{\Delta, p_{\min}, |\mathcal{S}|} \triangleq p_{\min} \cdot \left\{1, \left(1 + \frac{\Delta}{|\mathcal{S}|}\right), \left(1 + \frac{\Delta}{|\mathcal{S}|}\right)^2, \right.$$
$$\left. \ldots, \left(1 + \frac{\Delta}{|\mathcal{S}|}\right)^{\left\lfloor \frac{-\log p_{\min}}{\log\left(1 + \frac{\Delta}{|\mathcal{S}|}\right)} \right\rfloor}\right\}.$$

Our set of quantized distributions is denoted by $\mathcal{P}_{\Delta, p_{\min}}^{|\mathcal{S}|}$ and consists of all vectors in the simplex $\mathcal{P}^{|\mathcal{S}|}$ with at least $|\mathcal{S}| - 1$ components that belong to $\mathcal{G}_{\Delta, p_{\min}, |\mathcal{S}|}$. Clearly, the cardinality $|\mathcal{P}_{\Delta, p_{\min}}^{|\mathcal{S}|}|$ of this set is a function of only $|\mathcal{S}|$, $p_{\min}$ and Δ. Moreover, for any pmf $P_\theta \in \mathcal{P}^{|\mathcal{S}|}$ with the property that $\min_{s \in \mathcal{S}} P_\theta(s) > (1 - \Delta/|\mathcal{S}|)p_{\min}$, there exists a pmf $P_{\theta'} \in \mathcal{P}_{\Delta, p_{\min}}^{|\mathcal{S}|}$ such that $|P_\theta(s) - P_{\theta'}(s)| \leq \Delta P_{\theta'}(s)$ for all $s \in \mathcal{S}$. To see this, construct $P_{\theta'}$ by quantizing all but the greatest entry of P_θ to the nearest point in $\mathcal{G}_{\Delta, p_{\min}, |\mathcal{S}|}$ and set the remaining entry such that $\sum_{s \in \mathcal{S}} P_{\theta'}(s) = 1$.

Let $\mathcal{P}_{p_{\min}}^{|\mathcal{S}|}$ be the set of all probability mass functions on $\mathcal{S}$ whose minimal mass is smaller than $p_{\min}$. We can partition the remaining part of the simplex as

$$\mathcal{P}^{|\mathcal{S}|} \setminus \mathcal{P}_{p_{\min}}^{|\mathcal{S}|} = \bigcup_{P_S \in \mathcal{P}_{\Delta,p_{\min}}^{|\mathcal{S}|}} \mathcal{P}(P_s)$$

where the sets $\mathcal{P}(P_s) \subset \mathcal{P}^{|\mathcal{S}|}$ are disjoint and satisfy the property that for any $P_\theta \in \mathcal{P}(P_s)$ we have $|P_\theta(s) - P_S(s)| < \Delta P_S(s)$ for all $s \in \mathcal{S}$.

Let $R' \in \mathcal{R}_\Delta$ and $P_S \in \mathcal{P}_{\Delta,p_{\min}}^{|\mathcal{S}|}$. We will next show that almost every codebook in our ensemble satisfies the property that almost every subset of $2^{nR'}$ of its codewords achieves average distortion no greater than $D_{P_S}(R' - \delta)$ with respect to all sources whose pmf belongs to $\mathcal{P}(P_S)$. Since $|\mathcal{R}_\Delta| \cdot |\mathcal{P}_{\Delta,p_{\min}}^{|\mathcal{S}|}|$ does not increase with n, there must exist a codebook that simultaneously satisfies this property for all $R' \in \mathcal{R}_\Delta$ and $P_S \in \mathcal{P}_{\Delta,p_{\min}}^{|\mathcal{S}|}$. The theorem will then follow by taking $\Delta \to 0$, $p_{\min} \to 0$, and using the continuity of $D_{P_S}(R)$ w.r.t. R and P_S for a bounded distortion measure d [8].

Let $R' \in \mathcal{R}_\Delta$, $P_S \in \mathcal{P}_{\Delta,p_{\min}}^{|\mathcal{S}|}$ and $P_S' \in \mathcal{P}(P_s)$. We upper bound the average distortion of a given subset $(\mathcal{I}, R')$ of $\mathcal{C}$ over the ensemble of codebooks and a DMS $\mathbf{S}^n$ with pdf P_S'. To that end, rather than analyzing the average distortion attained by the encoder (3), we analyze the average distortion attained by the following suboptimal encoder. The suboptimal encoder first solves the minimization problem

$$P_{\hat{S}|S}^{R'} = \operatorname*{argmin}_{P_{\hat{S}|S}: I(S;\hat{S}) \leq R' - \delta} \sum_{s \in \mathcal{S}, \hat{s} \in \hat{\mathcal{S}}} P_S(s) P_{\hat{S}|S}(\hat{s}|s) d(s, \hat{s}),$$

and sets $P_{S\hat{S}}^{R'} = P_S P_{\hat{S}|S}^{R'}$ as the target joint pmf. Then, given a source sequence $\mathbf{s}^n$, it looks for the smallest index $m \in \mathcal{I}$ such that $(\mathbf{s}^n, \hat{\mathbf{s}}^n(m)) \in \mathcal{T}_\varepsilon^{(n)}(P_{S\hat{S}}^{R'})$ and sends it to the decoder. If no such index is found, the smallest $m \in \mathcal{I}$ is sent to the decoder. Note that this suboptimal encoder is not universal, as it requires knowledge of P_S. Nevertheless, its average distortion is by definition greater than that of the optimal encoder which is universal, and will therefore indeed serve as an upper bound on the distortion of the universal encoder.

Let $\mathbf{S}^n$ be a random source drawn i.i.d. according to P_S' and define $\hat{\mathbf{S}}^n$ as the reconstruction obtained by applying the suboptimal encoder (and the decoder described above). Define the error event $\mathcal{E}(\mathcal{I}, R')$ as

$$\mathcal{E}(\mathcal{I}, R') \triangleq \left\{ (\mathbf{S}^n, \hat{\mathbf{s}}^n(m)) \notin \mathcal{T}_\varepsilon^{(n)}(P_{S\hat{S}}^{R'}) \text{ for all } m \in \mathcal{I} \right\}. \tag{4}$$

By the definition of $P_{S\hat{S}}^{R'}$ and the definition of the typical set, we have that if $\mathcal{E}(\mathcal{I}, R')$ did not occur

$$d(\mathbf{S}^n, \hat{\mathbf{S}}^n) \leq (1 + \varepsilon) D_{P_S}(R' - \delta), \tag{5}$$

where $D_{P_S}(\cdot)$ is the distortion-rate function (1) with respect to P_S. Further, the error event satisfies $\mathcal{E}(\mathcal{I}, R') \subset \mathcal{E}_1 \cup \mathcal{E}_2(\mathcal{I}, R')$, where

$$\mathcal{E}_1 = \left\{ \mathbf{S}^n \notin \mathcal{T}_{\varepsilon'}^{(n)}(P_S) \right\}$$

$$\mathcal{E}_2(\mathcal{I}, R') = \left\{ \mathbf{S}^n \in \mathcal{T}_{\varepsilon'}^{(n)}(P_S), \right.$$
$$\left. \hat{\mathbf{s}}^n(m) \notin \mathcal{T}_\varepsilon^{(n)}(P_{S\hat{S}}^{R'}|\mathbf{S}^n) \text{ for all } m \in \mathcal{I} \right\}, \tag{6}$$

and $\Delta < \varepsilon' < \varepsilon$. Note that $\mathcal{E}_1$ is independent of $\mathcal{I}$ and that $\Pr(\mathcal{E}_1) \to 0$ as $n \to \infty$ by the (weak) law of large numbers and the fact that $\Delta < \varepsilon'$. The second error event does depend on $\mathcal{I}$. We will now show that its expected probability with respect to the ensemble of codebooks $\mathbb{E}_\mathcal{C}\left(\Pr(\mathcal{E}_2(\mathcal{I}, R'))\right)$ vanishes for almost all $\mathcal{I} \subset \{1, 2, \ldots, 2^{nR}\}$ of cardinality $2^{nR'}$.

Let $\mathcal{U}$ be a random subset of indices drawn from the uniform distribution on all subsets of $\{1, \ldots, 2^{nR}\}$ with cardinality $2^{nR'}$. By the random symmetric generation process of the codewords in $\mathcal{C}$, the value of $\mathbb{E}_{\mathcal{C}|\mathcal{U}}\left(\Pr(\mathcal{E}_2(\mathcal{U}, R')) | \mathcal{U} = \mathcal{I}\right)$ depends on the index set $\mathcal{I}$ only through its cardinality $2^{nR'}$. We therefore have

$$\mathbb{E}_{\mathcal{C},\mathcal{U}}\left(\Pr(\mathcal{E}_2(\mathcal{U}, R'))\right) = \mathbb{E}_\mathcal{U}\mathbb{E}_{\mathcal{C}|\mathcal{U}}\left(\Pr(\mathcal{E}_2(\mathcal{U}, R')|\mathcal{U})\right)$$

$$= \mathbb{E}_\mathcal{U}\left(\sum_{\mathbf{s}^n \in \mathcal{T}_{\varepsilon'}^{(n)}(P_S)} P'_{\mathbf{S}^n}(\mathbf{s}^n)\right.$$

$$\Pr\left(\hat{\mathbf{S}}^n(m) \notin \mathcal{T}_\varepsilon^{(n)}(P_{S\hat{S}}^{R'}|\mathbf{s}^n) \text{ for all } m \in \mathcal{U} \mid \mathcal{U}\right)\bigg)$$

$$= \mathbb{E}_\mathcal{U}\left(\sum_{\mathbf{s}^n \in \mathcal{T}_{\varepsilon'}^{(n)}(P_S)} P'_{\mathbf{S}^n}(\mathbf{s}^n)\right.$$

$$\prod_{m \in \mathcal{U}} \Pr\left(\hat{\mathbf{S}}^n(m) \notin \mathcal{T}_\varepsilon^{(n)}(P_{S\hat{S}}^{R'}|\mathbf{s}^n) | \mathcal{U}\right)\bigg)$$

$$= \sum_{\mathbf{s}^n \in \mathcal{T}_{\varepsilon'}^{(n)}(P_S)} P'_{\mathbf{S}^n}(\mathbf{s}^n) \left(\Pr\left(\hat{\mathbf{S}}^n(1) \notin \mathcal{T}_\varepsilon^{(n)}(P_{S\hat{S}}^{R'}|\mathbf{s}^n)\right)\right)^{2^{nR'}}$$
$$\tag{7}$$

By Lemma 1 we have

$$\Pr\left(\hat{\mathbf{S}}^n(1) \in \mathcal{T}_\varepsilon^{(n)}(P_{S\hat{S}}^{R'}|\mathbf{s}^n)\right) \geq 2^{-n(I(S;\hat{S}^{R'})+\delta(\epsilon))},$$

where $I(S; \hat{S}^{R'})$ is the mutual information between S and $\hat{S}$ under the target joint pmf $P_{S\hat{S}}^{R'}$. This implies that

$$\left(\Pr\left(\hat{\mathbf{S}}^n(1) \notin \mathcal{T}_\varepsilon^{(n)}(P_{S\hat{S}}^{R'}|\mathbf{s}^n)\right)\right)^{2^{nR'}}$$

$$\leq \left(1 - 2^{-n(I(S;\hat{S}^{R'})+\delta(\epsilon))}\right)^{2^{nR'}}$$

$$\leq \exp\left\{-2^{n(R'-I(S;\hat{S}^{R'})-\delta(\epsilon))}\right\} \tag{8}$$

$$\leq \exp\left\{-2^{n(\delta-\delta(\epsilon))}\right\}, \tag{9}$$

where (8) follows from the inequality $(1 - x)^k \leq e^{-kx}$ which holds for $x \in [0, 1]$, and (9) follows from the definition of $P_{\hat{S}|S}^{R'}$. Combining (7) and (9) we have

$$\mathbb{E}_{\mathcal{C},\mathcal{U}}\left(\Pr(\mathcal{E}_2(\mathcal{U}, R'))\right) \leq \exp\left\{-2^{n(\delta-\delta(\epsilon))}\right\}. \tag{10}$$

Thus, for any $\delta > 0$ we may take $0 < \Delta < \varepsilon' < \varepsilon$ sufficiently small such that $\delta(\varepsilon) < \delta$ and (10) can be made arbitrary small when increasing n. The proof is concluded by applying the following arguments:

- By Markov's inequality, (10) implies that for every $P'_S \in \mathcal{P}(P_S)$ and almost every subset $\mathcal{I} \subset \{1, 2, \dots, 2^{nR}\}$ of cardinality $2^{nR'}$ the expectation $\mathbb{E}_{\mathcal{C}}\left(\Pr\left(\mathcal{E}_2(\mathcal{I}, R')\right)\right)$ vanishes for large n.
- Applying Markov's inequality again, this time with respect to $\mathcal{C}$ we see that almost all codebooks in the ensemble satisfy that for every $P'_S \in \mathcal{P}(P_S)$ and almost every subset $\mathcal{I} \subset \{1, 2, \dots, 2^{nR}\}$ of cardinality $2^{nR'}$ the probability $\Pr\left(\mathcal{E}_2(\mathcal{I}, R')\right)$ vanishes for large n.
- As discussed above, $\Pr(\mathcal{E}_1) \to 0$ as $n \to \infty$ for every $P'_S \in \mathcal{P}(P_S)$, regardless of $\mathcal{I}$ and $\mathcal{C}$. By the union bound $\Pr\left(\mathcal{E}(\mathcal{I}, R')\right) \leq \Pr(\mathcal{E}_1) + \Pr\left(\mathcal{E}_2(\mathcal{I}, R')\right)$. We therefore have that almost all codebooks in the ensemble satisfy that for every $P'_S \in \mathcal{P}(P_S)$ and almost every subset $\mathcal{I} \subset \{1, 2, \dots, 2^{nR}\}$ of cardinality $2^{nR'}$ the probability $\Pr\left(\mathcal{E}(\mathcal{I}, R')\right)$ vanishes for large n.
- Since the distortion measure d is bounded, this together with (5) implies that almost all codebooks in the ensemble satisfy that for every $P'_S \in \mathcal{P}(P_S)$ and almost every subset $\mathcal{I} \subset \{1, 2, \dots, 2^{nR}\}$ of cardinality $2^{nR'}$ the average distortion approaches $(1+\varepsilon)D_{P_S}(R'-\delta)$ as n increases.
- Since $|\mathcal{R}_\Delta| \cdot |\mathcal{P}^{|\mathcal{S}|}_{\Delta, p_{\min}}|$ does not increase with n, there must exist a sequence of codebooks whose average distortion approaches $(1+\varepsilon)D_{P_S}(R'-\delta)$ for almost every subset $\mathcal{I} \subset \{1, 2, \dots, 2^{nR}\}$ of cardinality $2^{nR'}$, simultaneously for all $R' \in \mathcal{R}_\Delta$, $P_S \in \mathcal{P}^{|\mathcal{S}|}_{\Delta, p_{\min}}$ and $P'_S \in \mathcal{P}(P_S)$.
- The theorem follows by taking $\varepsilon \to 0$, $p_{\min} \to 0$ and using the continuity of the function $D_{P_S}(R)$ with respect to P_S and R for a bounded distortion measure d [8].

APPENDIX

Proof of Proposition 2. From Proposition 1 and the definition of ε-typical sequences, we have that $\pi(z|\mathbf{z}^n) \leq (1+\epsilon)P_Z(z)$ for any $\mathbf{z}^n \in \mathcal{T}^{(n)}_\varepsilon(P_{YZ}|\mathbf{y}^n)$. We can therefore write

$$\Pr\left(\mathbf{Z}^n \in \mathcal{T}^{(n)}_\varepsilon(P_{YZ}|\mathbf{y}^n)\right) = \sum_{\mathbf{z}^n \in \mathcal{T}^{(n)}_\varepsilon(P_{YZ}|\mathbf{y}^n)} Q(\mathbf{z}^n)$$

$$= \sum_{\mathbf{z}^n \in \mathcal{T}^{(n)}_\varepsilon(P_{YZ}|\mathbf{y}^n)} \int_{\theta \in \mathcal{P}^{|\mathcal{Z}|}} w(\theta) \prod_{i=1}^{n} P_\theta(z_i) d\theta$$

$$\geq \sum_{\mathbf{z}^n \in \mathcal{T}^{(n)}_\varepsilon(P_{YZ}|\mathbf{y}^n)} \int_{\theta \in \mathcal{P}^{|\mathcal{Z}|}} w(\theta) \prod_{z \in \mathcal{Z}} P_\theta(z)^{n(1+\varepsilon)P_Z(z)} d\theta$$

$$\geq (1-\varepsilon) 2^{n(1-\varepsilon)H(Z|Y)}$$

$$\cdot \int_{\theta \in \mathcal{P}^{|\mathcal{Z}|}} w(\theta) 2^{n(1+\epsilon)\sum_{z \in \mathcal{Z}} P_Z(z) \log P_\theta(z)} d\theta$$

$$= (1-\varepsilon) \int_{\theta \in \mathcal{P}^{|\mathcal{Z}|}} w(\theta) 2^{nE} d\theta,$$

where the last inequality follows from Proposition 1, and

$$E = (1-\varepsilon)H(Z|Y) + (1+\varepsilon) \sum_{z \in \mathcal{Z}} P_Z(z) \log P_\theta(z)$$

$$= -2\varepsilon H(Z|Y) + (1+\varepsilon)\left(H(Z|Y) - D(P_Z||P_\theta) - H(Z)\right)$$

$$\geq -(1+\varepsilon)\left(I(Y;Z) + D(P_Z||P_\theta) + 2\varepsilon H(Z|Y)\right).$$

as desired. ∎

Proof of Proposition 3. Without loss of generality, we may assume $\mathcal{Z} = \{1, \dots, |\mathcal{Z}|\}$ and that $P_Z(1) \leq P_Z(2) \leq \dots \leq P_Z(|\mathcal{Z}|)$. Note that under these assumptions $(1/|\mathcal{Z}|) \leq P_Z(|\mathcal{Z}|) \leq 1$, and $P_Z(|\mathcal{Z}| - 1) \leq \frac{1}{2}$. Let us define the perturbation set

$$\mathcal{U} = \left\{(u_1, \dots, u_{|\mathcal{Z}|}) : 0 \leq u_i < \xi, \text{ for } i = 1, \dots, |\mathcal{Z}| - 1 \right.$$

$$\left. u_{|\mathcal{Z}|} = -\sum_{i=1}^{|\mathcal{Z}|-1} u_i \right\},$$

and the set $\mathcal{V} = P_Z + \mathcal{U}$, where the sum is in the Minkowski sense. Clearly, $\mathcal{V} \subset \mathcal{P}^{|\mathcal{Z}|}$ for any $0 < \xi < 1/|\mathcal{Z}|^2$, and its Lebesgue measure on the simplex $\mathcal{P}^{|\mathcal{Z}|}$ is $|\mathcal{V}| = \xi^{|\mathcal{Z}|-1}$. In addition, for any $P_\theta \in \mathcal{V}$ we have

$$D(P_Z||P_\theta) = \sum_{z \in \mathcal{Z}} P_Z(z) \log \frac{P_Z(z)}{P_\theta(z)}$$

$$\leq P_Z(|\mathcal{Z}|) \log \frac{P_Z(|\mathcal{Z}|)}{P_Z(|\mathcal{Z}|) - \xi(|\mathcal{Z}| - 1)}$$

$$\leq \log \frac{1/|\mathcal{Z}|}{(1/|\mathcal{Z}|) - \xi(|\mathcal{Z}| - 1)}$$

$$\leq \log \frac{1}{1 - \xi|\mathcal{Z}|^2}.$$

∎

REFERENCES

[1] S. I. Gelfand and M. S. Pinsker, "Coding for channels with random parameters," *Problem of Control and Information Theory*, vol. 9, no. 1, pp. 19–31, 1980.
[2] C. Heegard and A. Gamal, "On the capacity of computer memory with defects," *IEEE Transactions on Information Theory*, vol. 29, no. 5, pp. 731–739, Sep 1983.
[3] T. Berger, *Rate distortion theory: A mathematical basis for data compression*. Englewood Cliffs, NJ: Prentice-Hall, 1971.
[4] T. Berger and J. Gibson, "Lossy source coding," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2693–2723, Oct 1998.
[5] J. Ziv, "Coding of sources with unknown statistics–II: Distortion relative to a fidelity criterion," *IEEE Transactions on Information Theory*, vol. 18, no. 3, pp. 389–394, May 1972.
[6] A. El Gamal and Y.-H. Kim, *Network information theory*. Cambridge University Press, 2011.
[7] N. Merhav and M. Feder, "Universal prediction," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2124–2147, Oct 1998.
[8] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic Press, 1982.