The output distribution of good lossy source codes

Victoria Kostina*, Sergio Verdú†

*California Institute of Technology
Pasadena, CA 91125, USA
vkostina@caltech.edu

†Princeton University
Princeton, NJ 08544, USA
verdu@princeton.edu

***Abstract*—This paper provides a necessary condition good rate-distortion codes must satisfy. Specifically, it is shown that as the blocklength increases, the distribution of the input given the output of a good lossy code converges to the distribution of the input given the output of the joint distribution achieving the rate-distortion function, in terms of the normalized conditional relative entropy. The result holds for stationary ergodic sources with subadditive distortion measures, both for fixed-length and variable-length compression. A similar necessary condition is given for lossy joint source-channel coding.**

I. Introduction

A lossy source coder operating at blocklength k assigns representation z^k to a given block of source outcomes, s^k. The quality of a lossy coder is measured by the tradeoff between the rate (or the total number of representation points) and the distortion between source and the representation. A lossy source code operating at a given fidelity is *good* if its rate approaches the information-theoretic minimum, i.e. the rate-distortion function. If the source blocks are distributed according to P_{S^k}, and the lossy coder assigns source blocks to their representations according to the conditional probability distribution $P_{Z^k|S^k}$, the joint distribution of the pairs of input/output blocks induced by that coder is $P_{S^k}P_{Z^k|S^k}$. This paper studies the properties of $P_{S^k}P_{Z^k|S^k}$. In particular, we show that the distributions generated by all good codes necessarily look alike and are close to the distribution that achieves the rate-distortion function (see Fig. 1). We give a precise characterization of that property, thereby providing a necessary condition that any good code must satisfy. Such a necessary condition aids in the search for practical codes, as it helps to discard potential candidates for good codes. Moreover, knowledge of distributional properties of good codes facilitates design of systems that include a source coding block as one of their components. For example, it was observed in [1] that a dispersion-optimal joint source-channel scheme can be implemented as a good lossy source code wrapped around a good channel code, provided that the channel code properly accounts for the statistics of the source encoder outputs. Knowing the output statistics of *any* good source code allows for universally almost-optimal designs of inner channel codes regardless of a particular implementation of the outer source code.

Distributions induced by good channel codes were studied in [2], [3]. Shannon [4, Sec. 25] was the first to comment on the fact that to maximize the transmission rate over an AWGN channel, the codewords must resemble a white Gaussian noise. Shamai and Verdú [2] showed that the distribution at the memoryless channel output induced by a capacity-achieving sequence of codes with vanishing error probability converges (in normalized relative entropy) to the capacity-achieving output distribution, a result later generalized to a non-vanishing maximal error probability by Polyanskiy and Verdú [3].

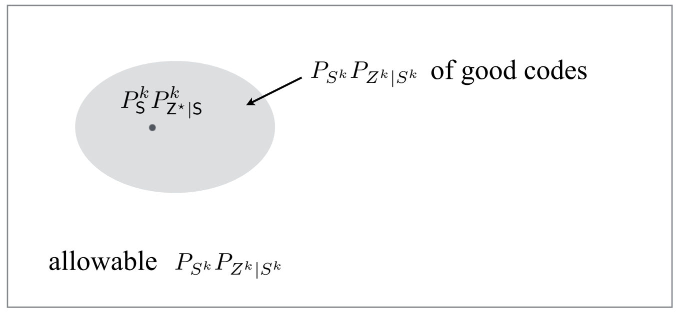

Fig. 1. The joint source/reproduction distribution of a good code not only satisfies the condition set out by the operational problem (in terms of distortion) but it belongs to a neighborhood of the joint distribution achieving the rate-distortion function.

Weissman and Ordentlich [5] studied the empirical marginal (per-letter) distributions induced by good source codes, i.e. the frequency of appearances of pairs of input/output letters (s, z) observed at the input and output of the lossy coder. In particular, they showed that for a stationary discrete memoryless source with separable distortion measure, any sequence of good codes operating at average distortion d has the following property: the frequency of appearances of pairs of input/output letters (s, z) converges almost surely to $P_{\mathsf{S}}P_{\mathsf{Z}^\star|\mathsf{S}}$, where $P_{\mathsf{Z}^\star|\mathsf{S}}$ is the probability kernel that attains the rate-distortion function $R(d)$. Kanlis et al. [6] showed that the type of most reproduction points of a good source code approaches $P_{\mathsf{Z}^\star}$, the marginal of $P_{\mathsf{S}}P_{\mathsf{Z}^\star|\mathsf{S}}$. Schieler and Cuff [7] studied actual joint blockwise (rather than empirical) distributions induced by good source codes and showed, in particular, that for discrete memoryless sources

$$\lim_{k\to\infty} \frac{1}{k} D\left(P_{S^k|Z^k} \| P^k_{\mathsf{S}|\mathsf{Z}^\star} | P_{Z^k}\right) = 0, \tag{1}$$

where $P_{S^k|Z^k}P_{Z^k} = P_{Z^k|S^k}P_{S^k}$ is the joint distribution between the input and the output block induced by the code, $P_{\mathsf{S}|\mathsf{Z}^\star}$ is the backward conditional distribution corresponding to the joint distribution achieving the rate-distortion function, and $D(P_{S^k|Z^k} \| P^k_{\mathsf{S}|\mathsf{Z}^\star} | P_{Z^k}) = D(P_{S^k|Z^k}P_{Z^k} \| P^k_{\mathsf{S}|\mathsf{Z}^\star}P_{Z^k})$ is the conditional relative entropy. Like [7], this paper focuses on the actual, not empirical, joint blockwise distributions of good codes. As we will see (1) will follow as a simple corollary to our main result.

In this paper, we consider a general stationary ergodic source with subadditive distortion measure and we show that for any sequence of codes operating at average distortion d,

$$D(P_{S^k|Z^k} \| P_{S^k|Z^{k\star}} | P_{Z^k}) \leq kR - \mathbb{R}_{S^k}(d) \tag{2}$$

for arbitrary k, where R is the rate of the code, and $\mathbb{R}_{S^k}(d)$ is the k-th order rate-distortion function:

$$\mathbb{R}_{S^k}(d) \triangleq \min_{\substack{P_{Z^k|S^k}: \\ \mathbb{E}\left[\mathsf{d}_k(S^k, Z^k)\right] \leq d}} I(S^k; Z^k), \tag{3}$$

where $\mathsf{d}_k(S^k, Z^k)$ is the distortion between S^k and Z^k, and $P_{Z^{k\star}|S^k}$ is the conditional distribution that achieves $\mathbb{R}_{S^k}(d)$. Of course, for good codes (2) implies convergence of the normalized conditional relative entropy to 0, as in (1). Furthermore, we generalize (2) to variable-length compression and to joint source-channel coding.

II. Single shot lossy compression

The lossy source coding problem can be abstracted as follows: the source $S \in \mathcal{M}$, where $\mathcal{M}$ is an abstract alphabet, is to be represented, under a rate constraint, by codewords living in the reproduction alphabet $\hat{\mathcal{M}}$. The fidelity of reproduction is quantified by the distortion measure $\mathsf{d}\colon \mathcal{M} \times \hat{\mathcal{M}} \mapsto [0, +\infty]$. A lossy code gives rise to a transition probability kernel $P_{Z|S}\colon \mathcal{M} \mapsto \hat{\mathcal{M}}$. We are interested in the properties $P_{Z|S}$ must necessarily have, provided that optimal rate-distortion tradeoffs are approached.

Throughout the paper, we assume that the target distortion $d \geq d_{\min}$, where $d_{\min}$ is the infimum of values at which the minimal mutual information quantity

$$\mathbb{R}_S(d) \triangleq \min_{\substack{P_{Z|S}: \\ \mathbb{E}[\mathsf{d}(S,Z)] \leq d}} I(S; Z) \tag{4}$$

is finite[1]. We also assume the mild condition that $P_{Z^\star|S}$ achieves the minimum in the right side of (4) while satisfying the constraint with equality.

The d-tilted information in $s \in \mathcal{M}$ [8], [9] can be defined as

$$\jmath_S(s, d) \triangleq \imath_{S;Z^\star}(s; z) + \lambda^\star \mathsf{d}(s, z) - \lambda^\star d, \tag{5}$$

where the usual information density is the logarithm of the Radon-Nikodym derivative of $P_{SZ^\star} = P_S P_{Z^\star|S}$ with respect to $P_S P_{Z^\star}$:[2]

$$\imath_{S;Z^\star}(s; z) \triangleq \log \frac{dP_{Z^\star S}}{dP_S \times dP_{Z^\star}}(s, z), \tag{6}$$

$z \in \text{supp}\,(P_{Z^\star})$, and

$$\lambda^\star \triangleq -\mathbb{R}'_S(d) \tag{7}$$

is the negative of the slope of rate-distortion function. Note that the value of the right side of (5) does not depend on the choice of $z \in \text{supp}\,(P_{Z^\star})$ [10], [11], i.e. it is a function of $s \in \mathcal{M}$ only. Throughout, to ensure that the notion of support is well-defined we assume that $\hat{\mathcal{M}}$ is a topological space.

An important property of the d-tilted information is that

$$\mathbb{R}_S(d) = \mathbb{E}\left[\jmath_S(S, d)\right]. \tag{8}$$

Our main tool is the following basic property of the probability kernel that achieves the minimum in the right side of (4).

Theorem 1. *Any $P_{Z|S}$ such that* [3]

$$\text{supp}\,(P_Z) \subseteq \text{supp}\,(P_{Z^\star}), \tag{9}$$

where P_Z is the output distribution induced by the code, satisfies

$$D(P_{S|Z} \| P_{S|Z^\star} | P_Z) = I(S; Z) - \mathbb{R}_S(d) + \lambda^\star \mathbb{E}\left[\mathsf{d}(S, Z)\right] - \lambda^\star d. \tag{10}$$

In particular, if $P_{Z|S}$ is such that

$$\mathbb{E}\left[\mathsf{d}(S, Z)\right] \leq d, \tag{11}$$

we may weaken (10) to conclude

$$D(P_{S|Z} \| P_{S|Z^\star} | P_Z) \leq I(S; Z) - \mathbb{R}_S(d). \tag{12}$$

Proof. Write

$$\begin{aligned} & I(S; Z) - D(P_{S|Z} \| P_{S|Z^\star} | P_Z) + \lambda^\star \mathbb{E}\left[\mathsf{d}(S, Z)\right] - \lambda^\star d \\ &= \mathbb{E}\left[\imath_{S;Z^\star}(S; Z)\right] + \lambda^\star \mathbb{E}\left[\mathsf{d}(S, Z)\right] - \lambda^\star d && (13) \\ &= \mathbb{E}\left[\jmath_S(S, d)\right] && (14) \\ &= \mathbb{R}_S(d), && (15) \end{aligned}$$

where to get (14) we used (5) and (9). □

Notice that (12) implies in particular that $D(P_{S|Z} \| P_{S|Z^\star} | P_Z) < \infty$ whenever $I(S; Z) < \infty$.

III. Block coding

A. Formal problem setup

This section treats block coding of a stationary ergodic source. The abstract single-shot setup in Section II specializes to $\mathcal{M} = \mathcal{A}^k$, $\hat{\mathcal{M}} = \hat{\mathcal{A}}^k$, $\mathsf{d}_k\colon \mathcal{A}^k \times \hat{\mathcal{A}}^k \mapsto [0, \infty]$. Going beyond separable distortion measures, we assume that the distortion measure is subadditive, i.e. that

$$\mathsf{d}_{n+m}\left((s_1^n, s_2^m), (z_1^n, z_2^m)\right) \leq \frac{n}{m+n} \mathsf{d}_n(s_1^n, z_1^n) + \frac{m}{m+n} \mathsf{d}_m(s_1^m, z_1^m). \tag{16}$$

We are interested in the asymptotic properties of $P_{Z^k|S^k}$ common to all codes that achieve the optimal rate-distortion

[1] We assume that $\mathbb{R}_S(d)$ is finite for some d.

[2] $P_{Z^\star}$ is the output distribution corresponding to the conditional probability distribution $P_{Z^\star|S}$ that achieves the rate-distortion function.

[3] This includes all codes $P_{Z|S}$ if $\text{supp}\,(P_{Z^\star}) = \hat{\mathcal{M}}$.

tradeoffs, in different senses we formalize next. We consider codes operating at a given average distortion:

$$\mathbb{E}\left[\mathsf{d}_k(S^k, Z^k)\right] \leq d. \tag{17}$$

We consider both fixed and average rate constraints:

(i) Fixed rate constraint: a fixed-length lossy code of rate R is a pair of random mappings $(P_{W|S^k}, P_{Z^k|W})$, where $W \in \{1, \ldots, \exp(kR)\}$.

(ii) Average rate constraint: a variable-length lossy code of average rate R is a pair of random mappings $(P_{W|S^k}, P_{Z^k|W})$, where $W \in \{1, 2, \ldots\}$, and $\mathbb{E}\left[\ell(W)\right] = kR$ where $\ell(w)$ is the length of the binary representation of integer w.

For a stationary ergodic source with subadditive distortion measure, the operational rate-distortion function, i.e. the minimum asymptotically achievable rate, fixed or average, compatible with average distortion d, satisfies [12, Lemma 10.6.2, Theorem 11.5.11]

$$R(d) = \inf_k \frac{1}{k}\mathbb{R}_{S^k}(d) = \lim_{k\to\infty} \frac{1}{k}\mathbb{R}_{S^k}(d), \tag{18}$$

where

$$\mathbb{R}_{S^k}(d) \triangleq \inf_{\substack{P_{Z^k|S^k}:\\ \mathbb{E}\left[\mathsf{d}_k(S^k, Z^k)\right]\leq d}} I\left(S^k; Z^k\right). \tag{19}$$

B. The main result

Our main result explores the restrictions imposed on $D(P_{S^k|Z^k}\|P_{S^k|Z^{k\star}}|P_{Z^k})$ by the constraints on rate and distortion.

Theorem 2. *Let $\{S^k\}_{k=1}^{\infty}$ be a stationary ergodic source with subadditive distortion measure. Let $P_{Z^{k\star}|S^k}$ achieve the infimum in (19). Let $\{P_{Z^k|S^k}\}_{k=1}^{\infty}$ be generated by a sequence of codes for average distortion d with rates R_k (fixed or average) such that $R_k \to R(d)$ as $k \to \infty$. Assume that*

$$\mathrm{supp}\left(P_{Z^k}\right) \subseteq \mathrm{supp}\left(P_{Z^{k\star}}\right). \tag{20}$$

Then,

$$\frac{1}{k} I(S^k; Z^k) \to R(d), \tag{21}$$

$$\frac{1}{k} D(P_{S^k|Z^k}\|P_{S^k|Z^{k\star}}|P_{Z^k}) \to 0. \tag{22}$$

Proof. An immediate consequence of (12) and (18) is

$$\frac{1}{k} D(P_{S^k|Z^k}\|P_{S^k|Z^{k\star}}|P_{Z^k}) \leq \frac{1}{k} I(S^k; Z^k) - R(d). \tag{23}$$

Note that (23) holds for all codes that meet the distortion constraint in (17), regardless of their rates.

To further upper-bound (23) when a rate constraint is imposed, we invoke the data processing inequality. For fixed-rate codes, we apply

$$I(S^k; Z^k) \leq kR_k \tag{24}$$

to the right side of (23) and use $R_k \to R(d)$ to obtain (21) and (22).

For variable-length codes, by data processing and [13, Lemma 3], for any $S - W - Z$ we have

$$\begin{aligned} I(S; Z) &\leq H(W) & (25)\\ &\leq \mathbb{E}\left[\ell(W)\right] + \log_2(\mathbb{E}\left[\ell(W)\right] + 1) + \log_2 e, & (26)\end{aligned}$$

and therefore

$$I(S^k; Z^k) \leq kR_k + \log_2(kR_k + 1) + \log_2 e. \tag{27}$$

Applying (27) to the right sides of (23) and using $R_k \to R(d)$ leads to both (21) and (22). □

For the special case of fixed-length lossy compression of a finite alphabet i.i.d. source with separable distortion, an alternative proof of (22) using the concept of coordination codes (introduced in [14]) follows from [7, (37) and Theorem 3].

Note that for any good deterministic code sequence, the entropy of the reproduction converges to the rate-distortion function:

$$\frac{1}{k} H(Z^k) \to R(d). \tag{28}$$

Indeed, $\frac{1}{k} H(Z^k) \leq R_k \to R(d)$ holds by the asymptotic optimality of the code, and $\frac{1}{k} H(Z^k) = \frac{1}{k} I(S^k; Z^k) \geq R(d)$ holds because the code is deterministic. Letting U_{Z^k} be the equiprobable distribution over the codebook, note that (22) and (28) imply that for any good deterministic code sequence,

$$\begin{aligned} D(P_{S^k Z^k}\|P_{S^k|Z^{k\star}} U_{Z^k}) &= \frac{1}{k} D(P_{S^k|Z^k}\|P_{S^k|Z^{k\star}}|P_{Z^k}) \\ &\quad + \frac{1}{k} D(P_{Z^k}\|U_{Z^k}) & (29)\\ &\to 0, & (30)\end{aligned}$$

an observation made by Schieler and Cuff [7] in the context of finite alphabet coordination codes.

On the other hand, Kanlis et al. [6, Proposition 2] demonstrated the existence of rate-distortion-achieving codes for the finite-alphabet i.i.d. source such that

$$\lim_{k\to\infty} \inf \frac{1}{k} D(P_{Z^k|S^k}\|P^k_{\mathsf{Z}^\star|\mathsf{S}}|P_{S^k}) \geq H(\mathsf{Z}^\star|\mathsf{S}), \tag{31}$$

thereby showing that, in contrast to (22),

$$\frac{1}{k} D(P_{Z^k|S^k}\|P_{Z^{k\star}|S^k}|P_{S^k}) \nrightarrow 0. \tag{32}$$

To shed some light onto why (31) holds, let us demonstrate the existence of a code sequence such that

$$\frac{1}{k} D(P_{Z^k}\|P^k_{\mathsf{Z}^\star}) \to H(\mathsf{Z}^\star|\mathsf{S}). \tag{33}$$

Then, (31) will follow by the data processing inequality. Consider a deterministic constant composition code with all codewords of type $P_{\mathsf{Z}^\star}$ (if the point masses of $P_{\mathsf{Z}^\star}$ are not multiples of $\frac{1}{k}$, consider instead the k-type closest to $P_{\mathsf{Z}^\star}$, in terms of Euclidean distance, viewing the probability mass function on a finite alphabet as a vector of length $|\mathcal{M}|$). It is known that such codes can achieve the rate-distortion function, see e.g. [8] for refined achievability results. Letting z^k be any output sequence of type $P_{Z^\star}$, write

$$\frac{1}{k} D(P_{Z^k} \| P_{\mathsf{Z}^\star}^k) = \frac{1}{k} \log \frac{1}{P_{Z^{k\star}}(z^k)} - \frac{1}{k} H(Z^k) \quad (34)$$
$$\to H(\mathsf{Z}^\star) - I(\mathsf{S}; \mathsf{Z}^\star) \quad (35)$$
$$= H(\mathsf{Z}^\star | \mathsf{S}), \quad (36)$$

where $\frac{1}{k} \log \frac{1}{P_{Z^{k\star}}(z^k)} \to H(\mathsf{Z}^\star)$ in (35) is by type counting, and $\frac{1}{k} H(Z^k) \to I(\mathsf{S}; \mathsf{Z}^\star)$ is due to (28).

C. Redundancy-optimal codes

The difference between the rate of the code and the rate-distortion function, $\Delta_k(d) \triangleq R_k - R_{S^k}(d)$, is referred to as the rate redundancy. Denote the minimum achievable rate redundancy among all codes operating at average distortion d and blocklength k by $\Delta_k^\star(d)$. A nonasymptotic refinement of (22) is the following: any redundancy-optimal code for a stationary ergodic source with subadditive distortion measure must satisfy, via (23) and (24),

$$\frac{1}{k} D(P_{S^k|Z^k} \| P_{S^k|Z^{k\star}} | P_{Z^k}) \leq \Delta_k^\star(d). \quad (37)$$

In fixed-length coding of an i.i.d. discrete source with a separable distortion measure, the work by Zhang et al. [15] implies that the minimum achievable rate redundancy $\Delta_k^\star(d)$ is equal to

$$\Delta_k^\star(d) = \frac{1}{2k} \log k + O\left(\frac{1}{k} \log \log k \right). \quad (38)$$

D. Codes operating under an excess distortion constraint

In lieu of the average distortion constraint (17), one might be interested in reproducing the source within distortion d, with high probability:

$$\mathbb{P}\left[\mathsf{d}_k(S^k, Z^k) > d \right] \leq \epsilon_k. \quad (39)$$

For bounded distortion measures (or more generally under a uniform integrability condition, see [16]) , convergence of the distortion in probability as $k \to \infty$ to d implies convergence of the average distortion to d and vice versa:

$$\mathsf{d}(S^k, Z^k) \xrightarrow{\mathbb{P}} d \Longleftrightarrow \mathbb{E}\left[\mathsf{d}(S^k, Z^k) \right] \to d. \quad (40)$$

Therefore, for bounded distortion measures, Theorem 2 continues to hold for sequences of codes that satisfy $\mathbb{P}\left[\mathsf{d}_k(S^k, Z^k) > d \right] \to 0$ instead of (17).

However, Theorem 2 need not hold if a small but nonvanishing excess-distortion probability is tolerated even as $k \to \infty$, i.e. if it is only asked that $\epsilon_k \leq \epsilon$. This behavior is similar to that of channel codes with nonvanishing average error probability [3]. To construct a simple counterexample, let $P_{\tilde{Z}^k|S^k}$ be a good lossy coder for the binary memoryless source such that the probability that the Hamming distance between the source and its representation exceeds $d < \frac{1}{2}$ is ϵ. Without loss of generality, we may assume that the all-zero vector 0^k and the all-one vector 1^k are not contained the codebook. Modify this code in the following way: if $\mathsf{d}(S^k, \tilde{Z}^k) \leq d$, output $\tilde{Z}^k$. If $\mathsf{d}(S^k, \tilde{Z}^k) > d$, output 0^k if the Hamming weight of S^k is exceeds $\frac{1}{2}k$ and output 1^k if the Hamming weight of S^k is less than or equal to $\frac{1}{2}k$. Denote the resulting conditional probability distribution by $P_{Z^k|S^k}$. Clearly,

$$\mathbb{P}\left[\mathsf{d}(S^k, Z^k) > d | Z^k \notin \{0^k, 1^k\} \right] = 0 \quad (41)$$
$$\mathbb{P}\left[\mathsf{d}(S^k, Z^k) \geq \tfrac{1}{2} | Z^k \in \{0^k, 1^k\} \right] = 1 \quad (42)$$
$$P_{Z^k}(\{0^k, 1^k\}) = \epsilon \quad (43)$$

Denote for brevity the set

$$\mathcal{E} \triangleq \left\{ (s^k, z^k) \in \mathcal{A}^k \times \hat{\mathcal{A}}^k \colon \mathsf{d}(s^k, z^k) \geq \tfrac{1}{2} \right\}. \quad (44)$$

Write

$$D(P_{S^k|Z^k} \| P_{S^k|Z^{k\star}} | P_{Z^k})$$
$$\geq d\left(P_{S^k Z^k | Z^k \in \{0^k, 1^k\}}(\mathcal{E}) \| P_{S^k Z^{k\star} | Z^{k\star} \in \{0^k, 1^k\}}(\mathcal{E}) \right) \epsilon \quad (45)$$
$$= \epsilon \log \frac{1}{P_{S^k Z^{k\star} | Z^{k\star} \in \{0^k, 1^k\}}(\mathcal{E})} \quad (46)$$
$$= k\, d\left(\tfrac{1}{2} \| d \right) \epsilon, \quad (47)$$

where (45) with $d(p\|q) \triangleq p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$ is by the data processing inequality and (43), (46) is due to (42) which holds by construction. Finally, (47) is by Cramér's large deviations theorem: conditioned on $Z^{k\star}$,

$$k\mathsf{d}(S^k, Z^{k\star}) = \sum_{i=1}^k 1\{S_i \neq Z_i^\star\} \quad (48)$$

has Bernouilli distribution with success probability d (e.g. [17]).

E. Lossy joint source-channel coding

Theorem 2 generalizes to the joint source-channel coding setup. In fixed-length JSCC, $S^k - X^n - Y^n - Z^k$, where $P_{Y^n|X^n}$ is fixed. In variable-length JSCC with feedback and termination, the encoder has access to Y^{n-1}, and the transmission stops when a special termination symbol is received, which is always decoded error-free [18]. We restrict attention to the discrete memoryless channel. The rates of 'good' JSCC for distortion d approach the asymptotic fundamental limit: $\frac{k}{n} \to \frac{C}{R(d)}$ as $k, n \to \infty$, where C is the channel capacity, or, for the variable-length setup, $\frac{k}{\ell} \to \frac{C}{R(d)}$ where ℓ is the average transmission time.

Theorem 3 (JSCC)**.** *Let $\{S^k\}_{k=1}^\infty$ be stationary ergodic source with subadditive distortion measure. Let $\{P_{Z^k|S^k}\}_{k=1}^\infty$ be generated by a sequence of good JSCC codes for average distortion d for the discrete memoryless channel (fixed length or variable length with feedback). Assume that (20) is satisfied. Then, both (21) and (22) must hold.*

Proof. For the fixed-rate setup, applying the data processing inequality

$$I(S^k; Z^k) \leq I(X^n; Y^n) \quad (49)$$
$$\leq nC \quad (50)$$

to the right side of (23) and taking the limit as $k, n \to \infty$ leads to both (21) and (22).

For variable-length coding over the DMC with feedback, we conclude from the proof of [18, Theorem 4] (replacing the right side of [18, (67)] by $I(S^k; Z^k)$) that

$$I(S^k; Z^k) \leq C\ell + \log(\ell + 1) + \log e. \tag{51}$$

Applying (27) to the right side of (23) and taking the limit as $k, \ell \to \infty$ leads to both (21) and (22). □

IV. Acknowledgement

The authors are grateful to Dr. Curt Schieler for helpful discussions.

References

[1] V. Kostina and S. Verdú, "Lossy joint source-channel coding in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 59, no. 5, pp. 2545–2575, May 2013.
[2] S. Shamai and S. Verdú, "The empirical distribution of good codes," *IEEE Transactions on Information Theory*, vol. 43, no. 3, pp. 836–846, 1997.
[3] Y. Polyanskiy and S. Verdú, "Empirical distribution of good channel codes with nonvanishing error probability," *IEEE Transactions on Information Theory*, vol. 60, no. 1, pp. 5–21, Jan. 2014.
[4] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, July and October 1948.
[5] T. Weissman and E. Ordentlich, "The empirical distribution of rate-constrained source codes," *IEEE Transactions on Information Theory*, vol. 51, no. 11, pp. 3718–3733, 2005.
[6] A. Kanlis, S. Khudanpur, and P. Narayan, "Typicality of a good rate-distortion code," *Problems of Information Transmission*, vol. 32, no. 1, pp. 96–103, 1996.
[7] C. Schieler and P. Cuff, "A connection between good rate-distortion codes and backward DMCs," in *Proceedings 2013 IEEE Information Theory Workshop*, Seville, Spain, Sep. 2013.
[8] V. Kostina and S. Verdú, "Fixed-length lossy compression in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 58, no. 6, pp. 3309–3338, June 2012.
[9] V. Kostina, "Lossy data compression: nonasymptotic fundamental limits," Ph.D. dissertation, Princeton University, Sep. 2013.
[10] R. Gallager, *Information theory and reliable communication*. John Wiley & Sons, Inc. New York, 1968.
[11] I. Csiszár, "On an extremum problem of information theory," *Studia Scientiarum Mathematicarum Hungarica*, vol. 9, no. 1, pp. 57–71, Jan. 1974.
[12] R. M. Gray, *Entropy and information theory*. Springer, 2011.
[13] N. Alon and A. Orlitsky, "A lower bound on the expected length of one-to-one codes," *IEEE Transactions on Information Theory*, vol. 40, no. 5, pp. 1670–1672, 1994.
[14] P. W. Cuff, H. H. Permuter, and T. M. Cover, "Coordination capacity," *IEEE Transactions on Information Theory*, vol. 56, no. 9, pp. 4181–4206, 2010.
[15] Z. Zhang, E. Yang, and V. Wei, "The redundancy of source coding with a fidelity criterion," *IEEE Transactions on Information Theory*, vol. 43, no. 1, pp. 71–91, Jan. 1997.
[16] T. S. Han, *Information-Spectrum Methods in Information Theory*. Springer, Berlin, 2003.
[17] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.
[18] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Feedback in the non-asymptotic regime," *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 4903–4925, 2011.