Fifty-third Annual Allerton Conference
Allerton House, UIUC, Illinois, USA
September 29 - October 2, 2015

Excess Distortion in Lossy Compression: Beyond One-Shot Analysis

Yuval Kochman
Hebrew University of Jerusalem
Jerusalem, Israel
Email: yuvalko@cs.huji.ac.il

Gregory W. Wornell
Massachusetts Institute of Technology
Cambridge, MA 02139
Email: gww@mit.edu

***Abstract*—The problem of finite-blocklength lossy compression under an excess-distortion constraint is considered. If the blocklength constraint comes from the length of the source sequence itself, the excess rate needed above the rate-distortion function decays inversely proportional to the square root of the blocklength, according to second-order (dispersion) analysis. We consider a different case, where the source emits a long sequence, but shorter sub-sequences are considered for reasons such as delay, complexity and smoothness of the reconstruction fidelity. We analyze the redundancy of the rate with respect to different constraints. We show that the rate redundancy with respect to the processing blocklength, i.e. the dimension of the quantizer used, decays much faster than the dispersion analysis suggests. Thus, one may use much shorter source codes without sacrificing second-order performance.**

I. Introduction

Imagine that a source sequence needs to be compressed, in order to be recovered subject to some fidelity criterion. It is well known that in an asymptotic sense, the optimal compression rate is given by the rate-distortion function (RDF). There may be several constraints that prohibit approaching this asymptotic limit.

1) Finite source block. It may be that the source only emits a sequence of some finite length ℓ.
2) Finite processing block. Due to delay and/or complexity constraints, only k source samples can be processed simultaneously.
3) Finite fidelity block. In order for the reconstruction to be useful, the distortion has to be low enough when averaging over n consecutive source samples.
4) Finite resource block. It may be that the rate constraint has to be enforced over some finite horizon, i.e., for any A consecutive processing blocks, representing $m = Ak$ source samples, the encoder will use at most mR bits.

Having understood that various blocklengths constrain performance, it is only natural to ask: which of these constraints limits performance? In this work we further develop the framework presented in [1], pointing at a rather surprising conclusion: in many cases, it is *not* the processing blocklength k. More precisely, the gap from the RDF due to finite processing blocklength decays as $\log k/k$, while with respect to the other constraints it may decay much slower, inversely proportional to the square root of the blocklength.

Specifically, we make a distinction between two fundamentally different scenarios. In a one-shot scenario, the source block ℓ is short enough, such that the other constraints play no role. Thus, without loss of generality one can assume $k = n = m = \ell$. In this setting, the excess-distortion asymptotics are well approximated by the dispersion [2], [3]: for distortion threshold D, excess-distortion probability ϵ and blocklength ℓ, the required coding rate is given by

$$R = R(D) + \sqrt{\frac{V(D)}{\ell}} Q^{-1}(\epsilon) + O\left(\frac{\log \ell}{\ell}\right), \quad (1)$$

where $R(D)$ and $V(D)$ are the source RDF and dispersion, respectively, and $Q^{-1}(\cdot)$ is the inverse Gaussian composite distribution function. Large-deviation type asymptotics have been derived decades earlier in [4].

Consider, on the other hand, a "many-shots" scenario, where the source blocklength $\ell \gg k, n, m$. The source produces a very long sequence, which is parsed by the encoder into many processing blocks, and also contains many fidelity and resource blocks. We note that it is a very important scenario in practice. Consider the compression of a long video, for example.

Practical encoders will not process the whole video jointly, but use much shorter processing blocks. Also, the fidelity is not measured over the whole source block: lost frames in one part of the video cannot be compensated for by excellent-quality reproduction in another part. If, in addition, the video is to be streamed over a communication channel with constraint over the data throughput over some time window, then the high coding rate of one part cannot be balanced with a low rate for another.

In order to facilitate tractable analysis, we ignore edge effects of the source sequence, and take in this regime the source blocklength ℓ to be infinite. Also, as a first stage we make the simplifying assumption of *synchroneous* fidelity blocks, where either the fidelity block is composed of an integer number of processing blocks, or vice versa. Within the remaining blocklengths k, m, n we find that the following tradeoff is achievable:

$$R = R(D) + \sqrt{\frac{V(D)}{\overline{n}}} Q^{-1}(\epsilon) + O\left(\frac{\log \underline{n}}{\underline{n}}\right), \quad (2)$$

where

$$\begin{aligned} \overline{n} &\triangleq \max(m, n) \\ \underline{n} &\triangleq \min(k, n). \end{aligned} \quad (3)$$

Perhaps the most important conclusion is, that the processing blocklength k never appears in the dispersion $(1/\sqrt{\overline{n}})$ term, but only in the logarithmic term which decays much faster. We see that if one considered designing a quantization scheme guided by (1), the required quantizer dimension k can now be reduced to be in the order of a square root of the estimate given by dispersion!

Within this scenario, we highlight two cases. In the first, $n \gg k = m$, the coding rate is strictly fixed for all processing blocks, but the fidelity is measured over a larger blocklength. Here,

$$R = R(D) + \sqrt{\frac{V(D)}{n}} Q^{-1}(\epsilon) + O\left(\frac{\log k}{k}\right)$$

can be achieved and is, in fact, optimal. Further, we know that under *average* distortion constraint, [5]

$$R \cong R(D) + \frac{\log k}{2k}. \quad (4)$$

Since this can be seen as the limit of an excess-distortion constraint with $n \to \infty$, (2) can be seen as a bridge between (1) and (4).

Next, consider the case $m \gg k = n$, where the distortion is measured over a blocklength which equals the processing one, but the rate constraint is more flexible. We have that

$$R \cong R(D) + \sqrt{\frac{V(D)}{m}} Q^{-1}(\epsilon) + O\left(\frac{\log k}{k}\right)$$

is achievable. Here, we make no claim for optimality. Specifically, with an average rate constraint, which can be seen as $m \to \infty$, and with zero excess-distortion probability, the behaviour was shown in [5] to be similar to (4); however, this is only with zero excess-distortion probability. Allowing excess distortion, using more elaborate rate allocation strategies, as was done by Kostina et al. [6] in the one-shot setting under an average rate constraint,[1] may improve the bounds even further.

From an operational point of view, synchronous fidelity blocks do not give the whole picture: in order for them to have an operational meaning, the fidelity measurement has to be aware of the parsing of the source sequence into processing blocks, which seems rather unjustified. Thus, we consider the corrections needed when considering *asynchronous* fidelity measurements. It can be easily shown that for $n > k$ these are at most $O(k/n)$,[2] thus our main conclusions, concerning the regime where $n \gg k$, remain unchanged.

The rest of this paper is organized as follows. After making some definitions, we turn in Section III to describe a universal encoder that will serve as our main building block. In Section IV we illustrate our ideas using an example where some of the technical difficulties are not present. In Section V we present our main results for a simplified, synchronized, setting, and then in Section VI we address their extension to the unsynchronized setting. Finally in Section VII we discuss the limitations of the analysis and possible extensions.

II. Definitions

The source is an infinite i.i.d. sequence $\ldots, X_0, X_1, \ldots$ where the symbols belong to some alphabet $\mathcal{X}$ and have some distribution P. The encoder is a function from $\mathcal{X}^k$ to an index, applied to processing blocks $X_{ak+1}, \ldots, X_{(a+1)k}$ for integer a. It is convenient to think of the index as a sequence of bits of length $R_a \cdot k$, where R_a is the rate of the

[1] Although the justification for considering average rate in a one-shot setting is not clear.

[2] We can also show an achievable rate with a correction of $O\left(k/n^{3/2}\right)$, see Section VI.

a-th block.[3] Let ζ be some positive integer. Then R_a is a function of the source within the a-th processing block and of past encoder cardinalities, chosen to satisfy the rate constraint:

$$\sum_{b=1}^{\zeta} kR_{A\zeta+b} \le \zeta kR \triangleq mR$$

for any integer A. The decoder is a function from the index to $\hat{\mathcal{X}}^k$, used to reconstruct an infinite sequence $\ldots, \hat{X}_0, \hat{X}_1, \ldots$ by placing each reconstructed processing block in the location of the original block. The fidelity is measured using a single-letter function $d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$, averaged over blocks:

$$d_{n,n_0}(X, \hat{X}) = \frac{1}{n} \sum_{i=n_0+1}^{n_0+n} d(X_i, \hat{X}_i). \tag{5}$$

The parsing into processing and fidelity blocks is demonstrated in Figure 1. We make a distinction between two settings:

1) Synchronized fidelity blocks. This model is suitable for, e.g., compress-and-forward relaying, where different blocklength constraints come from the delay and complexity requirements at the relay and the final destination, as well as the required smoothness of transmission rate over the channel. In this setting, either $n = \rho k$ for some integer $\rho \ge 1$, in which case the fidelity block spans ρ whole processing blocks, or $\rho = 1/\mu$ where $\mu \ge 1$ is an integer, in which case each processing block is composed of μ whole fidelity blocks.
2) Unsynchronized fidelity blocks. This model is suitable for the video compression problem described in the introduction. Since the distortion measurement should be oblivious of the parsing into processing blocks, the excess-distortion probability of a scheme with observation blocklength n is averaged over the offset between the start of processing and fidelity blocks:

$$p_e(n, D) = \frac{1}{k} \sum_{n_0=0}^{k-1} \Pr\left\{ d_{n,n_0}(X, \hat{X}) > D \right\}. \tag{6}$$

As the synchronized setting is easier to analyze, it is assumed until Section VI. As will become clear later, the two settings differ mostly when n and k are close.

The rate-distortion function (RDF) of a source with an i.i.d. distribution Q is given by

$$R(Q, D) = \min_{W(\hat{X}|X): E_{Q,W} d(X,\hat{X}) \le D} I(X; \hat{X}). \tag{7}$$

[3]We ignore everywhere the fact that this length may not be an integer, as it has no effect on the order of apprximation of interest.

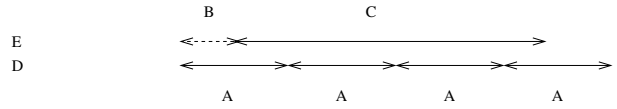

Fig. 1: Unsynchronized fidelity blocks: schematic view of the processing and fidelity blocks.

If Q is the source distribution P, we simply write $R(D)$. The inverse function is denoted by $D(R)$ or $D(Q, R)$.[4] We call a source-distortion pair (P, D) a *regular point* if $R(P, D)$ is differentiable with respect to D and twice differentiable with respect to P in a neighborhood of (P, D). For a regular point, the source dispersion is given by

$$V(P, D) = \mathrm{Var}\left\{ \left. \frac{\partial R(Q, D)}{\partial Q_i} \right|_{Q=P} \right\}, \tag{8}$$

where Q_i are the elements of the distribution Q.[5]

Our analysis is based on the method of types. We adopt the notation of Csiszár and Körner [7]: The *type* of a sequence $\mathbf{x} \in X^n$ is the vector Q over $\mathcal{X}$ whose elements are the relative frequencies of the alphabet letters in $\mathcal{X}$. $\mathcal{T}_n(\mathcal{X})$ denotes all the types of sequences in $\mathcal{X}^n$. The *type class* of the type $Q \in \mathcal{T}_n(\mathcal{X})$, denoted T_Q, is the set of all sequences $\mathbf{x} \in \mathcal{X}^n$ with type Q. For a reconstruction word $\hat{\mathbf{x}} \in \hat{\mathcal{X}}$, we say that $\mathbf{x}$ is D-covered by $\hat{\mathbf{x}}$ if $d(\mathbf{x}, \hat{\mathbf{x}}) \le D$.

III. A Simple Universal Encoder

In this section we present a simple encoder for a one-shot scenario with blocklength ℓ, which will serve as our main building block in the sequel. It is nearly optimal universally with respect to the source type. We start, then, with the optimal performance for a given type, following the formuation in [2].

Proposition 1 (Type covering): Let $Q \in \mathcal{T}_\ell(\mathcal{X})$ with a corresponding type class T_Q. Let $A(Q, \mathcal{C}, D)$ be the intersection of T_Q with the set of source sequences $\mathbf{x} \in \mathcal{X}^\ell$ which are D-covered by at least one of the words in a codebook $\mathcal{C}$ with rate R (i.e. $|\mathcal{C}| = e^{\ell R}$). Assume that (Q, D) is a regular point. Then there exists a constant $J = J(|\mathcal{X}|, |\hat{\mathcal{X}}|)$ such that:

1) There exists a codebook C_Q that completely D-covers T_Q (i.e. $A(Q, \mathrm{C}_Q, D) = T_Q$), where for

[4]If the inverse is not unique at rate R, we take the lowest distortion satisfying the equality.

[5]These derivatives point in general outside the simplex. They are defined with respect to the following function of the RDF (7): Q is not necessarily a probability distribution but $W(\hat{X}|X)$ is still a valid conditional distribution. Expectation and mutual information are still given as sums weighted by the elements of Q.

large enough ℓ,

$$\frac{1}{\ell}\log|\mathbf{C_q}| = R \le R(Q,D) + J\frac{\log\ell}{\ell}. \quad (9)$$

2) If $R(Q,D) > R$, the fraction of the type class that is D-covered by any code with rate R is bounded by

$$-\frac{1}{n}\log\frac{|A(Q,\mathcal{C}_n,D)|}{|T_Q|} \ge R(Q,D) - R + J\frac{\log\ell}{\ell}. \quad (10)$$

The first part of this proposition is a refinement of Berger's type-covering lemma [8], found in [9]. The second part is a corollary of [5, Lemma 3]. The universal performance is as follows.

Proposition 2 (Type-universal encoder): For any finite alphabets $\mathcal{X}$, $\hat{\mathcal{X}}$ there exists a constant $\tilde{J} = \tilde{J}(|\mathcal{X}|,|\hat{\mathcal{X}}|)$ such that for any large enough blocklength ℓ and for any $\{R_Q, D_Q\}$ defined on $Q \in \mathcal{T}_\ell(\mathcal{X})$ satsfying

$$R_Q \le R(Q,D_Q) + \tilde{J}\frac{\log\ell}{\ell}. \quad (11)$$

there exists a *single* codebook satisfying:

- It is uniquely decodable.
- If the source type is Q then the distortion is at most D_Q.
- If the source type is Q then the length is at most $\ell \cdot R_Q$.

This codebook is just a union of type-covering codes as in the direct part of Proposition 1, with a prefix that identifies the type. We can thus satisfy the conditions with:

$$\tilde{J} = J + \frac{\ell}{\log\ell}\cdot\frac{\log|\mathcal{T}_\ell(\mathcal{X})|}{\ell} \le J + |\mathcal{X}| + 1.$$

In light of the second part of Proposition 1, these universal codes have a rate penalty of $O(\log\ell/\ell)$ over the optimal code that knows the source type in advance.

We note that two particular variants of the encoder are of special interest: a *fixed-distortion* encoder where $D_Q = D$ for all Q and a *fixed-rate* encoder where $R_Q = R$ for all Q.

IV. A Motivating Example

We take the following example, which suffices to capture a lot of the essence. It is a version of "erasure" source/distortion. The source and reproduction are both ternary: $\mathcal{X} = \hat{\mathcal{X}} = \{0,1,E\}$. The source is 0 or 1 with probability $p/2$, E with probability $1-p$. The distortion measure is: $d(x,\hat{x}) = 1$ if $x \ne E$ and $y = E$, $d(x,\hat{x}) = 0$ if $x = E$ or $x = \hat{x}$, infinite distortion otherwise. That is, source symbols can be "erased" with unit cost, except when the source has the "don't care" value E, then it never suffers from distortion.

The rate-distortion function is simply

$$R(p,D) = p - D.$$

That is, it takes 1 bit to accurately describe a source symbol that matters. It is the linearity of this function that makes this source convenient; in the general case, we will need linearizations. Note also that for any distribution P over the ternary input alphabet with $P(0)+P(1) = p$, $R(P,D) \le R(p,D)$. Thus, in Proposition 2 we can replace $R(Q,D)$ by $R(q,D)$ where q is the empirical portion of source symbols that are not E.[6] Namely, we have for the fixed-rate and fixed-distortion variants of the universal encoder, and for distortion measured over the processing block:

$$R \le q - D_q + \tilde{J}\frac{\log k}{k} \quad (12)$$

$$R_q \le q - D + \tilde{J}\frac{\log k}{k}. \quad (13)$$

We analyze this source in synchronized setting, in three cases; together, they reflect the techniques used to achieve (2).

First consider the case $n = \rho k$ for $\rho > 1$, and assume that $m = k$; that was the case addressed in [1]. That is, the code has to be strictly fixed-rate, but the fidelity block includes many processing blocks. Under this, (6) becomes:

$$p_e(n,D) = \Pr\left\{\frac{1}{\rho}\sum_{j=1}^{\rho} D_{q_j} > D\right\},$$

where D_{q_j} is given by (12) with q_j being the portion of non-E symbols in the j-th processing block. Substituting (12) and denoting by

$$q^\rho = \frac{1}{\rho}\sum_{j=1}^{\rho} q_j,$$

we have:

$$p_e(n,D) = \Pr\left\{q^\rho > R + D + \tilde{J}\frac{\log k}{k}\right\}.$$

The source of improvement over one-shot analysis is now evident: In (1), the main redundancy term, the dispersion proportional to $1/\sqrt{n}$, is due to the

[6] This can also be shown to be optimal to the order of interest, as the dispersion "with respect to" the empirical distribution between the input letters 0 and 1 is zero.

variation of the source type; we have now managed to identify the relevant type q^ρ measured over the fidelity blocklength n, rather than the individual $\{q_j\}$ measured over the processing blocks. We thus use now CLT analysis over the fidelity block. Loosely speaking, q^ρ is approximately Gaussian with mean q and variance $V(p,D)/n$, where

$$V(p,D) = p(1-p). \tag{14}$$

More precisely, in order to ensure that $p_e(n,D) \leq \epsilon$ for a fixed ϵ, we need:[7]

$$R = p - D + \sqrt{\frac{V(p)}{n}} Q^{-1}(\epsilon) + O\left(\frac{\log k}{k}\right) + O\left(\frac{1}{n}\right). \tag{15}$$

Second, let $m = \zeta k$ for $\zeta > 1$, but $n = k$. Now, within the resource block, for any processing block we use the fixed-distortion encoder (13), until we have exhausted our rate budget ζR. When that happens, we can allocate zeros rate and suffer excess distortion in the remaining processing blocks. We make a very severe assumption that if the rate was exhausted at some point, all of the processing blocks within the resource block had excess distortion. Even with this, we have:

$$\begin{aligned} p_e(n,D) &\leq \Pr\left\{\frac{1}{\zeta}\sum_{j=1}^{\zeta} R_{q_j} > R\right\} \\ &\leq \Pr\left\{q^\zeta > R + D + \tilde{J}\frac{\log k}{k}\right\}. \end{aligned}$$

Applying again dispersion analysis, this time over the resource blocklength m, we can have $p_e(n,D) \leq \epsilon$ where:

$$R = p - D + \sqrt{\frac{V(p)}{m}} Q^{-1}(\epsilon) + O\left(\frac{\log k}{k}\right) + O\left(\frac{1}{m}\right). \tag{16}$$

Finally, consider the case where $m = k = \mu n$ for integer $\mu > 1$, i.e., the fidelity block is short. Obviously, one can work with encoders of dimension n and distortion D, as long as the rate constraint is kept over the original blocklength k. Thus, we can have the performance of (16), substituting k by n:

$$R = p - D + \sqrt{\frac{V(p)}{m}} Q^{-1}(\epsilon) + O\left(\frac{\log n}{n}\right) + O\left(\frac{1}{m}\right). \tag{17}$$

[7]

V. Main Result with Synchronized Fidelity

In this section we prove the achievable rate (2), and also show its optimality for the case $n > m = k$, in the synchronized setting. A main technical ingredient is the following.

Lemma 1: Let $f(Q)$ be some function of distributions over a finite alphabet $\mathcal{X}$. Let P be the source distribution and D be a distortion lebel, such that (P,D) is a regular point. Let $\overline{n} = \rho\underline{n}$ where $\rho \in \{1,2,\ldots\}$. Let $X^{\overline{n}}$ be a sequence drawn i.i.d.-P. Denote by Q_a, $a = 1,\ldots,A$ the types of the subsequence $X_{(a-1)A+1},\ldots,X_{aA}$. Let

$$g_a = f(Q_a) + \delta(\underline{n}).$$

Let ϵ be a given probability, and let $\Delta = \Delta(\underline{n},\overline{n})$ be chosen such that

$$\Pr\left\{\left[\frac{1}{A}\sum_{a=1}^{A} g_a - f(P)\right] > \Delta\right\} = \epsilon.$$

Then, as $\overline{n}$ and $\underline{n}$ grow,

$$\Delta = \sqrt{\frac{V}{\overline{n}}} Q^{-1}(\epsilon) + \delta(\underline{n}) + O\left(\frac{\log \underline{n}}{\underline{n}}\right).$$

Furthermore, the same result holds if ϵ is replaced by $\epsilon + g(n)$ as long as $g(n) = O(\log \underline{n}/\sqrt{\underline{n}})$.

We do not give here the proof; it is a straightforward (though tedious) generalization of [10, Theorem 1], which itself is the generalization of the analysis of [2] used to describe "dispersion-like" behavior of any function of distributions satisfying some technical conditions. The Lemma above is an adaptation to a setting where the distribution is drawn over a block rather than per sample. We are now ready to state our main result.

Theorem 1: Consider the synchronized fidelity setting, with blocklengths k, m, n. Let $\overline{n}$ and $\underline{n}$ be defined according to (3). Assume that (P,D) is a regular point. Then there exists a sequence of encoders and decoders satisfying the constraints, indexed by k, m, n, with rates satisfying:

$$R(k,m,n) = R(P,D) + \sqrt{\frac{V(P,D)}{\overline{n}}} Q^{-1}(\epsilon) + O\left(\frac{\log \underline{n}}{\underline{n}}\right). \tag{18}$$

Proof: Let $A = \overline{n}/\underline{n}$. We use A times an encoder with blocklength $\underline{n}$. We use the universal encoder of Proposition 2, with one of the following two modes of operation:

1) If $n \geq m$ we use a universal fixed-rate scheme of rate R, which achieves at the a-th block a distortion D_a satisfying:

$$R = R(Q_a, D_a) + \tilde{J}\frac{\log \underline{n}}{\underline{n}}, \tag{19}$$

where Q_a is the type measured over the a-th block. The processing and resource constraints are satisfied trivially. In order to have excess-distortion probability below ϵ, we require:

$$\Pr\left\{\frac{1}{A}\sum_{a=1}^{A} D_a > D\right\} \leq \epsilon. \tag{20}$$

2) Otherwise ($m > n$) we use a universal fixed-distortion scheme with distortion D and rate satisfying:

$$R_a = R(Q_a, D) + \tilde{J}\frac{\log \underline{n}}{\underline{n}}. \tag{21}$$

The processing and fidelity constraints are satisfied trivially. In order to satisfy the resource constraint, we check the cumulative length; if it will pass the allowed length with the new block, we stop the process. In that case, we make a worst-case assumption that any measurement over the block results in an excess-distortion event. Thus, we can bound the excess-distortion probability by ϵ, as long as:

$$\Pr\left\{\frac{1}{A}\sum_{a=1}^{A} R_a > R\right\} \leq \epsilon. \tag{22}$$

Now we would like to use Lemma 1. For (22) this is a direct process. For (20), however, we need a linearization of (19):

$$\begin{aligned} D_a &= R^{-1}\left(Q_a, R - \tilde{J}\frac{\log \underline{n}}{\underline{n}}\right) \\ &= R^{-1}\left(Q_a, R\right) - O\left(\frac{\log \underline{n}}{\underline{n}}\right) \end{aligned}$$

and then we have for the function

$$f(Q_a) = R^{-1}\left(Q_a, R\right).$$

Since

$$\frac{\partial f(Q_a)}{\partial Q_{ai}} = \frac{\partial R(P,D)/\partial Q_{ai}}{\partial R(P,D)/\partial D},$$

we have the dispersion:

$$V_f = \frac{V(P,D)}{(\partial R(P,D)/\partial D)^2}.$$

Applying Lemma 1, the distortion threshold is given by:

$$\begin{aligned} D = R^{-1}(P,R) + \sqrt{\frac{V(P)}{\overline{n}}} \cdot \frac{Q^{-1}(\epsilon)}{\partial R(P,D)/\partial D} \\ + O\left(\frac{\log \underline{n}}{\underline{n}}\right). \end{aligned}$$

Taking the RDF of both sides, the proof is completed. ■

The following states when the above rate is optimal.

Theorem 2: In the synchronized fidelity setting with $n > m = k$, for any sequence of encoders and decoders satisfying the constraints with excess distortion probability at most ϵ, the optimal rate grows according to (18), i.e.,

$$R(k,m,n) = R(P,D) + \sqrt{\frac{V(D)}{n}}Q^{-1}(\epsilon) + O\left(\frac{\log k}{k}\right).$$

Proof: Consider a genie-aided scenario, where at each processing block $a = 1, \ldots, A$ the decoder in informed of the source type Q_a. Following the converse proof in [2], if

$$R(Q_a, D) > R + (J+1)\frac{\log k}{k}$$

then the probability of *not* having an excess distortion event is at most $1/n$. Now we apply Lemma 1 with error probability $\epsilon - 1/n$, with linearizations as in the proof of Theorem 1 to get the desired result. ■

VI. Correction for Unsynchronized Fidelity

Until now we considered a synchronized fidelity setting, where the distortion measurement is aligned with the processing blocks. Although this is a convenient assumption, it has no justification from an operational point of view: after the reconstructed blocks are pasted back together by the decoder, the user should be oblivious of parsing. Thus, we should measure the excess-distortion probability according to (6).

How will the results change, then? A trivial way to bound $d(n, n_0)$ is to round the number of blocks. Let $\rho = n/k$ be a real number, then

$$\sum_{i=1}^{\lfloor \rho \rfloor} d_i \leq d(n, n_0) \leq \sum_{i=1}^{\lceil \rho \rceil + 1} d_i$$

where $\{d_i\}$ are distortions measured over independent processing blocks. Thus, the results of the previous section hold up to a correction of

$$O\left(\frac{1}{\rho}\right) = O\left(\frac{k}{n}\right).$$

We note that this is sufficient in order extend our main observation to the unsynchronized fidelity setting. Specifically, take $n > m = k$ and ask, how should the processing blocklength k grow in order to get the same rate redundancy as obtained by the traditional (one-shot) dispersion analysis where $k = n$? The rate redundancy is at most

$$O\left(\frac{1}{\sqrt{n}}\right) + O\left(\frac{\log k}{k}\right) + O\left(\frac{k}{n}\right).$$

Taking $k = O(n^{\alpha+1/2})$ for any $\alpha > 0$ would give redundancy

$$O\left(n^{\frac{1}{2}-\alpha}\right),$$

which can be made close to $O(1/k)$ by taking small α.

However, one can do better than this crude analysis. Consider the part of the fidelity block which is not contained in full processing blocks; this part consists of samples from (at most) two processing blocks. By arguments of coordinate randomization, an achievable performance is given by a process of random sampling from finite populations (the population being the per-symbol distortion over a processing block). For such a setting, a central-limit-theorem like analysis is known to hold (see, e.g., [11]). Thus the dispersion with respect to that part of the distortion may vary, leading to a correction term of the form

$$O\left(\frac{1}{\rho}\cdot\frac{1}{\sqrt{n}}\right) = O\left(\frac{k}{n^{3/2}}\right).$$

Further quantification of this correction term is the subject of a current research.

VII. Conclusion

In this work we have shown, that when compressing long source sequences, careful consideration of the different system constraints yields meaningful tradeoffs. From a system designer point of view, it was shown that the dimension of vector quantizer needed for a given distortion threshold and excess-distortion probability is much lower than that suggested by dispersion analysis.

We acknowledge that the results are not complete yet. More clever rate allocation may further improve performance, when the resource blocklength is longer than the processing blocklength. Furthermore, more careful analysis is needed in order to refine the analysis of the asynchronous setting. Also, the analysis need not be restricted to fixed excess-distortion probability, and large-deviation (exponent) results are of interest as well.

We note that at this stage the contribution has limited practical implication, as vector quantization is rarely applied to memoryless sources. In this respect, future extension of the work to sources with memory is highly needed. Also, the analysis is limited to schemes operating on source blocks. A natural extension would be to consider sequential schemes, limited by delay; this may have a dramatic effect, as is the case for feedback communications [12].

Beyond source coding, similar distinction between different constraints can be applied to channel and joint source-channel coding as well, shedding light on basic tradeoffs in communications. In fact, some initial joint source-channel results [13] precede the current work.

References

[1] Y. Kochman and G. W. Wornell. Lossy compression with a short processing block: Asymptotic analysis. In *IEEEI 2014, Eilat, Israel*, Dec.. 2014.

[2] A. Ingber and Y. Kochman. The dispersion of lossy source coding. In *Proc. of the Data Compression Conference*, Snowbird, Utah, March 2011.

[3] V. Kostina and S. Verdú. Fixed-length lossy compression in the finite blocklength regime. *IEEE Trans. Info. Theory*, 58(6):3309–3338, June 2012.

[4] K. Marton. Error exponent for source coding with a fidelity criterion. *IEEE Trans. Info. Theory*, IT-20:197–199, Mar. 1974.

[5] Z. Zhang, E.H. Yang, and V. Wei. The redundancy of source coding with a fidelity criterion - Part one: Known statistics. *IEEE Trans. Info. Theory*, IT-43:71–91, Jan. 1997.

[6] V. Kostina, Y. Polyanskiy, and S. Verdú. Variable-length compression allowing errors. In *ISIT 2014, Honolulu, HI*, July 2014.

[7] I. Csiszár and J. Korner. *Information Theory - Coding Theorems for Discrete Memoryless Systems*. Academic Press, New York, 1981.

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

[9] B. Yu and T.P. Speed. A rate of convergence result for a universal D-semifaithful code. 39(3):813 –820, may. 1993.

[10] A. Ingber, D. Wang, and Y. Kochman. Dispersion theorems via second order analysis of functions of distributions. In *46th Annual Conference on Information Sciences and Systems (CISS)*, Mar. 2012.

[11] T. Hoglund. Sampling from a finite population: a remainder term estimate. *Scandinavian Journal of Statistics*, pages 69–71, 1978.

[12] A. Sahai. Why do block length and delay behave differently if feedback is present? *IEEE Trans. Info. Theory*, 54(5):1860–1886, May 2008.

[13] Y. Kochman and G. W. Wornell. On uncoded transmission and blocklength. In *Information Theory Workshop (ITW), Lausanne, Switzerland*, Sep. 2012.