When is Shannon's lower bound tight at finite blocklength?

Victoria Kostina

***Abstract*—This paper formulates an abstract version of Shannon's lower bound that applies to abstract sources and arbitrary distortion measures and that recovers the classical Shannon lower bound as a special case. A necessary and sufficient condition for it to be attained exactly is presented. It is demonstrated that whenever that condition is met, the d-tilted information of the source adopts a simple, explicit representation that parallels Shannon's lower bound. That convenient representation simplifies the non-asymptotic analysis of achievable rate-distortion tradeoffs. In particular, if a memoryless source meets Shannon's lower bound with equality, then its rate-dispersion function is given simply by the varentropy of the source.**

***Index Terms*—Lossy source coding, rate-distortion function, Shannon's lower bound, finite blocklength regime, dispersion.**

I. Introduction

In the compression of a memoryless source with single-letter distribution P_X under a single-letter distortion measure $\mathsf{d}(\cdot,\cdot)$, the minimum achievable source coding rate $R(n,d,\epsilon)$ comparable with blocklength n and the probability ϵ of exceeding distortion d given by [1]

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

where Q is the complementary Gaussian cdf, $R(d)$ is the rate-distortion function of the source:

$$R(d) = \inf_{\substack{P_{Y|X}:\ \mathcal{X}\mapsto\mathcal{Y}\\ \mathbb{E}[\mathsf{d}(X,Y)]\le d}} I(X;Y), \quad (2)$$

and $\mathcal{V}(d)$ is a parameter we termed the rate-dispersion function. That parameter quantifies the overhead over the rate-distortion function incurred by the finite blocklength constraint. Dropping the remainder term in (1), we obtain a simple approximation to the minimum achievable coding rate. That approximation provides good accuracy even at short blocklengths, as evidenced by the numerical results in [1].

The rate-distortion and the rate-dispersion function are given by the mean and the variance of $\jmath_X(X,d)$, the d-tilted information, the random variable which is defined as

$$\jmath_X(x,d) \triangleq \log \frac{1}{\mathbb{E}\left[\exp\left\{\lambda^\star d - \lambda^\star \mathsf{d}(x,Y^\star)\right\}\right]}, \quad (3)$$

V. Kostina is with California Institute of Technology (e-mail: vkostina@caltech.edu). This work was supported in part by the National Science Foundation (NSF) under Grant CCF-1566567, and by the Simons Institute for the Theory of Computing.

where $\lambda^\star = -R'(d)$, and the expectation is with respect to the unconditional distribution of $Y^\star$, the random variable that attains the rate-distortion function, i.e. $R(d) = I(X;Y^\star)$. Thus, both the rate-dispersion and the rate-distortion function are described in terms of the solution to the convex optimization problem in (2). Although the convexity of the problem in (2) often allows for an efficient numerical computation of its optimum [2], closed-form expressions are rarely available.

The absence of an explicit expression for the d-tilted information motivates a closer look into the behavior of (3). This paper shows a necessary and sufficient condition for

$$\jmath_X(x,d) = \log\frac{1}{f_X(x)} - \phi(d) \quad (4)$$

to hold, where $\phi(d)$ is a term that depends only on the distortion measure and distortion threshold d, and f_X is the probability density function defined with respect to an appropriate base measure. For continuous X, the base measure can be taken to be the Lebesgue measure, and f_X then is the usual probability density function. For discrete sources, f_X in (4) particularizes to the probability mass function of X:

$$\jmath_X(x,d) = \log\frac{1}{P_X(x)} - \phi(d) \quad (5)$$

The value of $\jmath_X(x,d)$ can be loosely interpreted as the amount of information that needs to be stored about x in order to restore it with distortion d [1]. The explicit nature of (4) illuminates the tension between the likelihood of x and the target distortion: the likelier realization x is, the fewer bits are required to store it; the lower tolerable d is, the more bits are required in order to represent the source with that distortion.[1] We stress that this intuitively pleasing insight is not afforded by the general formula (3).

Whenever (4) holds for the single-letter distribution of a memoryless source, one can insert (4) into (1) to conclude that in those cases, the nonasymptotic fundamental limit is given simply by

$$R(n,d,\epsilon) = \underline{R}(d) + \sqrt{\frac{\mathcal{V}}{n}}Q^{-1}(\epsilon) + O\left(\frac{\log n}{n}\right), \quad (6)$$

where $\underline{R}(d)$ is Shannon's lower bound, and $\mathcal{V}$ is the varentropy of the source.

[1] $\phi(d)$ is strictly increasing in d.

To gain further insight into the form of (4), recall that Shannon's lower bound [3] states that the rate-distortion function is bounded below by the difference between the (differential) entropy of the source, and a term $\phi(d)$ that depends only on the distortion measure and distortion threshold d. Thus Shannon's lower bound is given by the expectation of (4). Due to its simplicity and because it becomes increasingly tight in the limit of low distortion [4], [5], Shannon's lower bound is often used as a convenient proxy for $R(d)$.

In this paper, we formulate an abstract Shannon's lower bound, which encompasses the original Shannon's lower bound as a special case and which does not impose any symmetry conditions on the distortion measure. We show that the abstract Shannon lower bound holds with equality if and only if (4) also holds. Thus, Shannon's lower bound has far-fledging implications and connections in the nonasymptotic analysis of lossy source coding that we explore in this paper. In particular, a nonasymptotic version of Shannon's lower bound is closely linked to a certain binary hypothesis test and to a covering of the high probability set of source realizations with distortion d-balls.

According to a classical result by Pinkston [6], for discrete sources with difference distortion measures, Shannon's lower bound is satisfied with equality in the range $0 \le d_c$, where $d_c > 0$ is a function of P_X only. Our necessary and sufficient condition implies that (5) and (6) also always hold for these sources. Leveraging this insight, we show new simple finite blocklength bounds for Pinkston's scenario. The new bounds lend themselves to a particularly straightforward second order analysis.

For continuous sources, Shannon's lower bound holds with equality only if a peculiar matching between a source density and distortion measure is present. A companion paper [7] shows that under the mean-square error distortion, as long as d is small enough and the source density f_X satisfies a smoothness condition, the d-tilted information in $X \in \mathbb{R}^n$ is closely *approximated* by (4), even if Shannon's lower bound does not hold with equality. An ArXiv preprint [8] extends the results of [7] to non-MSE distortion measures.

The rest of the paper is organized as follows. Section II introduces the abstract Shannon lower bound. Section III presents the necessary and sufficient condition for (4) to hold. Leveraging the tightness of Shannon's lower bound in a range of low distortions, Section IV shows new finite blocklength bounds for a discrete memoryless source with a difference distortion measure, together with their second order analysis.

II. Shannon's lower bound

The (informational) rate-distortion function is defined for random variable $X \in \mathcal{X}$ and distortion measure $\mathsf{d}\colon \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}_+$ as the solution to the convex optimization problem in (2). The function in (2) admits the following parametric representation.

Theorem 1 (Parametric representation of $R(d)$, [9])**.** *Suppose that the following basic assumptions are satisfied.*

(a) $R(d)$ is finite for some d, i.e. $d_{\min} < \infty$, where

$$d_{\min} \triangleq \inf\{d\colon\ R(d) < \infty\}. \tag{7}$$

(b) The distortion measure is such that there exists a finite set $E \subset \mathcal{Y}$ such that

$$\mathbb{E}\left[\min_{y\in E} \mathsf{d}(X,y)\right] < \infty. \tag{8}$$

For each $d > d_{\min}$, it holds that

$$R(d) = \max_{g(x),\ \lambda} \{-\mathbb{E}[\log g(X)] - \lambda d\}, \tag{9}$$

where the maximization is over $g(x) \ge 0$ and $\lambda \ge 0$ satisfying the constraint

$$\mathbb{E}\left[\frac{\exp(-\lambda \mathsf{d}(X,y))}{g(X)}\right] \le 1\ \forall y \in \mathcal{Y}. \tag{10}$$

Remark 1. The maximization over $g(x) \ge 0$ in (9) can be restricted to only $0 \le g(x) \le 1$ [9]. Equality in (10) holds for $P_{Y^\star}$-a.s. y.

Remark 2. The d-tilted information (defined in (3), [1]) can be alternatively defined as

$$\jmath_X(x,d) = -\log g(x) - \lambda^\star d, \tag{11}$$

where the pair $(g(\cdot), \lambda)$ attains the maximum in (9). So,

$$R(d) = \mathbb{E}[\jmath_X(X,d)]. \tag{12}$$

Furthermore, if the infimum in (2) is attained by some $Y^\star$, then

$$g(x) = \mathbb{E}[\exp(-\lambda^\star \mathsf{d}(x, Y^\star))] \tag{13}$$

leads to the definition in (3).

For finite alphabet sources, a parametric representation of $R(d)$ is contained in Shannon's paper [3]; both Gallager's [10, Theorem 9.4.1] and Berger's [11] texts contain parametric representations of $R(d)$ for discrete and continuous sources. However, it was Csiszár [9] who gave rigorous proofs of (9) in the following much more general setting: X belongs to a general abstract probability space, and the existence of the conditional distribution $P_{Y^\star|X}$ attaining $R(d)$ is not required.

Here, we leverage the result of Csiszár to state a generalization of Shannon's lower bound to abstract probability spaces.

Each choice of $\lambda \ge 0$ and g satisfying (10) gives rise to a lower bound to $R(d)$. Shannon's lower bound corresponds to a particular choice of (λ, g).

Let μ be a measure on $\mathcal{X}$ such that the distribution of X is absolutely continuous with respect to μ. Denote the density of the distribution of X with respect to μ (Radon-Nikodym derivative) by

$$f_X(x) \triangleq \frac{dP_X}{d\mu}(x). \tag{14}$$

$\Sigma \triangleq \sup_{y\in\mathcal{Y}} \int \exp(-\lambda \mathsf{d}(x,y))d\mu(x)$ $= \int \exp(-\lambda \mathsf{d}(x,y_\lambda))d\mu(x)$
$\frac{dP_{X\|Y^\star=y}}{d\mu}(x) \triangleq \frac{\exp(-\lambda \mathsf{d}(x,y))}{\int \exp(-\lambda \mathsf{d}(x,y))d\mu(x)}$
$\phi_\mu(d) \triangleq \log\Sigma + \lambda d$ $g(x) = f_X(x)\Sigma$
$\lambda > 0$: arbitrary

TABLE I: The choice of $(g(x), \lambda)$ in (9) that leads to the abstract Shannon's lower bound in Theorem 2.

The differential entropy with respect to μ can be defined as

$$h_\mu(X) \triangleq -\int d\mu(x) f_X(x) \log f_X(x) \tag{15}$$
$$= -D(f_X\|\mu). \tag{16}$$

If X is a continuous random variable, a natural choice for μ is the Lebesgue measure. Then, the density in (14) is known as the probability density function, and $h_\mu(X)$ is simply $h(X)$, the differential entropy of X. If X is a discrete random variable, a natural choice for μ is the counting measure. Then, the density in (14) is the probability mass function, and $h_\mu(x)$ is equal to $H(X)$, the entropy of X.

It is easy to verify that the choice of λ and g in Table I satisfies (10). The generalization of Shannon lower bound to abstract spaces and arbitrary distortion measures can now be stated as follows.

Theorem 2 (abstract Shannon's lower bound)**.** *Fix a measure μ such that the distribution of X is absolutely continuous with respect to μ. For all $d \geq d_{\min}$,*

$$R(d) \geq h_\mu(X) - \phi_\mu(d). \tag{17}$$

Theorem 2 provides a family of lower bounds parameterized by the choice of base measure μ. In classical versions of Shannon's lower bound, μ is a Lebesgue measure (or a counting measure, if the alphabet is discrete) and the distortion measure satisfies a symmetry condition, so that the integral in the definition of Σ in Table I does not depend on the choice of y. Shannon's original derivation [12] applied to continuous sources under the mean-square error distortion, and it did not use a parametric representation of $R(d)$. A decade later, Pinkston [6] derived a version of the bound for a finite alphabet source with a distortion measure such that all the columns of the per-letter distortion matrix $\mathsf{d}(x,y)$ consist of the same set of entries. A generalization of the discrete Shannon lower bound to distortion measures not satisfying any symmetry conditions is put forth by Gray [13]. The new bound in Theorem 2 is more general than these results and recovers them as special cases.

The right-side of (17) can be made equal to $R(d)$ by choosing μ as follows:

$$\frac{d\mu}{dP_X}(x) = \exp\left(\jmath_X(x,d)\right). \tag{18}$$

To verify that the choice in (18) results in equality in (17), observe that

$$h_\mu(X) = \mathbb{E}\left[\jmath_X(X,d)\right], \tag{19}$$

and that

$$\phi_\mu(d) = \log\Sigma + \lambda d \tag{20}$$
$$= \sup_{y\in\mathcal{Y}} \log \mathbb{E}\left[\exp\left(-\lambda \mathsf{d}(X,y) + \jmath_X(X,d)\right)\right] + \lambda d \tag{21}$$
$$= 0, \tag{22}$$

where to obtain (22) we used (10), (11) and Remark 1.

The long-standing appeal of Shannon's lower bound is that one can obtain a tight bound on the rate-distortion function even without the knowledge of the distribution that attains it, as (18) demands. For an illustration of such a calculation, suppose that $\mathcal{X}$ is a set endowed with a group operation, "$+$", satisfying the group axioms. Then, it makes sense to talk about $x+y$ and $x-y = x+(-y)$, where $-y$ is the inverse of y (according to the group operation). Distortion measures of the form

$$\mathsf{d}(x,y) = \mathsf{d}(x-y) \tag{23}$$

are called *difference* distortion measures. If $\mathcal{X} = \mathbb{R}^n$ and d is a difference distortion measure, then letting μ be the Lebesgue measure, we obtain

$$\Sigma_\lambda = \int \exp(-\lambda \mathsf{d}(x-y))dx, \tag{24}$$

regardless of the choice of y. So, we may set $y=0$, and obtain a particularly elegant form of the abstract Shannon lower bound - see Table II.

In the same fashion, if $\mathcal{X}$ is a discrete group, letting μ be the counting measure on $\mathcal{X}$, we notice that

$$\Sigma = \sum_{x\in\mathcal{X}} \exp(-\lambda \mathsf{d}(x-y)), \tag{25}$$

for all $y \in \mathcal{X}$. Therefore, we may let $y=0$ (the identity element of group $\mathcal{X}$) and obtain Pinkston's variant of Shannon's lower bound [6]. See Table II.

We proceed to list several examples of the calculation of Shannon's lower bound for difference distortion measures.

Example. In the special case of $X \in \mathbb{R}^n$ and mean-square error distortion, we recover the original Shannon's lower bound [12] as follows. Let d be the mean-square error distortion:

$$\mathsf{d}(x^n, y^n) = \frac{1}{n}\|x-y\|_2^2. \tag{26}$$

$$\Sigma \triangleq \int \exp(-\lambda \mathsf{d}(z)) d\mu(z)$$

$$\frac{dP_{Z_\lambda}}{d\mu}(z) \triangleq \frac{1}{\Sigma}\exp(-\lambda \mathsf{d}(z))$$

$$\phi_\mu(d) \triangleq h_\mu(Z_\lambda) = \log \Sigma + \lambda d$$

$$g(x) = f_X(x)\Sigma$$

$\lambda > 0$: solution to equation $\mathbb{E}[\mathsf{d}(Z_\lambda)] = d$

TABLE II: The choice of $(g(x), \lambda)$ in (9) that leads to Shannon's lower bound in the case where d is difference distortion measure. The base measure μ is understood to be the counting measure if $\mathcal{X}$ is a discrete group, and the Lebesgue measure if $\mathcal{X} = \mathbb{R}^n$.

A straightforward calculation using Table II reveals that,

$$\lambda = \frac{n}{2d}\log e \tag{27}$$

$$\phi_\mu(d) = \frac{n}{2}\log d + \frac{n}{2}\log(2\pi e), \tag{28}$$

so if X is a continuous real-valued random vector of length n,

$$R(d) \geq h(X) + \frac{n}{2}\log\frac{1}{d} - \frac{n}{2}\log(2\pi e). \tag{29}$$

Example. For weighted mean-square error distortion measure,

$$\mathsf{d}(x,y) = \frac{1}{n}\|W(x-y)\|_2^2, \tag{30}$$

where W is an invertible $n \times n$ matrix, Shannon's lower bound is given by

$$R(d) \geq h(X) + \frac{n}{2}\log\frac{1}{d} - \frac{n}{2}\log(2\pi e) + \log|\det W|. \tag{31}$$

Example. Let d be the scaled L^p norm distortion:

$$\mathsf{d}(x,y) = n^{-\frac{s}{p}}\|x-y\|_p^s, \tag{32}$$

where $s > 0$. A direct calculation using Table II shows that Shannon's lower bound is given by

$$R(d) \geq h(X) + \frac{n}{s}\log\frac{1}{d} - \frac{n}{p}\log n - \log b_{n,p} + \frac{n}{s}\log\frac{n}{se} - \log\Gamma\left(\frac{n}{s}+1\right), \tag{33}$$

where $b_{n,p}$ is the volume of a unit L_p ball:

$$b_{n,p} \triangleq \frac{\left(2\Gamma\left(\frac{1}{p}+1\right)\right)^n}{\Gamma\left(\frac{n}{p}+1\right)}. \tag{34}$$

Example. Assume that the alphabet is finite, $|\mathcal{X}| = m$, and consider the symbol error distortion

$$\mathsf{d}(z) = 1\{z = 0\}. \tag{35}$$

Then,

$$R(d) \geq H(X) - h(d) - d\log(m-1). \tag{36}$$

As it turns out, the abstract Shannon lower bound in Theorem 2 has a nonasymptotic kin expressed in terms of the Neyman-Pearson function.

The optimal performance achievable among all randomized tests $P_{W|X}: A \to \{0,1\}$ between measures P and Q on A is denoted by (1 indicates that the test chooses P):

$$\beta_\alpha(P,Q) = \min_{\substack{P_{W|X}: \\ P[W=1]\geq\alpha}} Q[W=1] \tag{37}$$

Note that the Neyman-Pearson function $\beta_\alpha(P,Q)$ is well defined even if P and Q are not probability measures.

An (M,d,ϵ) lossy compression code is a mapping $P_{\hat{X}|X}$, where $\hat{X}$ takes M values, and

$$\mathbb{P}\left[\mathsf{d}(X,\hat{X}) > d\right] \leq \epsilon. \tag{38}$$

In [1] we showed the following converse result.

Theorem 3 (Converse, [1])**.** *Let P_X be the source distribution defined on the alphabet $\mathcal{X}$. Any (M,d,ϵ) code must satisfy*

$$M \geq \sup_\mu \frac{\beta_{1-\epsilon}(P_X,\mu)}{\sup_{y\in\mathcal{Y}}\mu[\mathsf{d}(X,y)\leq d]}. \tag{39}$$

where the supremum is over all measures on $\mathcal{X}$.

Note the striking parallels between Theorem 3 and the abstract Shannon lower bound in Theorem 2. Both bounds require a choice of the base measure μ. The optimal binary hypothesis test in (39) is a function of $\log\frac{d\mu}{dP_X}(x)$ only, whose expectation is equal to $h_\mu(X)$, the first term in (17). Furthermore, by Markov's inequality, the μ-volume of the distortion d-ball is linked to $\phi_\mu(d)$, the second term in (17), as follows.

$$\mu[d(X,y)\leq d] = \int d\mu(x) 1\{\mathsf{d}(x,y)\leq d\} \tag{40}$$

$$\leq \int d\mu(x)\exp(\lambda d - \lambda\mathsf{d}(x,y)) \tag{41}$$

$$\leq \sup_{y\in\mathcal{Y}}\int d\mu(x)\exp(\lambda d - \lambda\mathsf{d}(x,y)) \tag{42}$$

$$= \exp(\phi_\mu(d)). \tag{43}$$

III. The necessary and sufficient condition

The following result pins down the necessary and sufficient condition for equality in (17) to hold.

Theorem 4. *Assume that the infimum in (2) is achieved by some $P_{Y^\star|X}$. Then, the following statements are equivalent.*

A. The rate-distortion function is equal to Shannon's lower bound,

$$R(d) = h_\mu(x) - \phi_\mu(d). \tag{44}$$

B. For P_X-a.s. x,

$$\jmath_X(x,d) = \log\frac{1}{f_X(x)} - \phi_\mu(d). \tag{45}$$

C. *The backward conditional distribution*[2] *that achieves* $R(d)$ *satisfies, for* $P_{Y^\star}$*-a.s.* y*,*

$$\frac{dP_{X|Y^\star=y}}{d\mu}(x) = \frac{\exp(-\lambda \mathsf{d}(x,y))}{\Sigma}. \tag{46}$$

Proof. B $\Rightarrow$ A is trivial. To show A $\Rightarrow$ B, note that the existence of $P_{Y^\star|X}$ that achieves the infimum in (2) implies differentiability of $R(d)$ [9]. It follows that the maximum in (9) is attained by a unique $g(x)$ [9]. Since A establishes that $g(x)$ that attains the maximum in (9) is that in Table I, B is immediate.

To show B $\Leftrightarrow$ C, recall the equivalent representation of $\jmath_X(x,d)$ [1]:

$$\jmath_X(x,d) = \log \frac{dP_{X|Y^\star=y}}{dP_X}(x) + \lambda \mathsf{d}(x,y) - \lambda d. \tag{47}$$

Equality in (47) holds for $P_{Y^\star}$-a.s. y. Comparing (45) and (47) we conclude the equivalence B $\Leftrightarrow$ C. □

The necessary and sufficient conditions in Theorem 4 assume a particularly simple form for difference distortion measures. In that case, statement C can be replaced by

C′. There exists a random variable $Y^\star$ such that

$$X = Y^\star + Z_\lambda, \tag{48}$$

where $Y^\star$ is independent of Z_λ, and Z_λ is defined in Table II.

Example. If X is equiprobable on a finite group, (44) always holds. Indeed, in that case, equiprobable $Y^\star$ satisfies (48).

Example. Gaussian source with mean-square error distortion satisfies the conditions of Theorem 4; indeed, $X = Y^\star + Z$, where $X \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$, $Y^\star \sim \mathcal{N}(0, (\sigma^2 - d)\mathbf{I}) \perp\!\!\!\perp Z \sim \mathcal{N}(0, d\mathbf{I})$.

Theorem 4 extends a result by Gerrish and Schultheiss [15], who showed that for the compression of a continuous random vector under the mean-square error distortion, the Shannon lower bound gives the actual value of rate-distortion function if and only if X can be written as the sum of two independent random vectors $X = Y^\star + Z$, where $Z \sim \mathcal{N}(0, d\mathbf{I})$. Theorem 4 also generalizes the backward-channel condition for equality in the Shannon lower bound given in [11, Theorem 4.3.1]. Unlike these classical results, Theorem 4 applies to abstract sources and does not enforce any symmetry assumptions on the distortion measure.

IV. Finite alphabet sources

Most continuous probability distributions do not meet the conditions of Theorem 4. In particular, an X with indecomposable distribution cannot satisfy (48), for any difference distortion measure. In contrast, as the following result shows, for finite alphabet sources Shannon's lower bound is always attained with equality, as long as target distortion d is not too large.

[2]That is, $P_{X|Y^\star}$ such that $P_X P_{Y^\star|X} = P_{X|Y^\star} P_X$.

Theorem 5 (Pinkston [6])**.** *Let* $X \in \mathcal{X}$*, where* $\mathcal{X}$ *is a group of order* m*. Let the distortion measure satisfy* (23) *and*

$$\mathsf{d}(0) = 0, \quad \mathsf{d}(z) > 0, z \neq 0. \tag{49}$$

Then, there exists a $d_c > 0$ *such that Shannon's lower bound is satisfied with equality for*

$$0 \leq \forall d \leq d_c. \tag{50}$$

Example. For symbol error distortion equality in (36) holds for all

$$0 \leq d \leq (m-1) \min P_X(x). \tag{51}$$

Theorem 5 was obtained by Pinkston [6] for a more general case in which all rows and columns of the per-letter distortion matrix $\mathsf{d}(x,y)$ consists of the same set of entries (balanced distortion measure). Gray [13] showed that the rate-distortion function equals Shannon's lower bound in the range of small distortions for stationary ergodic finite alphabet sources, generalizing and simplifying the proofs of Gray's previous results in [16] (binary Markov source with BER distortion and Gauss-Markov source) and [13] (finite state finite alphabet Markov sources).

Leveraging the necessary and sufficient conditions in Theorem 4, we conclude that under the conditions of Theorem 5, the d-tilted information is given by (5), and the output random variable $Y^\star$ that achieves the rate-distortion function satisfies (48).

Taking advantage of these observations, we proceed to show simple finite blocklength bounds for the case of finite alphabet source with difference distortion measures.

Let n be the blocklength. We adopt the notation of [17]:

- type of the string: $\mathbf{k} = (k_1, \ldots, k_m)$, $k_1 + \ldots + k_m = n$
- probability of a given string of type $\mathbf{k}$: $p^{\mathbf{k}} = P_X(1)^{k_1} \ldots P_X(m)^{k_m}$
- multinomial coefficient: $\binom{n}{\mathbf{k}} = \dfrac{n!}{k_1! \ldots k_m!}$

We assume that the distortion measure is of form

$$\mathsf{d}(x^n, y^n) = \frac{1}{n} \sum_{i=1}^{n} \mathsf{d}(x_i - y_i) \tag{52}$$

First, we state a nonasymptotic Shannon's lower bound.

Theorem 6 (Converse)**.** *Let* P_{X^n} *be the source distribution defined on* $\mathcal{A}^n$*, where* $\mathcal{A}$ *is a finite alphabet, and let* d *be a separable difference distortion measure, as in* (52)*. Any* (M, d, ϵ) *code must satisfy*

$$M \geq \frac{\beta_{1-\epsilon}(P_{X^n}, \mu)}{\mu\left[\mathsf{d}(X^n, \mathbf{0}) \leq d\right]} \tag{53}$$

where μ *is the counting measure.*

Proof. Observe that $\mu\left[\mathsf{d}(X^n, y^n) \leq d\right]$ does not depend on the choice of y^n, so we may put $y^n = \mathbf{0}$ in Theorem 3. □

As we will see in Theorem 9, in the range of distortions where Shannon's lower bound is tight, an asymptotic analysis of the converse in Theorem 6 leads to a tighter third order term than previously known.

To introduce our new achievability result, we recall first the exact performance of random coding [1]. Denote the d-ball centered at x^n by

$$B_d(x^n) \triangleq \{y^n \in \mathcal{B}^n \colon \mathsf{d}(x^n, y^n) \le d\}\,. \tag{54}$$

Theorem 7 (Achievability [1, Theorem 9])**.** *There exists an (M, d, ϵ) code with*

$$\epsilon \le \inf_{P_{Y^n}} \mathbb{E}\left[1 - P_{Y^n}(B_d(X^n))\right]^M \tag{55}$$

where the infimization is over all random variables defined on $\mathcal{B}^n$, independent of X^n.

The right side of (55) gives the exact performance of random coding. Consequently, it is impossible to tighten (55) using Shannon's random coding argument alone. Unfortunately, in general the computation of $P_{Y^n}(B_d(x^n))$ in (55) has exponential in n complexity. Polynomial complexity lower bounds on $P_{Y^n}(B_d(x^n))$ were proposed and computed in [1] individually for the cases of the Gaussian source, the binary memoryless source and the discrete memoryless source. In the proof of Theorem 8 below, we present a new lower bound on $P_{Y^n}(B_d(x^n))$ that can be computed in polynomial time for any discrete memoryless source with a difference distortion measure as long as the target distortion is below Pinkston's critical value d_c.

Theorem 8 (Achievability)**.** *Let $P_{X^n} = P_X \times \ldots P_X$, where X is defined on a finite alphabet $\mathcal{A}$, and let d be a separable difference distortion measure, as in (52), satisfying (49). Suppose that $d < d_c$. There exists an (M, d, ϵ) code that satisfies*

$$\epsilon \le \sum_{\mathbf{k}} \binom{n}{\mathbf{k}} p^{\mathbf{k}} \left(1 - \binom{n}{\mathbf{k}}^{-1} \binom{n}{\mathbf{t}(\mathbf{k})} \prod_{a=1}^{m} \binom{t_a(\mathbf{k})}{\mathbf{t}_a(\mathbf{k})} q^{\mathbf{t}(\mathbf{k})}\right)^M \tag{56}$$

where

- *p denotes the probability vector P_X;*
- *q is the probability vector $P_{Y^\star}$ in (48) (the single-letter output distribution that achieves the rate-distortion function);*
- *$\mathbf{t}(\mathbf{k}) = [nq_{\mathbf{k}}]$, where $q_{\mathbf{k}}$ is the output distribution that achieves the rate-distortion function for the input distribution $\frac{1}{n}\mathbf{k}$, and $t_a(\mathbf{k})$ is the a-th entry of $\mathbf{t}(\mathbf{k})$ [3];*
- *the b-th entry of $\mathbf{t}_a$ is given by [4]*

$$[P_{Z_\lambda}(a-b)t_a(\mathbf{k})]; \tag{57}$$

- *Z_λ is defined in Table II.*

Proof. Let $P_{Y^n} = P_{Y^\star} \times \ldots P_{Y^\star}$. Denote by $L_n(\mathbf{k}, \mathbf{t}, d)$ the number of binary strings of type $\mathbf{t}$ that lie within d-distortion d from a given string of type $\mathbf{k}$. In other words, $L_n(\mathbf{k}, \mathbf{t}, d)$ is the volume of the intersection of $B_d(x^n)$, where x^n has type $\mathbf{k}$, with type class $\mathbf{t}$.

We have

$$P_{Y^n}(B_d(x^n)) = \sum_{\mathbf{t}} L_n(\mathbf{k}, \mathbf{t}, d) q^{\mathbf{t}} \tag{58}$$

$$\ge L_n(\mathbf{k}, \mathbf{t}(\mathbf{k}), d) q^{\mathbf{t}(\mathbf{k})}. \tag{59}$$

It is easy to verify that

$$L_n(\mathbf{k}, \mathbf{t}, d) \triangleq \binom{n}{\mathbf{k}}^{-1} \binom{n}{\mathbf{t}} \sum \prod_{a=1}^{m} \binom{t_a}{\mathbf{t}_a} \tag{60}$$

where the summation is over all collections of t_a-types $\mathbf{t}_a = (t_{a,1}, \ldots, t_{a,m})$ such that

$$\mathsf{d}(x^n, y^n) = \frac{1}{n} \sum_{a=1}^{m} t_a \sum_{b=1}^{m} t_{a,b}\, \mathsf{d}(a, b) \tag{61}$$

$$\le d, \tag{62}$$

where the type of x^n is $\mathbf{k}$, the type of y^n is $\mathbf{t}$, and the conditional type of x^n given y^n is $\mathbf{t}_a$. Weakening (60) by keeping only the joint type on the boundary of the intersection of $B_d(x^n)$ with type class $\mathbf{t}$, we obtain

$$L_n(\mathbf{k}, \mathbf{t}(\mathbf{k}), d) \ge \binom{n}{\mathbf{k}}^{-1} \binom{n}{\mathbf{t}(\mathbf{k})} \prod_{a=1}^{m} \binom{t_a}{\mathbf{t}_a(\mathbf{k})}, \tag{63}$$

and an application of Theorem 7 concludes the proof. □

Particularized to the binary memoryless source with Hamming distortion, the bound in Theorem 8 is plotted in Fig. 1. Although not as tight as the achievability result from [1] also plotted in Fig. 1, the advantage of the bound in Theorem 8 is that it applies to *all* finite alphabet sources with difference distortion measure as long as $d < d_c$. In contrast, the bound from [1] is *tailored* to the binary source with Hamming distortion. Furthermore, the new bound in Theorem 8 is dispersion-optimal, and the proof thereof is much simpler than the prior refined asymptotic analyses of lossy source coding [1], [18]–[20]. Also plotted in Fig. 1 is the Gaussian approximation in (6), (65), the original Shannon's achievability result [3], and the converse in Theorem 6, which in this case coincides with the converse in [1, Theorem 20].

We already saw that the dispersion of lossy source coding is equal to the varentropy of the source whenever Shannon's lower bound is tight. The following theorem demonstrates that this result can be be obtained by a straightforward analysis of the bounds in Theorems 6 and 7. An added bonus is a tighter converse bound on the logarithmic term than previously known.

The coding rate of an (M, d, ϵ) code for an n-dimensional source is the ratio $\frac{\log M}{n}$. The minimum coding rate compatible with n, d and ϵ is denoted by $R(n, d, \epsilon)$.

[3] $[\cdot]$ denotes rounding off to the nearest integer so that the resulting n-vector is an n-type.

[4] Rounding off in (57) is carried out so that the resulting vector is $t_a(\mathbf{k})$-type, and $\frac{1}{n}\sum_{a=1}^{m} t_a(\mathbf{k}) \sum_{b=1}^{m} t_{a,b}(\mathbf{k}) \mathsf{d}(a, b) \le d$, where $t_{a,b}(\mathbf{k})$ denotes the b-th entry of $\mathbf{t}_a$.

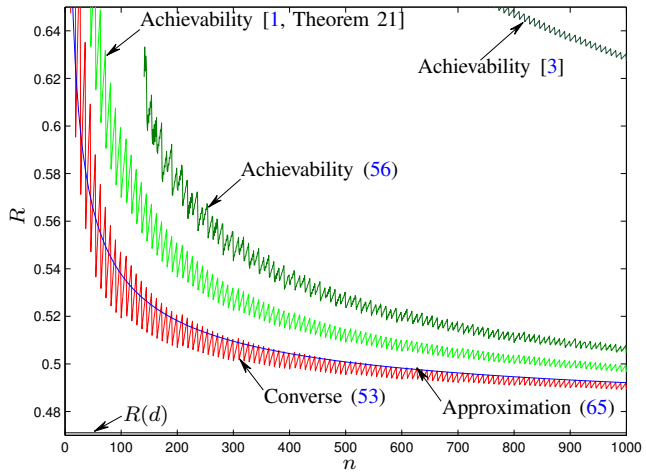

Fig. 1: Bounds to $R(n,d,\epsilon)$ and Gaussian approximation for binary memoryless source with $p=2/5$, $d=0.11$, $\epsilon=10^{-2}$.

Theorem 9 (Second order analysis)**.** *Let $P_{X^n}=P_X\times\ldots P_X$ be a probability distribution defined on a finite alphabet, and let d be a separable difference distortion measure, as in (52). Suppose that*

$$0\le d<d_c, \tag{64}$$

so that Shannon's lower bound holds with equality. Then, the minimum coding rate compatible with blocklength n and probability ϵ of exceeding threshold d expands as

$$R(n,d,\epsilon)=H(X)-\phi(d)+\sqrt{\frac{\mathcal{V}}{n}}Q^{-1}(\epsilon)+\frac{1}{n}\theta(\log n), \tag{65}$$

where $\mathcal{V}$ is the varentropy of the source. If $d>0$, the remainder term $\theta(\cdot)$ is bounded as

$$O(1)\le\theta(\log n)\le O(\log n), \tag{66}$$

whereas if $d=0$ [1, Theorem 31],

$$\theta(\log n)=-\frac{1}{2}\log n+O(1). \tag{67}$$

The following result will be useful in the proof of Theorem 9.

Lemma 1 ([21, Lemma 1])**.** *Let $X_1,\ldots,X_n$ be independent on $\mathcal{A}$ and distributed according to P_X. For all n and all $\gamma>0$, it holds that*

$$\mathbb{P}\left[\|\mathrm{type}(X^n)-nP_X\|^2>n^2\gamma\right]\le 2|\mathcal{A}|\exp\left(-\frac{n}{2}\frac{\gamma}{|\mathcal{A}|}\right), \tag{68}$$

where $\|\cdot\|$ denotes the Euclidean norm of its $|\mathcal{A}|$-dimensional vector argument.

We also use the "reverse Pinsker inequality" [22, Lemma 6.3]:

$$D(X\|\bar{X})\le\frac{\log e}{\min_{a\in\mathcal{A}}P_{\bar{X}}(a)}\|P_X-P_{\bar{X}}\|^2 \tag{69}$$

and Stirling's approximation of the multinomial coefficient:

$$\binom{n}{\mathbf{k}}=\frac{C}{n^{\frac{m}{2}-\frac{1}{2}}}\exp\left\{nH\left(\frac{\mathbf{k}}{n}\right)\right\} \tag{70}$$

where C is a constant.

Proof of Theorem 9. Fix $0<d<d_c$. To show the converse, recall that [23, Lemma 58], [1, (251)]

$$\log\beta_{1-\epsilon}(P_{X^n},\mu^n)=nH(X)+\sqrt{n\mathcal{V}}Q^{-1}(\epsilon)-\frac{1}{2}\log n+O(1). \tag{71}$$

The converse is immediate from[5]

$$\log\mu\left[\mathsf{d}(X^n,\mathbf{0})\le d\right]=nH(Z_\lambda)-\frac{1}{2}\log n+O(1), \tag{72}$$

where Z_λ is defined in (48). We proceed to show (72). Observe that

$$\mu\left[\mathsf{d}(X^n,\mathbf{0})\le d\right]=\sum_{\substack{\mathbf{j}:\\ \sum_{i=1}^m\mathsf{d}(i)j_i\le nd}}\binom{n}{\mathbf{j}} \tag{73}$$

By the definition of Z_λ in Table II,

$$d=\sum_{i=1}^m\mathsf{d}(i)P_{Z_\lambda}(i). \tag{74}$$

Consider the integer vector $[nP_{Z_\lambda}]$, where $[\cdot]$ denotes rounding off to the nearest integer so that $\sum_{i=1}^m\mathsf{d}(i)[nP_{Z_\lambda}(i)]\le d$. We may re-write (73) as

$$\mu\left[\mathsf{d}(X^n,\mathbf{0})\le d\right]=\sum_{\substack{\mathbf{\Delta}:\\ \sum_{i=1}^m\mathsf{d}(i)\Delta_i\le 0}}\binom{n}{[nP_{Z_\lambda}]+\mathbf{\Delta}}. \tag{75}$$

Now, a direct application of [17, (105)] to (75) yields (72).

To show the achievability, we apply (70) to (63) and plug into (59) to obtain

$$P_{Y^n}(B_d(x^n)) \tag{76}$$

$$\ge Cn^{-\frac{m-1}{2}}\exp\left(-nD\left(\frac{1}{n}\mathbf{t}(\mathbf{k})\|q\right)+n\phi(d)-nH\left(\frac{\mathbf{k}}{n}\right)\right) \tag{77}$$

Observe that

$$nH\left(\frac{\mathbf{k}}{n}\right)=\sum_{i=1}^n\log\frac{1}{P_X(x_i)}-nD\left(\frac{1}{n}\mathbf{k}\middle\|P_X\right). \tag{78}$$

The first term in (78) is a sum of i.i.d. random variables with mean $H(X)$ and variance $\mathcal{V}$. To evaluate the second term in (78), consider $\mathcal{T}$, the set of typical sentences of x^n:

$$\mathcal{T}\triangleq\left\{x^n\in\mathcal{A}^n\colon\left\|\frac{1}{n}\mathrm{type}(x^n)-P_X\right\|^2\le|\mathcal{A}|\frac{\log n}{n}\right\}. \tag{79}$$

[5] At the expense of weakening the lower bound in (66) to $-\frac{1}{2}\log n+O(1)$, one can simply apply (43) to evaluate $\log\mu\left[d(X^n,\mathbf{0})\le d\right]\le nH(Z_\lambda)$.

According to (68),

$$\mathbb{P}\left[X^n \in \mathcal{T}\right] \geq 1 - \frac{2|\mathcal{A}|}{\sqrt{n}}. \tag{80}$$

Due to the reverse Pinsker inequality in (69), as long as $x^n \in \mathcal{T}$,

$$D\left(\frac{\mathbf{k}}{n}\middle\| P_X\right) \leq \frac{|\mathcal{A}| \log e}{\min_{a\in\mathcal{A}} P_{\bar{X}}(a)} \frac{\log n}{n}. \tag{81}$$

Due to (48), $\mathbf{t}(\mathbf{k})$ is continuous in $\mathbf{k}$; so, for any $x^n \in \mathcal{T}$, we may apply the reverse Pinsker inequality again to conclude

$$D\left(\frac{1}{n}\mathbf{t}(\mathbf{k})\|q\right) \leq O\left(\frac{\log n}{n}\right). \tag{82}$$

Applying (68), (81) and (82) to (84), we conclude that for all $x^n \in \mathcal{T}$,

$$P_{Y^n}(B_d(x^n)) \tag{83}$$

$$\geq \exp\left(n\phi(d) - \sum_{i=1}^{n} \log \frac{1}{P_X(x_i)} + O\left(\log n\right)\right) \tag{84}$$

The achievability result is now obtained as follows. Applying $(1-x)^M \leq e^{-Mx}$ to (55), we conclude that there exists an (M, d, ϵ) code with

$$\epsilon \leq \mathbb{E}\left[e^{-MP_{Y^n}(B_d(X^n))}\right] \tag{85}$$

$$\leq \mathbb{E}\left[e^{-MP_{Y^n}(B_d(X^n))} 1\left\{X^n \in \mathcal{T}\right\}\right] + \frac{2|\mathcal{A}|}{\sqrt{n}} \tag{86}$$

Finally, we let

$$\log M = H(X) - \phi(d) + \sqrt{n\mathcal{V}} Q^{-1}\left(\epsilon - \frac{2|\mathcal{A}|}{\sqrt{n}}\right) + O\left(\log n\right), \tag{87}$$

insert (84) in (86), and apply the Berry-Esseen theorem in the same manner it is done in [1, (107)–(113)]. □

V. Conclusion

Shannon's lower bound provides a powerful tool to study the rate-distortion function. We proposed an abstract Shannon's lower bound (Theorem 2), which applies to sources defined on general probability spaces with arbitrary distortion measures. We presented the necessary and sufficient condition for Shannon's lower bound to be attained exactly (Theorem 4). Whenever Shannon's lower bound is attained exactly, the d-tilted information in x also admits a simple representation as the difference between the information in x and a term that depends only on tolerated distortion d (see (45)). All finite alphabet sources with difference distortion measures meet Shannon's lower bound with equality in a range of low distortions [6]. This implies in particular that the rate-dispersion function of a discrete memoryless source with difference distortion measure is given simply by the varentropy of the source, as long as the target distortion is low enough. The tightness of Shannon's lower bound also leads to simplified finite blocklength bounds.

References

[1] 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.

[2] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Transactions on Information Theory*, vol. 18, no. 4, pp. 460–473, 1972.

[3] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRE Int. Conv. Rec.*, vol. 7, no. 1, pp. 142–163, Mar. 1959, reprinted with changes in *Information and Decision Processes*, R. E. Machol, Ed. New York: McGraw-Hill, 1960, pp. 93-126.

[4] Y. N. Linkov, "Evaluation of ϵ-entropy of random variables for small ϵ," *Problems of Information Transmission*, vol. 1, pp. 18–26, 1965.

[5] T. Linder and R. Zamir, "On the asymptotic tightness of the Shannon lower bound," *IEEE Transactions on Information Theory*, vol. 40, no. 6, pp. 2026–2031, 1994.

[6] J. Pinkston, "An application of rate-distortion theory to a converse to the coding theorem," *IEEE Transactions on Information Theory*, vol. 15, no. 1, pp. 66–71, 1969.

[7] V. Kostina, "Data compression with low distortion and finite blocklength," in *Proceedings 53rd Annual Allerton Conference on Communication, Control and Computing*, Monticello, IL, Oct. 2015.

[8] ——, "Data compression with low distortion and finite blocklength," *ArXiv preprint*, Jan. 2016.

[9] I. Csiszár, "On an extremum problem of information theory," *Studia Scientiarum Mathematicarum Hungarica*, vol. 9, no. 1, pp. 57–71, Jan. 1974.

[10] R. Gallager, *Information theory and reliable communication*. John Wiley & Sons, Inc. New York, 1968.

[11] T. Berger, *Rate distortion theory*. Prentice-Hall, Englewood Cliffs, NJ, 1971.

[12] C. E. Shannon, "Probability of error for optimal codes in a Gaussian channel," *Bell Syst. Tech. J.*, vol. 38, no. 3, pp. 611–656, 1959.

[13] R. M. Gray, "Information rates of stationary ergodic finite-alphabet sources," *IEEE Transactions on Information Theory*, vol. 17, no. 5, pp. 516–523, 1971.

[14] L. Palzer and R. Timo, "A converse for lossy source coding in the finite blocklength regime," in *Proceedings International Zurich Seminar on Communications*, Mar. 2016, pp. 15–19.

[15] A. Gerrish and P. Schultheiss, "Information rates of non-Gaussian processes," *IEEE Transactions on Information Theory*, vol. 10, no. 4, pp. 265–271, Oct 1964.

[16] R. Gray, "Information rates of autoregressive processes," *IEEE Transactions on Information Theory*, vol. 16, no. 4, pp. 412 – 421, July 1970.

[17] W. Szpankowski and S. Verdú, "Minimum expected length of fixed-to-variable lossless compression without prefix constraints: memoryless sources," *IEEE Transactions on Information Theory*, vol. 57, no. 7, pp. 4017–4025, July 2011.

[18] 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.

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

[20] A. Ingber and Y. Kochman, "The dispersion of lossy source coding," in *Data Compression Conference (DCC)*, Snowbird, UT, Mar. 2011, pp. 53–62.

[21] V. Kostina and S. Verdú, "Nonasymptotic noisy lossy source coding," *IEEE Transactions on Information Theory*, vol. PP, 2016.

[22] I. Csiszár and Z. Talata, "Context tree estimation for not necessarily finite memory processes, via BIC and MDL," *IEEE Transactions on Information Theory*, vol. 52, no. 3, pp. 1007–1016, March 2006.

[23] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.