A new converse in rate-distortion theory

Victoria Kostina, Sergio Verdú
Dept. of Electrical Engineering, Princeton University, NJ, 08544, USA

***Abstract*—This paper shows new finite-blocklength converse bounds applicable to lossy source coding as well as joint source-channel coding, which are tight enough not only to prove the strong converse, but to find the rate-dispersion functions in both setups. In order to state the converses, we introduce the d-tilted information, a random variable whose expectation and variance (with respect to the source) are equal to the rate-distortion and rate-dispersion functions, respectively.**

***Index Terms*—Converse, finite blocklength regime, joint source-channel coding, lossy source coding, memoryless sources, rate-distortion theory, Shannon theory.**

I. Introduction

The fundamental problem in non-asymptotic rate-distortion theory is to estimate the minimum achievable source coding rate at any given blocklength k compatible with allowable distortion d. Similarly, in joint source-channel coding, channel blocklength n and distortion d are fixed, and one is interested in the maximum number of source samples k per channel block. While in certain special cases these non-asymptotic fundamental limits can be calculated precisely, in general no such formulas exist. Fortunately, non-asymptotic computable upper and lower bounds to the optimum code size can be obtained. This paper deals with converse, or impossibility, bounds that the rate of all codes sustaining a given fidelity level must satisfy.

The rest of the paper is organized as follows. Section II introduces the notion of d-tilted information. Sections III and IV focus on source coding and joint source-channel coding, respectively, while Section V discusses and generalizes the results of Sections III and IV. Section VI investigates the asymptotic analysis of the new converse bounds. Section VII considers lossy compression with side information at both compressor and decompressor.

II. d-tilted information

Denote by

$$\imath_{S;Z}(s;z) = \log \frac{dP_{Z|S=s}(z)}{dP_Z(z)} \tag{1}$$

the information density of the joint distribution P_{SZ} at $(s,z) \in \mathcal{M} \times \hat{\mathcal{M}}$. We can define the right side of (1) for a given $(P_{Z|S}, P_Z)$ even if there is no P_S such that the marginal of $P_S P_{Z|S}$ is P_Z. We use the same notation $\imath_{S;Z}$ for that more general function. Further, for a discrete random variable S, the information in outcome s is denoted by

$$\imath_S(s) = \log \frac{1}{P_S(s)} \tag{2}$$

Under appropriate conditions [1], the number of bits that it takes to represent s divided by $\imath_S(s)$ converges to 1 as these quantities go to infinity. Note that if S is discrete, then $\imath_{S;S}(s;s) = \imath_S(s)$.

For a given P_S and a distortion measure $\mathsf{d}\colon \mathcal{M} \times \hat{\mathcal{M}} \mapsto [0, +\infty]$, denote

$$\mathbb{R}_S(d) = \inf_{\substack{P_{Z|S}:\\ \mathbb{E}[\mathsf{d}(S,Z)] \le d}} I(S;Z) \tag{3}$$

We impose the following basic restrictions on the source and the distortion measure:

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

$$d_{\min} = \inf \{d\colon \ \mathbb{R}_S(d) < \infty\} \tag{4}$$

(b) The infimum in (3) is achieved by a unique $P_{Z^\star|S}$, and the distortion measure is finite-valued.[1]

The counterpart of (2) in lossy data compression, which in a certain (asymptotic) sense corresponds to the number of bits one needs to spend to encode s within distortion d, is the following.

Definition 1 ($\mathsf{d}-$tilted information)**.** *For $d > d_{\min}$, the d-tilted information in s is defined as*

$$\jmath_S(s,d) = \log \frac{1}{\mathbb{E}\left[\exp\left\{\lambda^\star d - \lambda^\star \mathsf{d}(s, Z^\star)\right\}\right]} \tag{5}$$

where the expectation is with respect to the unconditional distribution of $Z^\star$, and

$$\lambda^\star = -\mathbb{R}_S'(d) \tag{6}$$

A general proof of the following properties can be found using [2, Lemma 1.4].

Property 1. For $P_{Z^\star}$-almost every z,

$$\jmath_S(s,d) = \imath_{S;Z^\star}(s;z) + \lambda^\star \mathsf{d}(s,z) - \lambda^\star d \tag{7}$$

hence the name we adopted in Definition 1.

Property 2.

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

Property 3. For all $z \in \hat{\mathcal{M}}$,

$$\mathbb{E}\left[\exp\left\{\jmath_S(S,d) + \lambda^\star d - \lambda^\star \mathsf{d}(S,z)\right\}\right] \le 1 \tag{9}$$

This work was supported in part by the National Science Foundation (NSF) under Grant CCF-1016625 and by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under Grant CCF-0939370. The first author was supported in part by the Natural Sciences and Engineering Research Council of Canada.

[1] Restriction (b) is imposed for clarity of presentation. We will show in Section V that it can be dispensed with.

with equality for $P_{Z^\star}$-almost every z.

Remark 1. Using Property 1, it is easy to show that restriction (b) guarantees differentiability of $\mathbb{R}_S(d)$, thus (6) is well defined. Indeed, assume by contradiction that $\mathbb{R}_S(d)$ has two tangents at point d with slopes $-\lambda_1 \neq -\lambda_2$. Writing $\jmath_{S,1}(s,d)$ and $\jmath_{S,2}(s,d)$ for the corresponding d-tilted informations in (5), we invoke (7) to write

$$\imath_{S;Z^\star}(s,z) = \jmath_{S,1}(s,d) + \lambda_1 \mathsf{d}(s,z) - \lambda_1 d \tag{10}$$

$$= \jmath_{S,2}(s,d) + \lambda_2 \mathsf{d}(s,z) - \lambda_2 d \tag{11}$$

which implies that $\mathsf{d}(s,z)$ does not depend on z. But that means, via (7), that $\imath_{S;Z^\star}(s;z)$ does not depend on z, which leads to $P_{S|Z^\star} = P_S$, or $\mathbb{R}_S(d) = 0$. Thus $\mathbb{R}_S(d)$ can be non-differentiable only at the point where it vanishes.

Remark 2. While Definition 1 does not cover the case $d = d_{\min}$, for discrete random variables with $d(s,z) = 1\{s \neq z\}$ it is natural to define 0-tilted information as

$$\jmath_S(s,0) = \imath_S(s) \tag{12}$$

Example 1. Consider the stationary binary memoryless source (BMS) with $P_{\mathsf{S}}(1) = p \leq P_{\mathsf{S}}(0)$ and bit error rate distortion, $\mathsf{d}(s^k, z^k) = \frac{1}{k}\sum_{i=1}^k 1\{s_i \neq z_i\}$. It is well known that $P_{\mathsf{S}|\mathsf{Z}^\star}(1|0) = P_{\mathsf{S}|\mathsf{Z}^\star}(0|1) = d$, and a simple calculation using (7) and $\lambda^\star = k \log \frac{1-d}{d}$ yields, for $0 \leq d < p$,

$$\jmath_{S^k}(s^k, d) = \sum_{i=1}^k \imath_{\mathsf{S}}(s_i) - k h(d) \tag{13}$$

$$= w \log \frac{1}{p} + (k-w) \log \frac{1}{1-p} - k h(d) \tag{14}$$

where w is the number of ones in s^k, and $\jmath_{S^k}(s^k, d) = 0$ if $d \geq p$.

Example 2. For the stationary Gaussian memoryless source (GMS) with variance σ^2 and mean-square error distortion,[2]

$$\jmath_{S^k}(s^k, d) = \frac{k}{2} \log \frac{\sigma^2}{d} + \left(\frac{|s^k|^2}{\sigma^2} - k \right) \frac{\log e}{2} \tag{15}$$

if $0 < d < \sigma^2$, and 0 if $d \geq \sigma^2$.

The distortion d-ball around s is denoted by

$$B_d(s) = \{ z \in \hat{\mathcal{M}} \colon \ d(s,z) \leq d \} \tag{16}$$

The d-tilted information is closely related to the (unconditional) probability that $Z^\star$ falls within distortion d from S. Indeed, since $\lambda^\star > 0$, for an arbitrary P_Z we have by Markov's inequality,

$$P_Z(B_d(s)) = \mathbb{P}\left[d(s,Z) \leq d \right] \tag{17}$$

$$\leq \mathbb{E}\left[\exp\{ \lambda^\star d - \lambda^\star d(s,Z) \} \right] \tag{18}$$

where the probability measure is generated by the unconditional distribution of Z. Thus

$$\log \frac{1}{P_{Z^\star}(B_d(s))} \geq \jmath_S(s,d) \tag{19}$$

The "lossy AEP" [3] guarantees that under certain regularity conditions equality in (19) can be closely approached.

[2]We denote the Euclidean norm by $|\cdot|$, i.e. $|s^n|^2 = s_1^2 + \ldots + s_n^2$.

III. Lossy source coding

In fixed-length lossy compression, the output of a general source with alphabet $\mathcal{M}$ and source distribution P_S is mapped to one of the M codewords from the reproduction alphabet $\hat{\mathcal{M}}$. A lossy code is a (possibly randomized) pair of mappings $\mathsf{f}\colon \mathcal{M} \mapsto \{1, \ldots, M\}$ and $\mathsf{c}\colon \{1, \ldots, M\} \mapsto \hat{\mathcal{M}}$.

Let us introduce the following operational definition.

Definition 2. *An (M, d, ϵ) code for $\{\mathcal{M},\ \hat{\mathcal{M}},\ P_S,\ \mathsf{d}\}$ is a code with $|\mathsf{f}| = M$ such that $\mathbb{P}\left[\mathsf{d}\left(S, \mathsf{c}(\mathsf{f}(S))\right) > d\right] \leq \epsilon$.*

Note that in the conventional fixed-to-fixed (or block) setting $\mathcal{M}$ and $\hat{\mathcal{M}}$ are the $k-$fold Cartesian products of alphabets $\mathcal{S}$ and $\hat{\mathcal{S}}$, and an (M, d, ϵ) code for such a setting is called an (k, M, d, ϵ) code. We are now ready to state our nonasymptotic converse bound for lossy compression.

Theorem 1 (Converse, source coding)**.** *Fix $d > d_{\min}$. Any (M, d, ϵ) code must satisfy*

$$\epsilon \geq \sup_{\gamma \geq 0} \left\{ \mathbb{P}\left[\jmath_S(S,d) \geq \log M + \gamma \right] - \exp(-\gamma) \right\} \tag{20}$$

Proof: Let the compressor and the decompressor be the random transformations $P_{U|S}$ and $P_{Z|U}$, respectively, where U takes values in $\{1, \ldots, M\}$. Let Q_U be equiprobable on $\{1, \ldots, M\}$, and let $\bar{Z}$ be independent of S, with its distribution equal to the marginal of $P_{Z|U} Q_U$. We have, for any $\gamma \geq 0$

$$\mathbb{P}\left[\jmath_S(S,d) \geq \log M + \gamma \right] \tag{21}$$

$$= \mathbb{P}\left[\jmath_S(S,d) \geq \log M + \gamma, d(S,Z) > d \right]$$

$$+ \mathbb{P}\left[\jmath_S(S,d) \geq \log M + \gamma, d(S,Z) \leq d \right] \tag{22}$$

$$\leq \epsilon + \mathbb{P}\left[M \leq \exp\left(\jmath_S(S,d) - \gamma \right), d(S,Z) \leq d \right] \tag{23}$$

$$= \epsilon + \mathbb{P}\left[M \leq \exp\left(\jmath_S(S,d) - \gamma \right) 1\{ d(S,Z) \leq d \} \right] \tag{24}$$

$$\leq \epsilon + \frac{\exp(-\gamma)}{M} \mathbb{E}\left[\exp\left(\jmath_S(S,d) \right) 1\{ d(S,Z) \leq d \} \right] \tag{25}$$

$$\leq \epsilon + \frac{\exp(-\gamma)}{M} \mathbb{E}\left[\exp\left(\jmath_S(S,d) + \lambda^\star d - \lambda^\star \mathsf{d}(S,Z) \right) \right] \tag{26}$$

$$= \epsilon + \frac{\exp(-\gamma)}{M} \mathbb{E}\Bigg[\sum_{u=1}^M P_{U|S}(u|S)$$

$$\cdot \ \mathbb{E}\left[\exp\left(\jmath_S(S,d) + \lambda^\star d - \lambda^\star \mathsf{d}(S,Z) \right) | U = u, S \right] \Bigg] \tag{27}$$

$$\leq \epsilon + \exp(-\gamma) \, \mathbb{E}\Bigg[\sum_{u=1}^M \frac{1}{M}$$

$$\cdot \ \mathbb{E}\left[\exp\left(\jmath_S(S,d) + \lambda^\star d - \lambda^\star \mathsf{d}(S,Z) \right) | U = u, S \right] \Bigg] \tag{28}$$

$$= \epsilon + \exp(-\gamma) \, \mathbb{E}\left[\exp\left(\jmath_S(S,d) + \lambda^\star d - \lambda^\star \mathsf{d}(S,\bar{Z}) \right) \right] \tag{29}$$

$$\leq \epsilon + \exp(-\gamma) \tag{30}$$

where

- (25) is by Markov's inequality,
- (28) uses $P_{U|S}(u|s) \leq 1$,
- (29) is by the definition of $\bar{Z}$,
- (30) uses (9) averaged with respect to $P_{\bar{Z}}$.

∎

For lossy compression of a binary source with bit error rate distortion, a simple particularization of Theorem 1 using (14) leads to the following easily computable bound.

Corollary 1 (Converse, source coding, BMS)**.** *Consider a BMS with bias p and bit error rate distortion measure. Fix $0 \le d < p$, $0 < \epsilon < 1$. For any (k, M, d, ϵ) code,*

$$\epsilon \ge \sup_{\gamma \ge 0} \left\{ \mathbb{P}\left[g_k(W) \ge \log M + \gamma\right] - \exp\left(-\gamma\right) \right\} \quad (31)$$

$$g_k(w) = w \log \frac{1}{p} + (k - w) \log \frac{1}{1-p} - kh(d) \quad (32)$$

where W is binomial with success probability p and k degrees of freedom.

Similarly, the particularization of Theorem 1 to the Gaussian case using (15) yields the following result.

Corollary 2 (Converse, source coding, GMS)**.** *Consider a GMS with variance σ^2 and mean-square error distortion. Fix $0 < d < \sigma^2$ and $0 < \epsilon < 1$. For any (k, M, d, ϵ) code,*

$$\epsilon \ge \sup_{\gamma \ge 0} \left\{ \mathbb{P}\left[g_k(W) \ge \log M + \gamma\right] - \exp(-\gamma) \right\} \quad (33)$$

$$g_k(w) = \frac{k}{2} \log \frac{\sigma^2}{d} + \frac{w - k}{2} \log e \quad (34)$$

where $W \sim \chi_2^k$ (i.e. chi square distributed with k degrees of freedom).

IV. Lossy joint source-channel coding

A lossy source-channel code is a (possibly randomized) pair of mappings $\mathsf{f} \colon \mathcal{M} \mapsto \mathcal{X}$ and $\mathsf{c} \colon \mathcal{Y} \mapsto \hat{\mathcal{M}}$, where $\mathcal{X}$ and $\mathcal{Y}$ are the channel input and output alphabets, respectively.

Definition 3. *A (d, ϵ) code for $\{\mathcal{M},\ \mathcal{X},\ \mathcal{Y},\ \hat{\mathcal{M}},\ P_S,\ P_{Y|X},\ \mathsf{d}\}$ is a source-channel code with $\mathbb{P}\left[\mathsf{d}\left(S, \mathsf{c}(Y)\right) > d\right] \le \epsilon$ where $\mathsf{f}(S) = X$.*

The special case $d = 0$ and $\mathsf{d}(s, z) = 1\left\{s \neq z\right\}$ corresponds to almost-lossless compression. If, in addition, P_S is equiprobable on an alphabet of cardinality M, a $(0, \epsilon)$ code in Definition 3 corresponds to an (M, ϵ) channel code (defined in [4]). On the other hand, if the channel is an identity mapping on an alphabet of cardinality M, a (d, ϵ) code in Definition 3 corresponds to an (M, d, ϵ) lossy code (Definition 2).

In the conventional block setting $\mathcal{X}$ and $\mathcal{Y}$ are the $n-$fold Cartesian products of alphabets $\mathcal{A}$ and $\mathcal{B}$, $\mathcal{M}$ and $\hat{\mathcal{M}}$ are the $k-$fold Cartesian products of alphabets $\mathcal{S}$ and $\hat{\mathcal{S}}$, and a (d, ϵ) code for such a setting is called a (k, n, d, ϵ) code.

Theorem 1 can be generalized to the joint source-channel coding setup as follows.

Theorem 2 (Converse, joint source-channel coding)**.** *Fix $d > d_{\min}$. Any (d, ϵ) code for source S and channel $P_{Y|X}$ must satisfy*

$$\epsilon \ge \inf_{P_{X|S}} \sup_{\gamma > 0} \left\{ \sup_{P_{\bar{Y}}} \mathbb{P}\left[\jmath_S(S, d) - \imath_{X;\bar{Y}}(X; Y) \ge \gamma\right] - \exp\left(-\gamma\right) \right\} \quad (35)$$

$$\ge \sup_{\gamma > 0} \left\{ \sup_{P_{\bar{Y}}} \mathbb{E}\left[\inf_{x \in \mathcal{X}} \mathbb{P}\left[\jmath_S(S, d) - \imath_{X;\bar{Y}}(x; Y) \ge \gamma \mid S\right]\right] - \exp\left(-\gamma\right) \right\} \quad (36)$$

where in (35)*, $S - X - Y$, and in* (36)*, the probability measure is generated by $P_S P_{Y|X=x}$, and*

$$\imath_{X;\bar{Y}}(x; y) = \log \frac{dP_{Y|X}(y|x)}{dP_{\bar{Y}}(y)} \quad (37)$$

Proof: The encoder and the decoder are the random transformations $P_{X|S}$ and $P_{Z|Y}$, respectively. Fix an arbitrary $P_{\bar{Y}}$ on $\mathcal{Y}$. Let $\bar{Z}$ be independent of all other random variables, with distribution $dP_{\bar{Z}}(z) = \int dP_{Z|Y=y}(z) dP_{\bar{Y}}(y)$. We have,

$$\begin{aligned}
&\mathbb{P}\left[\jmath_S(S, d) - \imath_{X;\bar{Y}}(X, Y) \ge \gamma\right] \\
&\le \epsilon + \mathbb{P}\Big[\exp(\imath_{X;\bar{Y}}(X, Y)) \le \\
&\qquad \exp\left(\jmath_S(S, d) - \gamma\right) 1\left\{\mathsf{d}(S, Z) \le d\right\}\Big] && (38) \\
&\le \epsilon + \exp\left(-\gamma\right) \mathbb{E}\left[\exp\left(\jmath_S(S, d)\right) 1\left\{\mathsf{d}(S, \bar{Z}) \le d\right\}\right] && (39) \\
&\le \epsilon + \exp\left(-\gamma\right) \mathbb{E}\left[\exp\left(\jmath_S(S, d) + \lambda^\star d - \lambda^\star \mathsf{d}(S, \bar{Z})\right)\right] && (40) \\
&\le \epsilon + \exp\left(-\gamma\right) && (41)
\end{aligned}$$

where (38) is obtained in the same way as going from (21) to (24), (39) is by the change of measure argument, and (41) is due to (9). Optimizing over $\gamma > 0$ and $P_{\bar{Y}}$ to obtain the best possible bound for a given $P_{X|S}$, then choosing $P_{X|S}$ that gives the weakest bound, the code-independent converse (35) follows. To show (36), we first weaken (35) as,

$$\epsilon \ge \sup_{\gamma > 0} \left\{ \sup_{P_{\bar{Y}}} \inf_{P_{X|S}} \mathbb{P}\left[\jmath_S(S, d) - \imath_{X;\bar{Y}}(X; Y) \ge \gamma\right] - \exp\left(-\gamma\right) \right\} \quad (42)$$

and then observe that for any $P_{\bar{Y}}$,

$$\inf_{P_{X|S}} \mathbb{P}\left[\jmath_S(S, d) - \imath_{X;\bar{Y}}(X; Y) \ge \gamma\right] = \mathbb{E}\left[\inf_{x \in \mathcal{X}} \mathbb{P}\left[\jmath_S(S, d) - \imath_{X;\bar{Y}}(x; Y) \ge \gamma \mid S\right]\right] \quad (43)$$

∎

V. Theorems 1 and 2: discussion

A. Block coding

In the conventional block coding setting, fixing any three parameters in the quartet (k, M, d, ϵ) (in source coding) or (k, n, d, ϵ) (in source-channel coding), Theorems 1 and 2 can be used to compute converse bounds on the fourth parameter.

In addition, fixing the rate and d and letting the blocklength go to infinity, one can use Theorems 1 and 2 to obtain strong converses in their respective setups (Section VI-A). In the practically relevant setting in which design parameters d and ϵ are fixed, a simple Gaussian approximation to the optimum finite blocklenghth coding rate can be obtained via the analysis of the bounds in Theorems 1 and 2 using the Berry-Esseen Central Limit Theorem (Section VI-B).

B. Almost lossless data compression and transmission

Using (12), it is easy to check that both Theorem 1 and Theorem 2 still hold in the almost-lossless case, which corresponds to $d = 0$ and $d(s,z) = 1\{s \neq z\}$. In fact, Theorem 1 gives a pleasing generalization of the almost-lossless data compression converse bound [5], [6, Lemma 1.3.2]:

$$\epsilon \geq \sup_{\gamma \geq 0} \left\{ \mathbb{P}\left[\imath_S(S) \geq \log M + \gamma\right] - \exp(-\gamma) \right\} \tag{44}$$

Similarly, applying Theorem 2 to the almost-lossless source-channel coding setup, we conclude that any joint source-channel code with error probability not exceeding ϵ must satisfy (36) with $\jmath_S(S,d)$ replaced by $\imath_S(S)$, which can be regarded as a generalization of Wolfowitz's converse [7] to nonequiprobable source outputs [5].

C. Theorem 2 implies Theorem 1

Theorem 1 can be viewed as a particular case of the result in Theorem 2. Indeed, if $\mathcal{X} = \mathcal{Y} = \{1, \ldots, M\}$, $P_{Y|X}(m|m) = 1$, $P_{\bar{Y}}(1) = \ldots = P_{\bar{Y}}(M) = \frac{1}{M}$, then $\imath_{X;\bar{Y}}(x;Y) = \log M$ a.s. regardless of $x \in \mathcal{X}$, and (36) leads to (20).

D. Channels with input cost constraints

If the channel has input cost constraints, Theorem 2 still holds with the entire channel input alphabet $\mathcal{X}$ replaced by the subset of allowable channel inputs $\mathcal{F}$, where $\mathcal{F} \subseteq \mathcal{X}$.

E. Theorems 1 and 2 still hold even if condition (b) fails

Even if the rate-distortion function is not achieved by any output distribution, the definition of $\mathsf{d}-$tilted information can be extended appropriately, so that Theorems 1 and 2 still hold.

Instead of imposing restriction (b) of Section II, we require that the following condition holds.

(b′) The distortion measure is such that there exists a finite set $\mathcal{G} \subseteq \hat{\mathcal{M}}$ such that

$$\mathbb{E}\left[\min_{z \in \mathcal{G}} \mathsf{d}(S,z)\right] < \infty \tag{45}$$

Then the rate-distortion function admits the following general representation due to Csiszár [2].

Theorem 3 (Alternative representation of $\mathbb{R}(d)$ [2])**.** *Under the basic restrictions* (a), (b′), *for each $d > d_{\min}$, it holds that*

$$\mathbb{R}_S(d) = \max_{\alpha(s),\ \lambda} \left\{ \mathbb{E}\left[\alpha(S)\right] - \lambda d \right\} \tag{46}$$

where the maximization is over $\alpha(s) \geq 0$ and $\lambda \geq 0$ satisfying the constraint

$$\mathbb{E}\left[\exp\left\{\alpha(S) - \lambda d(S,z)\right\}\right] \leq 1 \ \forall z \in \hat{\mathcal{M}} \tag{47}$$

Let $(\alpha^\star(s),\ \lambda^\star)$ achieve the maximum in (9) for some $d > d_{\min}$, and define the $\mathsf{d}-$tilted information in s by

$$\jmath_S(s,d) = \alpha^\star(s) - \lambda^\star d \tag{48}$$

Note that (9), the only property of $\mathsf{d}-$tilted information we used in the proof of Theorems 1 and 2, still holds due to (47), thus both theorems remain true.

VI. Asymptotic analysis

In addition to the basic conditions (a), (b) of Section II, in this section we impose the following restrictions.

(i) The source is stationary and memoryless, $P_{S^k} = P_{\mathsf{S}} \times \ldots \times P_{\mathsf{S}}$.
(ii) The distortion measure is separable, $d(s^k, z^k) = \frac{1}{k}\sum_{i=1}^{k} d(s_i, z_i)$.
(iii) The distortion level satisfies $d > d_{\min}$, where $d_{\min}$ is defined in (4).
(iv) The supremum $\sup_{P_{\mathsf{X}}} I(\mathsf{X};\mathsf{Y})$ is attained by a unique $P_{\mathsf{X}^\star}$.
(v) The channel is stationary and memoryless, $P_{Y^n|X^n} = P_{\mathsf{Y}|\mathsf{X}} \times \ldots \times P_{\mathsf{Y}|\mathsf{X}}$, either discrete with finite input alphabet, or Gaussian with a maximal power constraint.

A. Strong converses

The strong converse for source coding on a general alphabet under a distortion constraint was previously obtained in [8]. In joint source-channel coding, if the source and the channel are both discrete or both Gaussian, strong converses can be obtained via the error exponent results in [9], [10], however no strong converses as general as that in [8] are known for joint source-channel coding. As we show next, Theorems 1 and 2 lead to very general strong converses in their respective setups.

Due to (i) and (ii), $P^\star_{Z^k} = P^\star_{\mathsf{Z}} \times \ldots \times P^\star_{\mathsf{Z}}$, and the $\mathsf{d}-$tilted information single-letterizes, that is, for a.e. s^k,

$$\jmath_{S^k}(s^k, d) = \sum_{i=1}^{k} \jmath_{\mathsf{S}}(s_i, d) \tag{49}$$

In the lossy compression setup, weakening (20) by choosing $\gamma = k\tau$ for some fixed $\tau > 0$ and choosing $\log M = kR(d) - 2\gamma$, we obtain

$$\epsilon \geq \left\{ \mathbb{P}\left[\sum_{i=1}^{k} \jmath_{\mathsf{S}}(S_i, d) \geq kR(d) - k\tau\right] - \exp(-k\tau) \right\} \tag{50}$$

Using (8), we conclude that the right side of (50) tends to 1 as k increases by the law of large numbers.

In the joint source-channel coding setup, we choose γ as above, choose $P_{\bar{\mathsf{Y}}} = P_{\mathsf{Y}^\star}$, where $P_{\mathsf{Y}^\star}$ is the channel output distribution generated by the capacity-achieving input distribution $P_{\mathsf{X}^\star}$, let $P_{Y^n} = P_{\mathsf{Y}^\star} \times \ldots \times P_{\mathsf{Y}^\star}$, and for each n, choose k so that $kR(d) = nC + 3\gamma$. Then, (36) weakens to

$$\begin{aligned} \epsilon & \\ \geq & \, \mathbb{E}\left[\inf_{x^n \in \mathcal{A}^n} \mathbb{P}\left[\sum_{i=1}^{k} \jmath_{\mathsf{S}}(S_i, d) - \sum_{j=1}^{n} \imath_{\mathsf{X};\mathsf{Y}^\star}(x_i; Y_i) \geq k\tau \mid S^k\right]\right] \\ - & \, \exp\left(-k\tau\right) \end{aligned} \tag{51}$$

$$\begin{aligned} \geq & \inf_{x^n \in \mathcal{A}^n} \mathbb{P}\left[\sum_{j=1}^{n} \imath_{\mathsf{X};\mathsf{Y}^\star}(x_i; Y_i) \leq nC + k\tau\right] \\ & \cdot \mathbb{P}\left[\sum_{i=1}^{k} \jmath_{\mathsf{S}}(S_i, d) \geq nC + 2k\tau\right] - \exp\left(-k\tau\right) \end{aligned} \tag{52}$$

Recalling (8) and $\mathbb{E}\left[\imath_{\mathsf{X};\mathsf{Y}^\star}(\mathsf{x}, \mathsf{Y}) | \mathsf{X} = \mathsf{x}\right] \leq C$ with equality for $P_{\mathsf{X}^\star}$-a.e. x, we conclude using the law of large numbers that the right side of (52) tends to 1 as $k, n \to \infty$ for a fixed ratio $\frac{k}{n} > \frac{C}{R(d)}$.

B. Gaussian approximation analysis

In addition to restrictions (i)-(v), we assume that the following holds.

(vi) The excess-distortion probability satisfies $0 < \epsilon < 1$.
(vii) The random variable $\jmath_{\mathsf{S}}(\mathsf{S}, d)$ has finite absolute third moment.

Theorem 4 (Gaussian approximation, source coding)**.** *Under restrictions* (i)–(iii), (vi)–(vii), *any* (k, M, d, ϵ) *code satisfies*

$$\log M \geq kR(d) + \sqrt{k\mathcal{V}(d)}Q^{-1}\left(\epsilon\right) - \frac{1}{2}\log k + O\left(1\right) \tag{53}$$

where $\mathcal{V}(d)$ *is the source rate-dispersion function [11] given by*

$$\mathcal{V}(d) = \mathrm{Var}\left[\jmath_{\mathsf{S}}(\mathsf{S}, d)\right] \tag{54}$$

Proof: Consider the case $\mathcal{V}(d) > 0$, so that

$$B = 6\frac{\mathbb{E}\left[\left|\jmath_{\mathsf{S}}(\mathsf{S}, d) - R(d)\right|^3\right]}{\left[\mathrm{Var}\left[\jmath_{\mathsf{S}}(\mathsf{S}, d)\right]\right]^{\frac{3}{2}}} \tag{55}$$

is finite according to restriction (vii), and Berry-Esseen bound [12, Ch. XVI.5 Theorem 2] applies to $\sum_{i=1}^{k} \jmath_{\mathsf{S}}(S_i, d)$. Let $\gamma = \frac{1}{2}\log k$ in (20), and choose

$$\log M = kR(d) + \sqrt{k\mathcal{V}(d)}Q^{-1}\left(\epsilon_k\right) - \gamma \tag{56}$$

$$\epsilon_k = \epsilon + \exp(-\gamma) + \frac{B}{\sqrt{k}} \tag{57}$$

so that $\log M$ can be written as the right side of (53). Substituting (49) and (56) in (20), we conclude that for any (M, d, ϵ') code it must hold that

$$\begin{aligned} \epsilon' \geq \mathbb{P}\left[\sum_{i=1}^{k} \jmath_{\mathsf{S}}(S_i, d) \geq kR(d) + \sqrt{k\mathcal{V}(d)}Q^{-1}\left(\epsilon_k\right)\right] \\ - \exp(-\gamma) \end{aligned} \tag{58}$$

The proof for $\mathcal{V}(d) > 0$ is complete upon noting that the right side of (58) is lower bounded by ϵ by the Berry-Esseen bound in view of (57).

If $\mathcal{V}(d) = 0$, it follows that $\jmath_{\mathsf{S}}(\mathsf{S}, d) = R(d)$ almost surely. Choosing $\gamma = \log\frac{1}{1-\epsilon}$ and $\log M = kR(d) - \gamma$ in (20) it is obvious that $\epsilon' \geq \epsilon$. ∎

Using Theorem 2, the following generalization of Theorem 4 can be shown for joint source-channel coding.

Theorem 5 (Gaussian approximation, joint source-channel coding)**.** *Under restrictions* (i)–(vii), *any* (k, n, d, ϵ) *code must satisfy*

$$nC - kR(d) \geq \sqrt{nV + k\mathcal{V}(d)}Q^{-1}\left(\epsilon\right) + O\left(\log(n+k)\right) \tag{59}$$

where V is the channel dispersion [4] given by

$$V = \mathrm{Var}\left[\imath_{\mathsf{X}^\star;\mathsf{Y}^\star}(\mathsf{X}^\star; \mathsf{Y}^\star)\right] \tag{60}$$

and $\mathcal{V}(d)$ *is the source dispersion given by* (54)*.*

Matching achievability results can be proven [11], [13] ensuring that the reverse inequalities in (53) and (59) also hold (up to the $O\left(\log k\right)$ term).

VII. Lossy source coding with side information

The results of Sections III and VI can be easily generalized to a setting in which both the compressor and the decompressor have access to side information $Y \in \mathcal{Y}$ which is statistically dependent on the source outcome $S \in \mathcal{M}$. A lossy code with side information at the compressor and the decompressor is a (possibly randomized) pair of mappings $\mathsf{f}\colon \mathcal{M} \times \mathcal{Y} \mapsto \{1, \ldots, M\}$ and $\mathsf{c}\colon \{1, \ldots, M\} \times \mathcal{Y} \mapsto \hat{\mathcal{M}}$.

Definition 4. *An* (M, d, ϵ) *code for* $\{\mathcal{M},\, \mathcal{Y},\, \hat{\mathcal{M}},\, P_{SY},\, \mathsf{d}\}$ *is a code with* $|\mathsf{f}| = M$ *such that* $\mathbb{P}\left[\mathsf{d}\left(S, \mathsf{c}(\mathsf{f}(S, Y), Y)\right) > d\right] \leq \epsilon$.

The notion of tilted information introduced in Section II is easily generalized when side information is available at the compressor and the decompressor. In parallel with (1), (2), denote

$$\imath_{S;Z|Y}(s; z|y) = \log\frac{dP_{SZ|Y=y}}{d(P_{S|Y=y} \times P_{Z|Y=y})}(s, z) \tag{61}$$

$$\imath_{S|Y}(s|y) = \log\frac{1}{P_{S|Y=y}(s)} \tag{62}$$

For a given P_{SY} and a distortion measure $\mathsf{d}\colon \mathcal{M} \times \hat{\mathcal{M}} \mapsto [0, +\infty]$, denote

$$\mathbb{R}_{S|Y}(d) = \inf_{\substack{P_{Z|S,Y}\colon \mathcal{M}\times\mathcal{Y}\mapsto\hat{\mathcal{M}} \\ \mathbb{E}[\mathsf{d}(S,Z)]\leq d}} I(S; Z|Y) \tag{63}$$

Under the same basic restrictions (a) and (b) in Section II (replacing $\mathbb{R}_S(d)$ by $\mathbb{R}_{S|Y}(d)$ and $P_{Z^\star|S}$ by $P_{Z^\star|SY}$), Definition 1 is generalized as follows.

Definition 5 ($\mathsf{d}-$tilted information with side information)**.** *For* $d > d_{\min}$*, the* d*-tilted information in s with side information y at the compressor and the decompressor is defined as*

$$\jmath_{S|Y}(s, d|y) = \log\frac{1}{\mathbb{E}\left[\exp\left\{\lambda^\star d - \lambda^\star \mathsf{d}(s, Z^\star)\right\} \mid Y = y\right]} \tag{64}$$

where the expectation is with respect to $P_{Z^\star|Y=y}$, *and*

$$\lambda^\star = -\mathbb{R}'_{S|Y}(d) \tag{65}$$

The counterparts of Properties 1–3 in Section II are stated below.

Property 1. For $P_{Z^\star|Y=y}$-almost every z,

$$\jmath_{S|Y}(s, d|y) = \imath_{S;Z^\star|Y}(s; z|y) + \lambda^\star \mathsf{d}(s, z) - \lambda^\star d \tag{66}$$

Property 2. $\mathbb{R}_{S|Y}(d) = \mathbb{E}\left[\jmath_{S|Y}(S, d|Y)\right]$

Property 3. For any $P_{Z|Y}$ where $S - Y - Z$,

$$\mathbb{E}\left[\exp\left\{\lambda^\star d - \lambda^\star \mathsf{d}(S, Z) + \jmath_{S|Y}(S, d|Y)\right\}\right] \leq 1 \tag{67}$$

In parallel with (12), for discrete random variables with $d(s, z) = 1\{s \neq z\}$ we define 0-tilted information with side information at the decompressor as

$$\jmath_{S|Y}(s, 0|y) = \imath_{S|Y}(s|y) \tag{68}$$

In the following stationary memoryless examples, $P_{S^k} = P_{\mathsf{S}} \times \ldots \times P_{\mathsf{S}}$ and $P_{Y^k} = P_{\mathsf{Y}} \times \ldots \times P_{\mathsf{Y}}$.

Example 3. If S and Y are binary equiprobable with bit error rate distortion and $\mathbb{P}[\mathsf{S} = 1|\mathsf{Y} = 0] = \mathbb{P}[\mathsf{S} = 0|\mathsf{Y} = 1] = p$, then

$$\jmath_{S^k|Y^k}(s^k, d|y^k) = \imath_{S^k|Y^k}(s^k|y^k) - kh(d) \tag{69}$$

if $0 \leq d < \min(p, 1-p)$, and 0 if $d \geq \min(p, 1-p)$.

Example 4. Let S and Y are zero-mean jointly Gaussian with variances σ^2_{S} and σ^2_{Y} and correlation coefficient $-1 < \rho < 1$. Denoting the variance of S conditioned on Y by $\sigma^2_{\mathsf{S}|\mathsf{Y}} = \sigma^2_{\mathsf{S}}(1 - \rho^2)$, we have

$$\jmath_{S^k|Y^k}(s^k, d|y^k) = \frac{k}{2}\log\frac{\sigma^2_{\mathsf{S}|\mathsf{Y}}}{d} + \left(\frac{|x^k - \rho\frac{\sigma_{\mathsf{S}}}{\sigma_{\mathsf{Y}}}y^k|^2}{\sigma^2_{\mathsf{S}|\mathsf{Y}}} - k\right)\frac{\log e}{2} \tag{70}$$

if $0 < d < \sigma^2_{\mathsf{S}|\mathsf{Y}}$, and 0 if $d \geq \sigma^2_{\mathsf{S}|\mathsf{Y}}$.

We now state the counterpart of Theorem 1.

Theorem 6 (Converse, source coding with side information)**.** *Fix* $d > d_{\min}$. *Any* (M, d, ϵ) *code must satisfy*

$$\epsilon \geq \sup_{\gamma \geq 0}\left\{\mathbb{P}\left[\jmath_{S|Y}(S, d|Y) \geq \log M + \gamma\right] - \exp(-\gamma)\right\} \tag{71}$$

Proof: Let the encoder and the decoder be transformations $P_{U|SY}$ and $P_{Z|UY}$, where U takes values in $\{1, \ldots, M\}$. Let $\bar{U}$ be equiprobable on $\{1, \ldots, M\}$ independent of all other random variables, and let $\bar{Z}$ be such that $S - Y - \bar{Z}$ and $P_{\bar{Z}|Y}$ is the marginal of $P_{\bar{U}} P_{Z|UY}$. We have, for any $\gamma \geq 0$

$$\mathbb{P}\left[\jmath_{S|Y}(S, d|Y) \geq \log M + \gamma\right]$$

$$\leq \epsilon + \frac{\exp(-\gamma)}{M}\mathbb{E}\Bigg[\sum_{u=1}^{M} P_{U|SY}(u|S, Y)$$

$$\cdot\, \mathbb{E}\left[\exp\left(\jmath_{S|Y}(S, d|Y)\right) 1\{\mathsf{d}(S, Z) \leq d\} \mid S, Y\right]\Bigg] \tag{72}$$

$$\leq \epsilon + \exp(-\gamma)\,\mathbb{E}\left[\exp\left(\jmath_{S|Y}(S, d|Y) + \lambda^\star d - \mathsf{d}(S, \bar{Z})\right)\right] \tag{73}$$

$$= \epsilon + \exp(-\gamma) \tag{74}$$

where (72) is obtained in the same way as going from (21) to (27), (73) is by $P_{U|SY}(u, s, y) \leq 1$ and the definition of $\bar{Z}$, and (74) uses (67). ■

Similar to Theorems 1 and 2, with the substitution in (68), Theorem 6 also applies to almost-lossless data compression (i.e. $d = 0$ and $d(s, z) = 1\{s \neq z\}$). In addition, Theorem 6 leads to a strong converse and a counterpart of Theorem 4 for source coding with side information.

VIII. Conclusion

Using the representation of the solution to the rate-distortion minimization problem (3) via the so-called d-tilted information, we have shown nonasymptotic general converse bounds for source coding (Theorem 1), joint source-channel coding (Theorem 2) and source coding with side information (Theorem 6). The new bounds are tight enough not only to show the strong converse, but to find the corresponding rate-dispersion functions once coupled with the achievability bounds in [11], [13].

References

[1] S. Verdú and T. S. Han, "The role of the asymptotic equipartition property in noiseless source coding," *IEEE Transactions on Information Theory*, vol. 43, no. 3, pp. 847–857, 1997.

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

[3] 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.

[4] 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.

[5] S. Verdú, "ELE528: Information theory lecture notes," *Princeton University*, 2009.

[6] T. S. Han, *Information spectrum methods in information theory*. Springer, Berlin, 2003.

[7] J. Wolfowitz, "Notes on a general strong converse," *Information and Control*, vol. 12, no. 1, pp. 1–4, 1968.

[8] J. Kieffer, "Strong converses in source coding relative to a fidelity criterion," *IEEE Transactions on Information Theory*, vol. 37, no. 2, pp. 257–262, Mar. 1991.

[9] I. Csiszár, "On the error exponent of source-channel transmission with a distortion threshold," *IEEE Transactions on Information Theory*, vol. 28, no. 6, pp. 823–828, 1982.

[10] Y. Zhong, F. Alajaji, and L. Campbell, "Joint source–channel coding excess distortion exponent for some memoryless continuous-alphabet systems," *IEEE Transactions on Information Theory*, vol. 55, no. 3, pp. 1296–1319, 2009.

[11] V. Kostina and S. Verdú, "Fixed-length lossy compression in the finite blocklength regime," *IEEE Transactions on Information Theory, to appear*, 2012.

[12] W. Feller, *An introduction to probability theory and its applications*, 2nd ed. John Wiley & Sons, 1971, vol. II.

[13] V. Kostina and S. Verdú, "Lossy joint source-channel coding in the finite blocklength regime," in *IEEE International Symposium on Information Theory*, Cambridge, MA, to appear, Jul. 2012.