Variable-length compression allowing errors

Victoria Kostina
Princeton University
vkostina@princeton.edu

Yury Polyanskiy
MIT
yp@mit.edu

Sergio Verdú
Princeton University
verdu@princeton.edu

***Abstract*—This paper studies the fundamental limits of the minimum average length of variable-length compression when a nonzero error probability ϵ is tolerated. We give non-asymptotic bounds on the minimum average length in terms of Erokhin's rate-distortion function and we use those bounds to obtain a Gaussian approximation on the speed of approach to the limit which is quite accurate for all but small blocklengths:**

$$(1-\epsilon)kH(\mathsf{S}) - \sqrt{\frac{kV(\mathsf{S})}{2\pi}}e^{-\frac{(Q^{-1}(\epsilon))^2}{2}}$$

where $Q^{-1}(\cdot)$ is the functional inverse of the Q-function and $V(\mathsf{S})$ is the source dispersion. A nonzero error probability thus not only reduces the asymptotically achievable rate by a factor of $1-\epsilon$, but also this asymptotic limit is approached from *below*, i.e. a larger source dispersion and shorter blocklengths are beneficial. Further, we show that variable-length lossy compression under excess distortion constraint also exhibits similar properties.

I. Introduction and summary of results

Let S be a discrete random variable to be compressed into a variable-length binary string. We denote the set of all binary strings (including the empty string) by $\{0,1\}^\star$ and the length of a string $a \in \{0,1\}^\star$ by $\ell(a)$. The codes considered in this paper fall under the following paradigm.

Definition 1 ((L,ϵ) code)**.** *A variable length (L,ϵ) code for source S defined on a finite or countably infinite alphabet $\mathcal{M}$ is a pair of possibly random transformations $P_{W|S}\colon \mathcal{M} \mapsto \{0,1\}^\star$ and $P_{\hat{S}|W}\colon \{0,1\}^\star \mapsto \mathcal{M}$ such that*[1]

$$\mathbb{P}\left[S \neq \hat{S}\right] \leq \epsilon \tag{1}$$

$$\mathbb{E}\left[\ell(W)\right] \leq L \tag{2}$$

The corresponding fundamental limit is

$$L_S^\star(\epsilon) \triangleq \inf\left\{L\colon \exists \textit{ an } (L,\epsilon) \textit{ code}\right\} \tag{3}$$

Lifting the prefix condition in variable-length coding is discussed in [1], [2]. In particular, in the zero-error case we have [3], [4]

$$H(S) - \log_2(H(S)+1) - \log_2 e \leq L_S^\star(0) \tag{4}$$

$$\leq H(S)\,, \tag{5}$$

while [1] shows that in the finite alphabet i.i.d. case (with a non-lattice distribution P_{S}, otherwise $o(1)$ becomes $O(1)$)

$$L_{S^k}^\star(0) = k\,H(\mathsf{S}) - \frac{1}{2}\log_2\left(8\pi e V(\mathsf{S})k\right) + o(1) \tag{6}$$

where $V(\mathsf{S})$ is the *varentropy* of P_{S}, namely the variance of the information

$$\imath_{\mathsf{S}}(\mathsf{S}) = \log_2 \frac{1}{P_{\mathsf{S}}(\mathsf{S})}. \tag{7}$$

Under the rubric of "weak variable-length source coding," T. S. Han [5], [6, Section 1.8] considers the asymptotic fixed-to-variable ($\mathcal{M} = \mathcal{S}^k$) almost-lossless version of the foregoing setup with vanishing error probability and prefix encoders. Among other results, Han showed that the minimum average length $L_{S^k}(\epsilon)$ of prefix-free encoding of a stationary ergodic source with entropy rate H behaves as

$$\lim_{\epsilon\to 0}\lim_{k\to\infty}\frac{1}{k}L_{S^k}(\epsilon) = H. \tag{8}$$

Koga and Yamamoto [7] treated variable-length prefix codes with non-vanishing error probability and showed that for finite alphabet i.i.d. sources with distribution P_{S},

$$\lim_{k\to\infty}\frac{1}{k}L_{S^k}(\epsilon) = (1-\epsilon)H(\mathsf{S}). \tag{9}$$

The benefit of variable length vs. fixed length in the case of given ϵ is clear from (9): indeed, the latter satisfies a strong converse and therefore any rate below the entropy is fatal. Allowing both nonzero error and variable-length coding is interesting not only conceptually but on account on several important generalizations. For example, the variable-length counterpart of Slepian-Wolf coding considered e.g. in [8] is particularly relevant in universal settings, and has a radically different (and practically uninteresting) zero-error version. Another substantive important generalization where nonzero error is inevitable is variable-length joint source-channel coding without or with feedback. For the latter, Polyanskiy et al. [9] showed that allowing a nonzero error probability boosts the ϵ-capacity of the channel, while matching the transmission length to channel conditions accelerates the rate of approach to that asymptotic limit. The use of nonzero error compressors is also of interest in hashing [10].

The purpose of this paper is to give non-asymptotic bounds on the fundamental limit (3), and to apply those bounds to analyze the speed of approach to the limit in (9), which also holds without the prefix condition. The key quantity in the

This work was supported in part by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under Grant CCF-0939370.

[1] Note that L need not be an integer.

non-asymptotic bounds is Erokhin's function [11]:

$$\mathbb{H}(S,\epsilon) \triangleq \min_{P_{Z|S}\colon \mathbb{P}[S\neq Z]\leq\epsilon} I(S;Z) \tag{10}$$

$$= \sum_{m=1}^{M} P_S(m)\log_2\frac{1}{P_S(m)} - (1-\epsilon)\log_2\frac{1}{1-\epsilon} - (M-1)\eta\log_2\frac{1}{\eta} \tag{11}$$

with the integer M and $\eta > 0$ determined by ϵ through

$$\sum_{m=1}^{M} P_S(m) = 1-\epsilon+(M-1)\eta \tag{12}$$

In particular, $\mathbb{H}(S,0) = H(S)$, and if S is equiprobable on an alphabet of M letters, then

$$\mathbb{H}(S,\epsilon) = \log_2 M - \epsilon\log_2(M-1) - h(\epsilon)\,, \tag{13}$$

where $h(x) = x\log_2\frac{1}{x} + (1-x)\log_2\frac{1}{1-x}$ is the binary entropy function.

Our non-asymptotic bounds are:

Theorem 1. *If $0 \leq \epsilon < 1 - P_S(1)$, then the minimum achievable average length satisfies*

$$\mathbb{H}(S,\epsilon) - \log_2(\mathbb{H}(S,\epsilon)+1) - \log_2 e \leq L_S^\star(\epsilon) \tag{14}$$

$$\leq \mathbb{H}(S,\epsilon) + \epsilon\log_2(H(S)+\epsilon) + \epsilon\log_2\frac{e}{\epsilon} + 2\,h(\epsilon). \tag{15}$$

If $\epsilon > 1 - P_S(1)$, then $L_S^\star(\epsilon) = 0$.

Note that we recover (4) and (5) by particularizing Theorem 1 to $\epsilon = 0$.

For memoryless sources we show that the speed of approach in the limit in (9) is given by the following result.

Theorem 2. *Assume that:*

- $P_{S^k} = P_\mathsf{S} \times \ldots \times P_\mathsf{S}$.
- $\mathbb{E}\left[|\imath_\mathsf{S}(\mathsf{S})|^3\right] < \infty$.

For any $0 \leq \epsilon \leq 1$ and $k \to \infty$ we have

$$\left.\begin{matrix} L_{S^k}^\star(\epsilon) \\ \mathbb{H}(S^k,\epsilon)\end{matrix}\right\} = (1-\epsilon)kH(\mathsf{S}) - \sqrt{\frac{kV(\mathsf{S})}{2\pi}}e^{-\frac{(Q^{-1}(\epsilon))^2}{2}} + \theta(k) \tag{16}$$

where Q^{-1} is the inverse of the complementary standard Gaussian cdf. The remainder term in (16) *satisfies*

$$-\log_2 k + O\left(1\right) \leq \theta(k) \leq O\left(1\right). \tag{17}$$

Therefore, not only $\epsilon > 0$ allows for a $(1-\epsilon)$ reduction in asymptotic rate (as found in [7]), but larger source dispersion is beneficial. This curious property is further discussed in Section III-A.

We also generalize the setting to allow a general distortion measure in lieu of the Hamming distortion in (1). More precisely, we replace (1) by the excess probability constraint $\mathbb{P}\left[\mathsf{d}\left(S,Z\right) > d\right] \leq \epsilon$. In this setting, refined asymptotics of minimum achievable lengths of variable-length lossy prefix codes almost surely operating at distortion d was studied in [12] (pointwise convergence) and in [13], [14] (convergence in mean). For fixed-length lossy compression, finite-blocklength bounds were shown in [15], and the asymptotic expansion for a stationary memoryless source in [15] (see also [16] for the finite-alphabet case). Our main result in that case is that (16) generalizes simply by replacing $H(\mathsf{S})$ and $V(\mathsf{S})$ by the corresponding rate-distortion and rate-dispersion functions.

II. Non-asymptotic bounds

A. Optimal code

In the zero-error case the optimum variable-length compressor without prefix constraints $\mathsf{f}_S^\star$ is known explicitly [3], [17][2]: a deterministic mapping that assigns the elements in $\mathcal{M}$ (labeled without loss of generality as the positive integers) ordered in decreasing probabilities to $\{0,1\}^\star$ ordered lexicographically. The decoder is just the inverse of this injective mapping. This code is optimal in the strong stochastic sense that the cumulative distribution function of the length of any other code cannot lie above that achieved with $\mathsf{f}_S^\star$. The length function of the optimum code is [3]:

$$\ell(\mathsf{f}_S^\star(m)) = \lfloor\log_2 m\rfloor. \tag{18}$$

In order to generalize this code to the nonzero-error setting, we take advantage of the fact that in our setting error detection is not required at the decoder. This allows us to retain the same decoder as in the zero-error case. As far as the encoder is concerned, to save on length on a given set of realizations which we are willing to fail to recover correctly, it is optimal to assign them all to $\varnothing$. Moreover, since we have the freedom to choose the set that we want to recover correctly (subject to a constraint on its probability $\geq 1-\epsilon$) it is optimal to include all the most likely realizations (whose encodings according to $\mathsf{f}_S^\star$ are shortest). If we are fortunate enough that ϵ is such that $\sum_{m=1}^{M} P_S(m) = 1-\epsilon$ for some M, then the code $\mathsf{f}(m) = \mathsf{f}_S^\star(m)$, if $m = 1,\ldots,M$ and $\mathsf{f}(m) = \varnothing$, if $m > M$ is optimal. To describe an optimum construction that holds without the foregoing fortuitous choice of ϵ, let M be the minimum M such that $\sum_{m=1}^{M} P_S(m) \geq 1-\epsilon$, let $\eta = \lfloor\log_2 M\rfloor$, and let $\mathsf{f}(m) = \mathsf{f}_S^\star(m)$, if $\lfloor\log_2 m\rfloor < \eta$ and $\mathsf{f}(m) = \varnothing$, if $\lfloor\log_2 m\rfloor > \eta$, and assign the outcomes with $\lfloor\log_2 m\rfloor = \eta$ to $\varnothing$ with probability α and to the lossless encoding $\mathsf{f}_S^\star(m)$ with probability $1-\alpha$, which is chosen so that[3]

$$\epsilon = \alpha \sum_{\substack{m\in\mathcal{M}\colon\\ \lfloor\log_2 m\rfloor = \eta}} P_S(m) + \sum_{\substack{m\in\mathcal{M}\colon\\ \lfloor\log_2 m\rfloor > \eta}} P_S(m) \tag{19}$$

$$= \mathbb{E}\left[\varepsilon^\star(S)\right] \tag{20}$$

[2] The construction in [17] omits the empty string.

[3] It does not matter exactly how the encoder implements randomization as long as conditioned on $\lfloor\log_2 S\rfloor = \eta$, the probability that S is mapped to $\varnothing$ is α. In the deterministic code with the fortuitous choice of ϵ described above, α is the ratio of the probabilities of the sets $\{m\in\mathcal{M}\colon m > M, \lfloor\log_2 m\rfloor = \eta\}$ to $\{m\in\mathcal{M}\colon \lfloor\log_2 m\rfloor = \eta\}$.

where

$$\varepsilon^\star(m) = \begin{cases} 0 & \ell(\mathsf{f}_S^\star(m)) < \eta \\ \alpha & \ell(\mathsf{f}_S^\star(m)) = \eta \\ 1 & \ell(\mathsf{f}_S^\star(m)) > \eta \end{cases} \tag{21}$$

We have shown that the output of the optimal encoder has structure

$$W(m) = \begin{cases} f_S^\star(m) & \langle \ell(\mathsf{f}_S^\star(m)) \rangle_\epsilon > 0 \\ \varnothing & \text{otherwise} \end{cases} \tag{22}$$

where the ϵ-cutoff random transformation acting on a real-valued random variable X is defined as

$$\langle X \rangle_\epsilon \triangleq \begin{cases} X & X < \eta \\ \eta & X = \eta \text{ (w. p. } 1-\alpha) \\ 0 & X = \eta \text{ (w. p. } \alpha) \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

where $\eta \in \mathbb{R}$ and $\alpha \in [0,1)$ are uniquely determined from

$$\mathbb{P}\left[X > \eta\right] + \alpha\, \mathbb{P}\left[X = \eta\right] = \epsilon \tag{24}$$

We have also shown that the minimum average length is given by

$$\begin{aligned} L_S^\star(\epsilon) &= \mathbb{E}\left[\langle \ell(\mathsf{f}_S^\star(S)) \rangle_\epsilon\right] && (25) \\ &= L_S^\star(0) - \max_{\varepsilon(\cdot) : \mathbb{E}\left[\varepsilon(S)\right] \le \epsilon} \mathbb{E}\left[\varepsilon(S) \ell(\mathsf{f}_S^\star(S))\right] && (26) \\ &= L_S^\star(0) - \mathbb{E}\left[\varepsilon^\star(S) \ell(\mathsf{f}_S^\star(S))\right] && (27) \end{aligned}$$

where the optimization is over $\varepsilon \colon \mathbb{Z}^+ \mapsto [0,1]$, and the optimal error profile $\varepsilon^\star(\cdot)$ that achieves (26) is given by (21).

An immediate consequence is that in the region of large error probability $\epsilon > 1 - P_S(1)$, $M = 1$, all outcomes are mapped to $\varnothing$, and therefore, $L_S^\star(\epsilon) = 0$. At the other extreme, if $\epsilon = 0$, then $M = |\mathcal{M}|$ and [2]

$$L_S^\star(0) = \mathbb{E}[\ell(\mathsf{f}_S^\star(S))] = \sum_{i=1}^{\infty} \mathbb{P}[S \ge 2^i] \tag{28}$$

B. Achievability bound

In principle, it may seem surprising that $L_S^\star(\epsilon)$ is connected to $\mathbb{H}(S,\epsilon)$ in the way dictated by Theorem 1, which implies that whenever the unnormalized quantity $\mathbb{H}(S,\epsilon)$ is large it must be close to the minimum average length. After all, the objectives of minimizing the input/output dependence and minimizing the description length of $\hat{S}$ appear to be disparate, and in fact (22) and the conditional distribution achieving (10) are quite different: although in both cases S and its approximation coincide on the most likely outcomes, the number of retained outcomes is different, and to lessen dependence, errors in the optimizing conditional in (22) do not favor $m = 1$ or any particular outcome of S. To prove (15), we let $E = 1\{S \neq \hat{S}\}$ and proceed to lower bound the mutual information between S and $\hat{S}$:

$$\begin{aligned} I(S;\hat{S}) &= I(S;\hat{S},\ell(\mathsf{f}_S^\star(S)) - I(S;\ell(\mathsf{f}_S^\star(S))|\hat{S}) && (29) \\ &= H(S) - H(\ell(\mathsf{f}_S^\star(S))|\hat{S}) - H(S|\hat{S},\ell(\mathsf{f}_S^\star(S))) && (30) \\ &= H(S) - H(\ell(\mathsf{f}_S^\star(S))|\hat{S},E) \\ &\quad - I(E;\ell(\mathsf{f}_S^\star(S))|\hat{S}) - H(S|\hat{S},\ell(\mathsf{f}_S^\star(S)) && (31) \\ &\ge L_S^\star(\epsilon) + H(S) - L_S^\star(0) - \epsilon \log_2(H(S) + \epsilon) \\ &\quad - \epsilon \log_2 \frac{e}{\epsilon} - 2\, h(\epsilon) && (32) \end{aligned}$$

where (32) follows from $(E;\ell(\mathsf{f}_S^\star(S))|\hat{S}) \le h(\epsilon)$ and the following chains (33)-(35) and (37)-(41).

$$\begin{aligned} &H(S|\hat{S},\ell(\mathsf{f}_S^\star(S))) \\ &\le \mathbb{E}\left[\varepsilon^\star(\ell(\mathsf{f}_S^\star(S)))\ell(\mathsf{f}_S^\star(S))\right] + \mathbb{E}\left[h(\varepsilon^\star(\ell(\mathsf{f}_S^\star(S))))\right] && (33) \\ &= L_S^\star(0) - L_S^\star(\epsilon) + \mathbb{E}\left[h(\varepsilon^\star(\ell(\mathsf{f}_S^\star(S))))\right] && (34) \\ &\le L_S^\star(0) - L_S^\star(\epsilon) + h(\epsilon) && (35) \end{aligned}$$

where (34) follows from (25); (35) follows from (20) and the concavity of $h(\cdot)$; and (33) is by Fano's inequality: since conditioned on $\ell(\mathsf{f}_S^\star(S)) = i$, S can have at most 2^i values:

$$H(S|\hat{S},\ell(\mathsf{f}_S^\star(S)) = i) \le \; i\,\varepsilon^\star(i) + h(\varepsilon^\star(i)) \tag{36}$$

Consider next the second quantity in (31)

$$\begin{aligned} &\frac{1}{\epsilon} H(\ell(\mathsf{f}_S^\star(S))|\hat{S},E) \\ &\le H(\ell(\mathsf{f}_S^\star(S))|\hat{S},E=1) && (37) \\ &\le H(\ell(\mathsf{f}_S^\star(S))|S \neq \hat{S}) && (38) \\ &\le \log_2(1 + \mathbb{E}\left[\ell(\mathsf{f}_S^\star(S))|S \neq \hat{S}\right]) + \log_2 e && (39) \\ &\le \log_2\left(1 + \frac{\mathbb{E}\left[\ell(\mathsf{f}_S^\star(S))\right]}{\epsilon}\right) + \log_2 e && (40) \\ &\le \log_2 \frac{e}{\epsilon} + \log_2(\epsilon + H(S))\,, && (41) \end{aligned}$$

where (37) follows since $H(\ell(\mathsf{f}_S^\star(S))|\hat{S},E=0) = 0$, (38) is because conditioning decreases entropy, (39) follows by maximizing entropy under the mean constraint (achieved by the geometric distribution), (40) follows by upper-bounding

$$\mathbb{P}[S \neq \hat{S}]\, \mathbb{E}\left[\ell(\mathsf{f}_S^\star(S))|S \neq \hat{S}\right] \le \mathbb{E}\left[\ell(\mathsf{f}_S^\star(S))\right]$$

and (41) uses (5) and (28). Finally, since the right side of (32) does not depend on $\hat{S}$, it must be a lower bound on the minimum (10):

$$\begin{aligned} \mathbb{H}(S,\epsilon) &\ge L_S^\star(\epsilon) + H(S) - L_S^\star(0) - \epsilon \log_2(H(S) + \epsilon) \\ &\quad - \epsilon \log_2 \frac{e}{\epsilon} - 2\, h(\epsilon) && (42) \end{aligned}$$

which leads to (15) via (5).

C. Converse

The entropy of the output string $W \in \{0,1\}^\star$ of the optimum random compressor in Section II-A satisfies

$$H(W) \ge I(S;W) = I(S;\hat{S}) \ge \mathbb{H}(S,\epsilon) \tag{43}$$

where the rightmost inequality holds in view of (10) and $\mathbb{P}[S \neq \hat{S}] = \epsilon$. Noting that the identity mapping $W \mapsto W \mapsto W$ is a lossless variable-length code, we lower-bound its average length as

$$
\begin{aligned}
H(W) - \log_2(H(W)+1) - \log_2 e &\leq L_W^\star(0) && (44)\\
&\leq \mathbb{E}[\ell(W)] && (45)\\
&= L_S^\star(\epsilon) && (46)
\end{aligned}
$$

where (44) follows from (4). The function of $H(W)$ in the left side of (44) is monotonically increasing if $H(W) > \log_2 \frac{e}{2} = 0.44$ bits and it is positive if $H(W) > 3.66$ bits. Therefore, it is safe to further weaken the bound in (44) by invoking (43). This concludes the proof of (14). By applying [1, Theorem 1] to W, we can get a sharper lower bound (which is always positive)

$$\psi^{-1}(\mathbb{H}(S,\epsilon)) \leq L_S^\star(\epsilon) \tag{47}$$

where ψ^{-1} is the inverse of the monotonic function on the positive real line:

$$\psi(x) = x + (1+x)\log_2(1+x) - x\log_2 x. \tag{48}$$

D. ϵ-cutoff of information

We show that

$$
\begin{aligned}
L_S^\star(\epsilon) &\approx \mathbb{H}(S,\epsilon) \pm \log H(S) && (49)\\
&\approx \mathbb{E}\left[\langle \imath_S(S)\rangle_\epsilon\right] \pm \log H(S) && (50)
\end{aligned}
$$

The random variable $\langle \imath_S(S)\rangle_\epsilon$ is a gateway to a refined asymptotic analysis of the quantities $L_S^\star(\epsilon)$ and $\mathbb{H}(S,\epsilon)$, both of whose analyses are challenging.

Similarly to (26), we have the variational characterization:

$$\mathbb{E}\left[\langle \imath_S(S)\rangle_\epsilon\right] = H(S) - \max_{\varepsilon(\cdot)\colon \mathbb{E}\left[\varepsilon(S)\right]\leq\epsilon} \mathbb{E}\left[\varepsilon(S)\imath_S(S)\right] \tag{51}$$

where $\varepsilon(\cdot)$ takes values in $[0,1]$.

Noting that the ordering $P_S(1) \geq P_S(2) \geq \ldots$ implies

$$\lfloor \log_2 m \rfloor \leq \imath_S(m) \tag{52}$$

and comparing (26) and (51), we obtain via (52):

$$
\begin{aligned}
\mathbb{E}\left[\langle \imath_S(S)\rangle_\epsilon\right] + L_S^\star(0) - H(S) &\leq L_S^\star(\epsilon) && (53)\\
&\leq \mathbb{E}\left[\langle \imath_S(S)\rangle_\epsilon\right] && (54)
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
&\mathbb{E}\left[\langle \imath_S(S)\rangle_\epsilon\right] - \epsilon\log(H(S)+\epsilon) - 2\,h(\epsilon) - \epsilon\log\frac{e}{\epsilon}\\
&\leq \mathbb{H}(S,\epsilon) && (55)\\
&\leq \mathbb{E}\left[\langle \imath_S(S)\rangle_\epsilon\right] && (56)
\end{aligned}
$$

Indeed, (55) follows from (42) and (53). Showing (56) involves defining a suboptimal choice (in (10)) of

$$Z = \begin{cases} S & \langle \imath_S(S)\rangle_\epsilon > 0\\ \bar{S} & \langle \imath_S(S)\rangle_\epsilon = 0 \end{cases} \tag{57}$$

where $P_{S\bar{S}} = P_S P_S$.

III. Asymptotics for memoryless sources

If the source is memoryless, the information in S^k is a sum of i.i.d. random variables

$$\imath_{S^k}(S^k) = \sum_{i=1}^{k} \imath_{\mathsf{S}}(S_i) \tag{58}$$

and Theorem 2 follows via an application of the following lemma to the bounds in (53)–(56).

Lemma 1. *Let $X_1, X_2, \ldots$ be a sequence of independent random variables with a common distribution P_{X} and a finite third absolute moment. Then for any $0 \leq \epsilon \leq 1$ and $k \to \infty$ we have*

$$
\begin{aligned}
\mathbb{E}\left[\left\langle \sum_{i=1}^{k} X_i \right\rangle_\epsilon\right] &= (1-\epsilon)k\mathbb{E}\left[\mathsf{X}\right] - \sqrt{\frac{k\mathrm{Var}\left[\mathsf{X}\right]}{2\pi}}e^{-\frac{(Q^{-1}(\epsilon))^2}{2}}\\
&\quad + O\left(1\right) && (59)
\end{aligned}
$$

Because of space limitations we refer the reader to [18] for the proof of Lemma 1. To gain some insight into the form of (59), note that if X is Gaussian, then

$$\mathbb{E}\left[\langle X\rangle_\epsilon\right] = (1-\epsilon)\mathbb{E}\left[X\right] - \sqrt{\frac{\mathrm{Var}\left[X\right]}{2\pi}}e^{-\frac{(Q^{-1}(\epsilon))^2}{2}}, \tag{60}$$

while the central limit theorem suggests

$$\sum_{i=1}^{k} X_i \overset{d}{\approx} \mathcal{N}\left(k\mathbb{E}\left[\mathsf{X}\right], k\mathrm{Var}\left[\mathsf{X}\right]\right) \tag{61}$$

A. Discussion

Like (6), but in contrast to [15], [19], [20], Theorem 2 exhibits an unusual phenomenon in which the dispersion term improves the achievable average rate. As illustrated in Fig. 1, a nonzero error probability ϵ decreases the average achievable rate as the source outcomes falling into the shaded area are assigned length 0. The more stretched the distribution of the encoded length the bigger is the gain, thus, remarkably, shorter blocklengths allow to achieve a lower average rate.

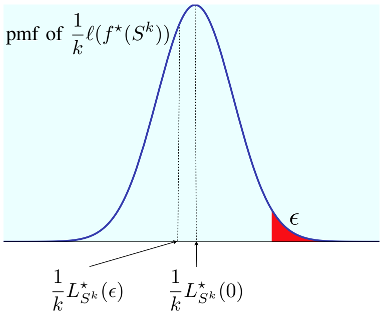
pmf of $\frac{1}{k}\ell(f^\star(S^k))$
ϵ
$\frac{1}{k}L_{S^k}^\star(\epsilon)$
$\frac{1}{k}L_{S^k}^\star(0)$

Fig. 1. The benefit of nonzero ϵ and dispersion.

For a source of biased coin flips, Fig. 2 depicts the exact average rate of the optimal code as well as the approximation in (16) in which the remainder $\theta(k)$ is taken to be that in [21, (9)]. Both curves are monotonically increasing in k.

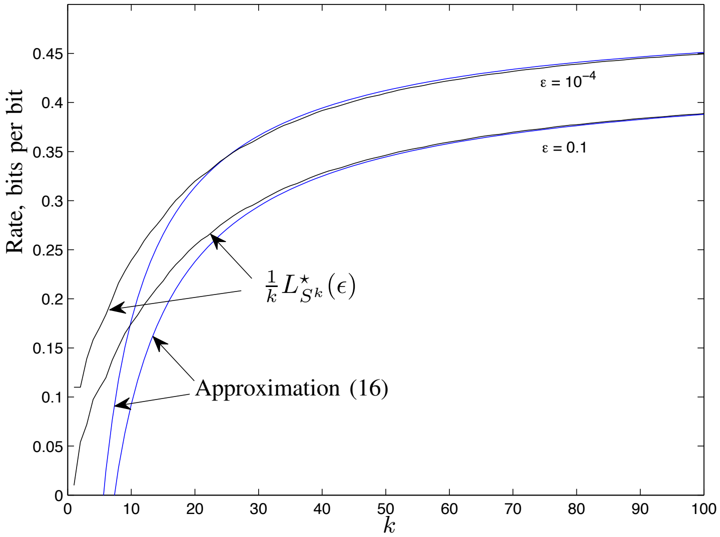

Fig. 2. Average rate achievable for variable-rate almost lossless encoding of a memoryless binary source with bias $p = 0.11$ and two values of ϵ. For $\epsilon < 10^{-4}$, the resulting curves are almost indistinguishable from the $\epsilon = 10^{-4}$ curve.

IV. Lossy variable-length compression

In the basic setup of lossy compression, we are given a source alphabet $\mathcal{M}$, a reproduction alphabet $\widehat{\mathcal{M}}$, a *distortion measure* $\mathsf{d} \colon \mathcal{M} \times \widehat{\mathcal{M}} \mapsto [0, +\infty]$ to assess the fidelity of reproduction, and a probability distribution of the object S to be compressed.

Definition 2 ((L, d, ϵ) code)**.** *A variable-length* (L, d, ϵ) *lossy code for* $\{S, \mathsf{d}\}$ *is a pair of random transformations* $P_{W|S} \colon \mathcal{M} \mapsto \{0,1\}^\star$ *and* $P_{Z|W} \colon \{0,1\}^\star \mapsto \mathcal{M}$ *such that*

$$\mathbb{P}\left[\mathsf{d}\left(S, Z\right) > d\right] \le \epsilon \tag{62}$$

$$\mathbb{E}\left[\ell(W)\right] \le L \tag{63}$$

The fundamental limit of interest is the minimum achievable average length compatible with the given tolerable error ϵ:

$$L^\star_S(d, \epsilon) \triangleq \inf\left\{L \colon \exists \text{ an } (L, d, \epsilon) \text{ code}\right\} \tag{64}$$

Denote the minimal mutual information quantity

$$\mathbb{R}_S(d, \epsilon) \triangleq \min_{\substack{P_{Z|S}\colon \\ \mathbb{P}[\mathsf{d}(S,Z) > d] \le \epsilon}} I(S; Z) \tag{65}$$

We assume that the following assumptions are satisfied.

(i) The source $\{S_i\}$ is stationary and memoryless.
(ii) The distortion measure is separable.
(iii) The distortion level satisfies $d_{\min} < d < d_{\max}$, where $d_{\min} = \inf\{d \colon R(d) < \infty\}$, and $d_{\max} = \inf_{\mathsf{z} \in \widehat{\mathcal{M}}} \mathbb{E}\left[\mathsf{d}(\mathsf{S}, \mathsf{z})\right]$, where the expectation is with respect to the unconditional distribution of S.
(iv) $\mathbb{E}\left[\mathsf{d}^{12}(\mathsf{S}, \mathsf{Z}^\star)\right] < \infty$ where the expectation is with respect to $P_\mathsf{S} \times P_{\mathsf{Z}^\star}$, where $\mathsf{Z}^\star$ achieves the rate-distortion function.

The lossy counterpart of Theorem 2 is the following (for the proof see [18]).

Theorem 3. *Assume that the rate-distortion function* $R(d)$ *is achieved by* $\mathsf{Z}^\star$*. Under assumptions* (i)–(iv)*, for any* $0 \le \epsilon \le 1$

$$\left.\begin{array}{l} L^\star_{S^k}(d, \epsilon) \\ \mathbb{R}_{S^k}(d, \epsilon) \end{array}\right\} = (1-\epsilon)kR(d) - \sqrt{\frac{k\mathcal{V}(d)}{2\pi}}\, e^{-\frac{(Q^{-1}(\epsilon))^2}{2}} + O\left(\log k\right) \tag{66}$$

where $\mathcal{V}(d)$ *is the rate-dispersion function [15].*

References

[1] 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, 2011.
[2] S. Verdú and I. Kontoyiannis, “Optimal lossless data compression: Non-asymptotics and asymptotics,” *IEEE Transactions on Information Theory*, vol. 60, no. 2, pp. 777–795, Feb. 2014.
[3] N. Alon and A. Orlitsky, “A lower bound on the expected length of one-to-one codes,” *IEEE Transactions on Information Theory*, vol. 40, no. 5, pp. 1670–1672, 1994.
[4] A. D. Wyner, “An upper bound on the entropy series,” *Inf. Contr.*, vol. 20, no. 2, pp. 176–181, 1972.
[5] T. S. Han, “Weak variable-length source coding,” *IEEE Transactions on Information Theory*, vol. 46, no. 4, pp. 1217–1226, 2000.
[6] ——, *Information spectrum methods in information theory.* Springer, Berlin, 2003.
[7] H. Koga and H. Yamamoto, “Asymptotic properties on codeword lengths of an optimal fixed-to-variable code for general sources,” *IEEE Transactions on Information Theory*, vol. 51, no. 4, pp. 1546–1555, April 2005.
[8] A. Kimura and T. Uyematsu, “Weak variable-length Slepian-Wolf coding with linked encoders for mixed sources,” *IEEE Transactions on Information Theory*, vol. 50, no. 1, pp. 183–193, 2004.
[9] Y. Polyanskiy, H. V. Poor, and S. Verdú, “Feedback in the non-asymptotic regime,” *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 4903–4925, 2011.
[10] H. Koga, “Source coding using families of universal hash functions,” *IEEE Transactions on Information Theory*, vol. 53, no. 9, pp. 3226–3233, Sep. 2007.
[11] V. Erokhin, “Epsilon-entropy of a discrete random variable,” *Theory of Probability and Applications*, vol. 3, pp. 97–100, 1958.
[12] 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.
[13] 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.
[14] E. Yang and Z. Zhang, “On the redundancy of lossy source coding with abstract alphabets,” *IEEE Transactions on Information Theory*, vol. 45, no. 4, pp. 1092–1110, May 1999.
[15] 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.
[16] A. Ingber and Y. Kochman, “The dispersion of lossy source coding,” in *Data Compression Conference (DCC)*, Snowbird, UT, Mar. 2011, pp. 53–62.
[17] S. Leung-Yan-Cheong and T. Cover, “Some equivalences between shannon entropy and kolmogorov complexity,” *IEEE Transactions on Information Theory*, vol. 24, no. 3, pp. 331–338, 1978.
[18] V. Kostina, Y. Polyanskiy, and S. Verdú, “Variable-length compression allowing errors (extended),” *ArXiv preprint*, Jan. 2014.
[19] V. Strassen, “Asymptotische abschätzungen in Shannon’s informationstheorie,” in *Proceedings 3rd Prague Conference on Information Theory*, Prague, 1962, pp. 689–723.
[20] 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.
[21] W. Szpankowski, “A one-to-one code and its anti-redundancy,” *IEEE Transactions on Information Theory*, vol. 54, no. 10, pp. 4762–4766, 2008.