Variable-length compression allowing errors

Victoria Kostina, Yury Polyanskiy, Sergio Verdú

Abstract

This paper studies the fundamental limits of the minimum average length of lossless and lossy variable-length compression, allowing a nonzero error probability ϵ, for lossless compression. 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 standard Gaussian complementary cdf, 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 this asymptotic limit is approached from *below*, i.e. a larger source dispersion and shorter blocklengths are beneficial. Variable-length lossy compression under excess distortion constraint is shown to exhibit similar properties.

Index Terms

Variable-length compression, lossless compression, lossy compression, single-shot, finite-blocklength regime, rate-distortion theory, dispersion, Shannon theory.

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.

This work was supported in part by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under Grant CCF-0939370. This paper was presented in part at ISIT 2014 [1].

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

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

Lifting the prefix condition in variable-length coding is discussed in [2], [3]. In particular, in the zero-error case we have [4], [5]

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

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

while [2] shows that in the 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 [6], [7, 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

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

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

Koga and Yamamoto [8] characterized asymptotically achievable rates of variable-length prefix codes with non-vanishing error probability and, in particular, 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 [9] 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. [10] 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 [11].

The purpose of Section II 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. Specifically, we show that (cf. (4)–(5))

$$L^{\star}_S(\epsilon) = \mathbb{H}(S,\epsilon) + O\left(\log_2 H(S)\right) \tag{10}$$

$$= \mathbb{E}\left[\langle \imath_S(S)\rangle_\epsilon\right] + O\left(\log_2 H(S)\right) \tag{11}$$

where

$$\mathbb{H}(S,\epsilon) \stackrel{\triangle}{=} \min_{\substack{P_{Z|S}:\\ \mathbb{P}[S\neq Z]\le\epsilon}} I(S;Z) \tag{12}$$

is Erokhin's function [12], and the ϵ-cutoff random transformation acting on a real-valued random variable X is defined as

$$\langle X\rangle_\epsilon \triangleq \begin{cases} X & X<\eta \\ \eta & X=\eta \text{ (w. p. } 1-\alpha) \\ 0 & X=\eta \text{ (w. p. } \alpha) \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

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

$$\mathbb{P}[X > \eta] + \alpha \, \mathbb{P}[X = \eta] = \epsilon. \tag{14}$$

While η and α satisfying (14) are not unique in general, any such pair defines the same $\langle X \rangle_\epsilon$ up to almost-sure equivalence.

The code that achieves (10) essentially discards "rich" source realizations with $\imath_S(S) > \eta$ and encodes the rest losslessly assigning them in the order of decreasing probabilities to the elements of $\{0,1\}^\star$ ordered lexicographically.

For memoryless sources with $S_i \sim \mathsf{S}$ we show that the speed of approach to the limit in (9) is given by the following result.

$$\left.\begin{array}{l} L^\star_{S^k}(\epsilon) \\ \mathbb{H}(S^k, \epsilon) \\ \mathbb{E}\left[\left\langle \imath_{S^k}(S^k) \right\rangle_\epsilon\right] \end{array}\right\} = (1-\epsilon)kH(\mathsf{S}) - \sqrt{\frac{kV(\mathsf{S})}{2\pi}} e^{-\frac{(Q^{-1}(\epsilon))^2}{2}} + O(\log k) \tag{15}$$

To gain some insight into the form of (15), note that if the source is memoryless, the information in S^k is a sum of i.i.d. random variables, and by the central limit theorem

$$\imath_{S^k}(S^k) = \sum_{i=1}^k \imath_{\mathsf{S}}(S_i) \tag{16}$$

$$\overset{d}{\approx} \mathcal{N}(kH(\mathsf{S}), kV(\mathsf{S})) \tag{17}$$

while for Gaussian X

$$\mathbb{E}[\langle X \rangle_\epsilon] = (1-\epsilon)\mathbb{E}[X] - \sqrt{\frac{\mathrm{Var}[X]}{2\pi}} e^{-\frac{(Q^{-1}(\epsilon))^2}{2}} \tag{18}$$

Our result in (15) underlines that not only does $\epsilon > 0$ allow for a $(1-\epsilon)$ reduction in asymptotic rate (as found in [8]), but, in contrast to [13]–[16], larger source dispersion is beneficial. This curious property is further discussed in Section II-E.

In Section III, we 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 $\mathbb{P}[\mathsf{d}(S,Z) > d] \leq \epsilon$. In this setting, refined asymptotics of minimum achievable lengths of variable-length lossy prefix codes almost surely operating at distortion d was studied in [17] (pointwise convergence) and in [18], [19] (convergence in mean). Our main result in the lossy

case is that (15) generalizes simply by replacing $H(\mathsf{S})$ and $V(\mathsf{S})$ by the corresponding rate-distortion and rate-dispersion functions, replacing Erokhin's function by

$$\mathbb{R}_S(d,\epsilon) \triangleq \min_{\substack{P_{Z|S}:\\ \mathbb{P}[\mathsf{d}(S,Z)>d]\le\epsilon}} I(S;Z), \tag{19}$$

and replacing the ϵ-cutoff of information by that of d-tilted information [15], $\langle \jmath_S(S,d)\rangle_\epsilon$. Moreover, we show that the (d,ϵ)-entropy of S^k [20] admits the same asymptotic expansion. If only deterministic encoding and decoding operations are allowed, the basic bounds (4), (5) generalize simply by replacing the entropy by the (d,ϵ)-entropy of S. In both the almost-lossless and the lossy case we show that the optimal code is "almost deterministic" in the sense that randomization is performed on at most one codeword of the codebook. Enforcing deterministic encoding and decoding operations ensues a penalty of at most 0.531 bits on average achievable length.

II. Almost lossless variable length compression

A. Optimal code

In the zero-error case the optimum variable-length compressor without prefix constraints $\mathsf{f}_S^\star$ is known explicitly (e.g. [4], [21])[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 [4]:

$$\ell(\mathsf{f}_S^\star(m)) = \lfloor \log_2 m \rfloor. \tag{20}$$

Note that the ordering $P_S(1) \ge P_S(2) \ge \ldots$ implies

$$\lfloor \log_2 m \rfloor \le \imath_S(m). \tag{21}$$

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

[2]The construction in [21] omits the empty string.

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 optimal code is $\mathsf{f}(m) = \mathsf{f}_S^\star(m)$, if $m = 1, \ldots, M$ and $\mathsf{f}(m) = \varnothing$, if $m > M$.[3]

Formally, for a given encoder $P_{W|S}$, the optimal decoder is always deterministic and we denote it by g. Consider $w_0 \in \{0,1\}^\star \setminus \varnothing$ and source realization m with $P_{W|S=m}(w_0) > 0$. If $\mathsf{g}(w_0) \neq m$, the average length can be decreased, without affecting the probability of error, by setting $P_{W|S=m}(w_0) = 0$ and adjusting $P_{W|S=m}(\varnothing)$ accordingly. This argument implies that the optimal encoder has at most one source realization m mapping to each $w_0 \neq \varnothing$. Next, let $m_0 = \mathsf{g}(\varnothing)$ and by a similar argument conclude that $P_{W|S=m_0}(\varnothing) = 1$. But then, interchanging m_0 and 1 leads to the same or better probability of error and shorter average length, which implies that the optimal encoder maps 1 to $\varnothing$. Continuing in the same manner for $m_0 = \mathsf{g}(0), \mathsf{g}(1), \ldots, \mathsf{g}(\mathsf{f}_S^\star(M))$, we conclude that the optimal code maps $\mathsf{f}(m) = \mathsf{f}_S^\star(m)$, $m = 1, \ldots, M$. Finally, assigning the remaining source outcomes whose total mass is ϵ to $\varnothing$ shortens the average length without affecting the error probability, so $\mathsf{f}(m) = \varnothing$, $m > M$ is optimal.

We proceed to describe an optimum construction that holds without the foregoing fortuitous choice of ϵ. Let M be the smallest integer 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[4]

$$\epsilon = \alpha \sum_{\substack{m \in \mathcal{M}: \\ \lfloor \log_2 m \rfloor = \eta}} P_S(m) + \sum_{\substack{m \in \mathcal{M}: \\ \lfloor \log_2 m \rfloor > \eta}} P_S(m) \tag{22}$$

$$= \mathbb{E}\left[\varepsilon^\star(S)\right] \tag{23}$$

[3]Jelinek [22, Sec 3.4] provided an asymptotic analysis of a scheme in which a vanishing portion of the least likely source outcomes is mapped to the same codeword, while the rest of the source outcomes are encoded losslessly.

[4]It does not matter how the encoder implements randomization on the boundary 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

$$\varepsilon^\star(m) = \begin{cases} 0 & \ell(\mathsf{f}_S^\star(m)) < \eta \\ \alpha & \ell(\mathsf{f}_S^\star(m)) = \eta \\ 1 & \ell(\mathsf{f}_S^\star(m)) > \eta \end{cases} \tag{24}$$

We have shown that the output of the optimal encoder has structure[5]

$$W(m) = \begin{cases} f_S^\star(m) & \langle \ell(\mathsf{f}_S^\star(m)) \rangle_\epsilon > 0 \\ \varnothing & \text{otherwise} \end{cases} \tag{25}$$

and that the minimum average length is given by

$$L_S^\star(\epsilon) = \mathbb{E}\left[\langle \ell(\mathsf{f}_S^\star(S)) \rangle_\epsilon\right] \tag{26}$$

$$= L_S^\star(0) - \max_{\varepsilon(\cdot):\mathbb{E}[\varepsilon(S)]\le\epsilon} \mathbb{E}\left[\varepsilon(S)\ell(\mathsf{f}_S^\star(S))\right] \tag{27}$$

$$= L_S^\star(0) - \mathbb{E}\left[\varepsilon^\star(S)\ell(\mathsf{f}_S^\star(S))\right] \tag{28}$$

where the optimization is over $\varepsilon\colon \mathbb{Z}^+ \mapsto [0,1]$, and the optimal error profile $\varepsilon^\star(\cdot)$ that achieves (27) is given by (24).

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, $L_{S,\mathsf{det}}^\star(\epsilon) = 0$. At the other extreme, if $\epsilon = 0$, then $M = |\mathcal{M}|$ and [3]

$$L_S^\star(0) = \mathbb{E}[\ell(\mathsf{f}_S^\star(S))] = \sum_{i=1}^{\infty} \mathbb{P}[S \ge 2^i] \tag{29}$$

Denote by $L_{S,\mathsf{det}}(\epsilon)$ the minimum average length comparable with error probability ϵ if randomized codes are not allowed. It satisfies the bounds

$$L_S^\star(\epsilon) \le L_{S,\mathsf{det}}(\epsilon) \tag{30}$$

$$\le L_S^\star(\epsilon) + \phi(\min\left\{\epsilon, e^{-1}\right\}), \tag{31}$$

where

$$\phi(x) \triangleq x \log_2 \frac{1}{x}. \tag{32}$$

[5]If error detection is required and $\epsilon \ge P_S(1)$, then $\mathsf{f}_S^\star(m)$ in the right side of (25) is replaced by $\mathsf{f}_S^\star(m+1)$. Similarly, if error detection is required and $P_S(j) > \epsilon \ge P_S(j+1)$, $\mathsf{f}_S^\star(m)$ in the right side of (25) is replaced by $\mathsf{f}_S^\star(m+1)$ as long as $m \ge j$, and $\varnothing$ in the right side of (25) is replaced by $\mathsf{f}_S^\star(j)$.

Note that $0 \le \phi(x) \le e^{-1}\log_2 e \approx 0.531$ bits on $x \in [0,1]$, where the maximum is achieved at $x = e^{-1}$.

To show (31), observe that the optimal encoder needs to randomize at most one element of $\mathcal{M}$. Indeed, let $m_0 \in \mathcal{M}$ be the minimum of m_0 satisfying

$$\mathbb{P}\left[S > m_0 | \lfloor \log_2 S \rfloor = \eta\right] \le \alpha \tag{33}$$

and map all $\{m > m_0 \colon \lfloor \log_2 m \rfloor = \eta\}$ to $\varnothing$, all $\{m < m_0 \colon \lfloor \log_2 m \rfloor = \eta\}$ to $\mathsf{f}_S^\star(m)$, and map m_0 to $\varnothing$ with probability $\alpha^- \triangleq (\alpha - \mathbb{P}\left[S > m_0 | \lfloor \log_2 S \rfloor = \eta\right]) \frac{\mathbb{P}[\lfloor \log_2 S \rfloor = \eta]}{P_S(m_0)}$, and to $\mathsf{f}_S^\star(m_0)$ otherwise. Clearly this construction achieves both (23) and (26). Using (21), it follows that

$$L^\star_{S,\mathtt{det}}(\epsilon) = L^\star_S(\epsilon) + \alpha^- P_S(m_0)\ell(\mathsf{f}^\star_S(m_0)) \tag{34}$$

$$\le L^\star_S(\epsilon) + \alpha^- P_S(m_0)\log_2 \frac{1}{P_S(m_0)} \tag{35}$$

To obtain (31), notice that $\alpha^- P_S(m_0) \le \epsilon$, and if $P_S(m_0) > \epsilon$ we bound

$$\alpha^- P_S(m_0)\log_2 \frac{1}{P_S(m_0)} \le \epsilon \log_2 \frac{1}{\epsilon}. \tag{36}$$

Otherwise, since the function $\phi(p)$ is monotonically increasing on $p \le e^{-1}$ and decreasing on $p > e^{-1}$, maximizing it over $[0,\epsilon]$ we obtain (31).

Variants of the variational characterization (27) will be important throughout the paper. In general, for $X \in \mathbb{R}$

$$\mathbb{E}\left[\langle X \rangle_\epsilon\right] = \min_{\varepsilon(\cdot):\mathbb{E}\left[\varepsilon(X)\right] \le \epsilon} \mathbb{E}\left[(1-\varepsilon(X))X\right] \tag{37}$$

where the optimization is over $\varepsilon \colon \mathbb{R} \mapsto [0,1]$.

B. Erokhin's function

As made evident in (10), Erokhin's function [12] plays an important role in characterizing the nonasymptotic limit of variable-length lossless data compression allowing nonzero error probability. In this subsection, we point out some of its properties.

Erokhin's function is defined in (12), but in fact, the constraint in (12) is achieved with equality:

$$\mathbb{H}(S,\epsilon) = \min_{\substack{P_{Z|S}: \\ \mathbb{P}[S \ne Z] = \epsilon}} I(S;Z) \tag{38}$$

Indeed, given $\mathbb{P}[S \neq Z] \leq \epsilon$ we may define Z' such that $S \to Z \to Z'$ and $\mathbb{P}[S \neq Z'] = \epsilon$ (for example, by probabilistically mapping non-zero values of Z to $Z' = 0$).

Furthermore, Erokhin's function can be parametrically represented as follows [12].

$$\mathbb{H}(S, \epsilon) = \sum_{m=1}^{M} P_S(m) \log_2 \frac{1}{P_S(m)} - (1-\epsilon) \log_2 \frac{1}{1-\epsilon} - (M-1)\eta \log_2 \frac{1}{\eta} \tag{39}$$

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

$$\sum_{m=1}^{M} P_S(m) = 1 - \epsilon + (M-1)\eta \tag{40}$$

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

$$\mathbb{H}(S, \epsilon) = \log_2 M - \epsilon \log_2 (M-1) - h(\epsilon)\,, \tag{41}$$

As the following result shows, Erokhin's function is bounded in terms of the expectation of the ϵ-cutoff of information, $\langle \imath_S(S) \rangle_\epsilon$, which is easier to compute and analyze than the exact parametric solution in (39).

Theorem 1 (Bounds to $\mathbb{H}(S, \epsilon)$)**.** *If $0 \leq \epsilon < 1 - P_S(1)$, Erokhin's function satisfies*

$$\mathbb{E}\left[\langle \imath_S(S) \rangle_\epsilon\right] - \epsilon \log_2 (L_S^\star(0) + \epsilon) - 2\, h(\epsilon) - \epsilon \log_2 \frac{e}{\epsilon} \leq \mathbb{H}(S, \epsilon) \tag{42}$$

$$\leq \mathbb{E}\left[\langle \imath_S(S) \rangle_\epsilon\right] \tag{43}$$

If $\epsilon > 1 - P_S(1)$, then $\mathbb{H}(S, \epsilon) = 0$.

Proof. The bound in (42) follows from (71) and (45) below. Showing (43) involves defining a suboptimal choice (in (12)) of

$$Z = \begin{cases} S & \langle \imath_S(S) \rangle_\epsilon > 0 \\ \bar{S} & \langle \imath_S(S) \rangle_\epsilon = 0 \end{cases} \tag{44}$$

where $P_{S\bar{S}} = P_S P_S$, and noting that $I(S; Z) \leq D(P_{Z|S} \| P_S | P_S) = \mathbb{E}\left[\langle \imath_S(S) \rangle_\epsilon\right]$, where $D(\cdot \| \cdot | \cdot)$ denotes conditional relative entropy.

□

Figure 1 plots the bounds to $\mathbb{H}(S^k, \epsilon)$ in Theorem 1 for biased coin flips.

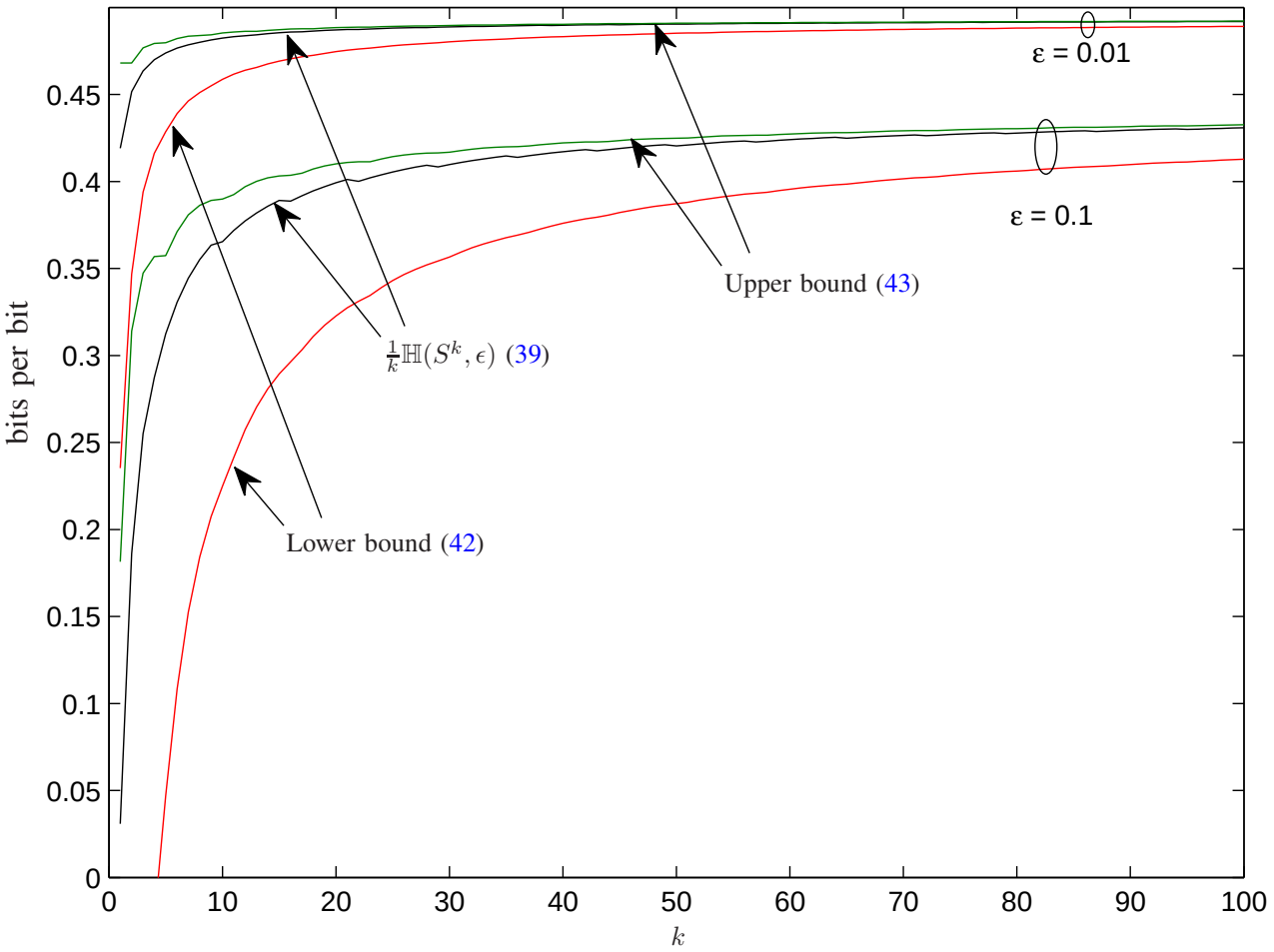

Fig. 1. Bounds to Erokhin's function for a memoryless binary source with bias $p = 0.11$.

C. Non-asymptotic bounds

Expression (26) is not always convenient to work with. The next result tightly bounds $L^\star(\epsilon)$ in terms of the ϵ-cutoff of information, $\langle \imath_S(S) \rangle_\epsilon$, a random variable which is easier to deal with.

Theorem 2 (Bounds to $L_S^\star(\epsilon)$)**.** *If $0 \leq \epsilon < 1 - P_S(1)$, then the minimum achievable average length satisfies*

$$\mathbb{E}\left[\langle \imath_S(S) \rangle_\epsilon\right] + L_S^\star(0) - H(S) \leq L_S^\star(\epsilon) \tag{45}$$

$$\leq \mathbb{E}\left[\langle \imath_S(S) \rangle_\epsilon\right] \tag{46}$$

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

Proof. Due to (37), we have the variational characterization:

$$\mathbb{E}\left[\langle \imath_S(S)\rangle_\epsilon\right] = H(S) - \max_{\varepsilon(\cdot):\mathbb{E}\left[\varepsilon(S)\right]\le\epsilon} \mathbb{E}\left[\varepsilon(S)\imath_S(S)\right] \tag{47}$$

where $\varepsilon(\cdot)$ takes values in $[0,1]$. We obtain (45)–(46) comparing (27) and (47) via (21). □

Example. If S is equiprobable on an alphabet of cardinality M, then

$$\langle \imath_S(S)\rangle_\epsilon = \begin{cases} \log_2 M \text{ w. p. } 1-\epsilon \\ 0 \text{ otherwise} \end{cases} \tag{48}$$

The next result, in which the role of entropy is taken over by Erokhin's function, generalizes the bounds in (4) and (5) to $\epsilon > 0$.

Theorem 3 (Relation between $L_S^\star(\epsilon)$ and $\mathbb{H}(S,\epsilon)$)**.** *If $0 \le \epsilon < 1 - P_S(1)$, then the minimum achievable average length satisfies*

$$\mathbb{H}(S,\epsilon) - \log_2(\mathbb{H}(S,\epsilon)+1) - \log_2 e \le L_S^\star(\epsilon) \tag{49}$$

$$\le \mathbb{H}(S,\epsilon) + \epsilon \log_2(H(S)+\epsilon) + \epsilon \log_2 \frac{e}{\epsilon} + 2\,h(\epsilon) \tag{50}$$

where $\mathbb{H}(S,\epsilon)$ is defined in (12), and $h(x) = x\log_2\frac{1}{x} + (1-x)\log_2\frac{1}{1-x}$ is the binary entropy function.

Note that we recover (4) and (5) by particularizing Theorem 3 to $\epsilon = 0$.

Proof. We first show the converse bound (49). The entropy of the output string $W \in \{0,1\}^\star$ of an arbitrary compressor $S \to W \to \hat{S}$ with $\mathbb{P}\left[S \ne \hat{S}\right] \le \epsilon$ satisfies

$$H(W) \ge I(S;W) = I(S;\hat{S}) \ge \mathbb{H}(S,\epsilon) \tag{51}$$

where the rightmost inequality holds in view of (12). Noting that the identity mapping $W \mapsto W \mapsto W$ is a lossless variable-length code, we lower-bound its average length as

$$H(W) - \log_2(H(W)+1) - \log_2 e \le L_W^\star(0) \tag{52}$$

$$\le \mathbb{E}[\ell(W)] \tag{53}$$

where (52) follows from (4). The function of $H(W)$ in the left side of (52) 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 (52) by invoking (51). This concludes the proof of (49). By applying [2, Theorem 1] to W, we can get a sharper lower bound (which is always positive)

$$\psi^{-1}(\mathbb{H}(S,\epsilon)) \leq L_S^\star(\epsilon) \tag{54}$$

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

$$\psi(x) = x + (1+x)\log_2(1+x) - x\log_2 x. \tag{55}$$

To show the achievability bound (50), fix $P_{Z|S}$ satisfying the constraint in (38). Denote for brevity

$$\Lambda \triangleq \ell(\mathsf{f}_S^\star(S)) \tag{56}$$

$$E \triangleq 1\{S \neq Z\} \tag{57}$$

$$\varepsilon(i) \stackrel{\triangle}{=} \mathbb{P}[S \neq Z | \Lambda = i] \tag{58}$$

We proceed to lower bound the mutual information between S and Z:

$$I(S;Z) = I(S;Z,\Lambda) - I(S;\Lambda|Z) \tag{59}$$

$$= H(S) - H(\Lambda|Z) - H(S|Z,\Lambda) \tag{60}$$

$$= H(S) - I(\Lambda;E|Z) - H(\Lambda|Z,E) - H(S|Z,\Lambda) \tag{61}$$

$$\geq L_S^\star(\epsilon) + H(S) - L_S^\star(0) - \epsilon\log_2(L_S^\star(0)+\epsilon) - \epsilon\log_2\frac{e}{\epsilon} - 2\,h(\epsilon) \tag{62}$$

where (62) follows from $I(\Lambda;E|Z) \leq h(\epsilon)$ and the following chains (63)-(64) and (66)-(70).

$$H(S|Z,\Lambda) \leq \mathbb{E}\left[\varepsilon(\Lambda)\Lambda + h(\varepsilon(\Lambda))\right] \tag{63}$$

$$\leq L_S^\star(0) - L_S^\star(\epsilon) + h(\epsilon) \tag{64}$$

where (63) is by Fano's inequality: conditioned on $\Lambda = i$, S can have at most 2^i values, so

$$H(S|Z,\Lambda=i) \leq \; i\,\varepsilon(i) + h(\varepsilon(i)) \tag{65}$$

and (64) follows from (27), (38) and the concavity of $h(\cdot)$.

The third term in (61) is upper bounded as follows.

$$H(\Lambda|Z,E) = \epsilon H(\Lambda|Z,E=1) \tag{66}$$

$$\leq \epsilon H(\Lambda|S \neq Z) \tag{67}$$

$$\leq \epsilon \left(\log_2(1+\mathbb{E}\left[\Lambda|S\neq Z\right]) + \log_2 e\right) \tag{68}$$

$$\leq \epsilon \left(\log_2\left(1+\frac{\mathbb{E}\left[\Lambda\right]}{\epsilon}\right) + \log_2 e\right) \tag{69}$$

$$= \epsilon \log_2 \frac{e}{\epsilon} + \epsilon(\log_2(L_S^\star(0)) + \epsilon)\,, \tag{70}$$

where (66) follows since $H(\Lambda|Z,E=0)=0$, (67) is because conditioning decreases entropy, (68) follows by maximizing entropy under the mean constraint (achieved by the geometric distribution), (69) follows by upper-bounding

$$\mathbb{P}[S\neq Z]\,\mathbb{E}\left[\Lambda|S\neq Z\right] \leq \mathbb{E}\left[\Lambda\right]$$

and (70) applies (29).

Finally, since the right side of (62) does not depend on Z, we may minimize the left side over $P_{Z|S}$ satisfying the constraint in (38) to obtain

$$L_S^\star(\epsilon) \leq \mathbb{H}(S,\epsilon) + L_S^\star(0) - H(S) + \epsilon\log_2(L_S^\star(0)+\epsilon) + 2\,h(\epsilon) + \epsilon\log_2\frac{e}{\epsilon} \tag{71}$$

which leads to (50) via Wyner's bound (5).

□

Remark 1. The following stronger version of (4) is shown in [4, Lemma 3]:

$$H(S) \leq L_S^\star(0) + \log_2(L_S^\star(0)+1) + \log_2 e \tag{72}$$

which, via the same reasoning as in (51)–(53), leads to the following strengthening of (49):

$$\mathbb{H}(S,\epsilon) \leq L_S^\star(\epsilon) + \log_2(L_S^\star(\epsilon)+1) + \log_2 e \tag{73}$$

Together, Theorems 1, 2, and 3 imply that as long as the quantities $L_S^\star(\epsilon)$, $\mathbb{H}(S,\epsilon)$ and $\mathbb{E}\left[\langle \imath_S(S)\rangle_\epsilon\right]$ are not too small, they are close to each other.

In principle, it may seem surprising that $L_S^\star(\epsilon)$ is connected to $\mathbb{H}(S,\epsilon)$ in the way dictated by Theorem 3, 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 (25) and the conditional distribution achieving (12) 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 (12) do not favor $m = 1$ or any particular outcome of S.

D. Asymptotics for memoryless sources

Theorem 4. *Assume that:*

- $P_{S^k} = P_{\mathsf{S}} \times \ldots \times P_{\mathsf{S}}$.
- *The third absolute moment of* $\imath_{\mathsf{S}}(\mathsf{S})$ *is finite.*

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

$$\left.\begin{array}{l} L^{\star}_{S^k}(\epsilon) \\ \mathbb{H}(S^k, \epsilon) \\ \mathbb{E}\left[\left\langle \imath_{S^k}(S^k)\right\rangle_{\epsilon}\right] \end{array}\right\} = (1-\epsilon)kH(\mathsf{S}) - \sqrt{\frac{kV(\mathsf{S})}{2\pi}} e^{-\frac{(Q^{-1}(\epsilon))^2}{2}} + \theta(k) \tag{74}$$

where the remainder term satisfies

$$-\log_2 k + O\left(\log_2 \log_2 k\right) \leq \theta(k) \leq O\left(1\right) \tag{75}$$

Proof. If the source is memoryless, the information in S^k is a sum of i.i.d. random variables as indicated in (16), and Theorem 4 follows by applying Lemma 1 below to the bounds in Theorem 2. □

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*

$$\mathbb{E}\left[\left\langle \sum_{i=1}^{k} X_i \right\rangle_{\epsilon}\right] = (1-\epsilon)k\mathbb{E}\left[\mathsf{X}\right] - \sqrt{\frac{k\mathrm{Var}\left[\mathsf{X}\right]}{2\pi}} e^{-\frac{(Q^{-1}(\epsilon))^2}{2}} + O\left(1\right) \tag{76}$$

Proof. Appendix A. □

Remark 2. Applying (6) to (45), for finite alphabet sources the lower bound on $L^{\star}_{S^k}(\epsilon)$ is improved to

$$\theta(k) \geq -\frac{1}{2}\log_2 k + O\left(1\right) \tag{77}$$

For $\mathbb{H}(S^k, \epsilon)$, the lower bound is in fact $\theta(k) \geq -\epsilon \log_2 k + O\left(1\right)$, while for $\mathbb{E}\left[\left\langle \imath_{S^k}(S^k)\right\rangle_{\epsilon}\right]$, $\theta(k) = O\left(1\right)$.

Remark 3. If the source alphabet is finite, we can sketch an alternative proof of Theorem 4 using the method of types. By concavity and symmetry, it is easy to see that the optimal coupling that achieves $\mathbb{H}(S^k, \epsilon)$ satisfies the following property: the error profile

$$\epsilon(s^k) \stackrel{\triangle}{=} \mathbb{P}[Z^k \neq S^k | S^k = s^k] \tag{78}$$

is constant on each k-type (see [23, Chapter 2] for types). Denote the type of s^k as $\hat{P}_{s^k}$ and its size as $M(s^k)$. We then have the following chain:

$$\begin{aligned} I(S^k; Z^k) &= I(S^k, \hat{P}_{S^k}; Z^k) && (79)\\ &= I(S^k; Z^k | \hat{P}_{S^k}) + O(\log k) && (80)\\ &\geq \mathbb{E}\left[(1-\epsilon(S^k))\log M(S^k)\right] + O(\log k) && (81) \end{aligned}$$

where (80) follows since there are only polynomially many types and (81) follows from (41). Next, (81) is to be minimized over all $\epsilon(S^k)$ satisfying $\mathbb{E}\left[\epsilon(S^k)\right] \leq \epsilon$. The solution (of this linear optimization) is easy: $\epsilon(s^k)$ is 1 for all types with $M(s^k)$ exceeding a certain threshold, and 0 otherwise. In other words, we get

$$\mathbb{H}(S^k, \epsilon) = (1-\epsilon)\mathbb{E}\left[\log M(S^k) | M(S^k) \leq \gamma\right] + O(\log k)\,, \tag{82}$$

where γ is chosen so that $\mathbb{P}[M(S^k) > \gamma] = \epsilon$. Using the relation between type size and its entropy, we have

$$M(s^k) = H(\hat{P}_{s^k}) + O(\log k) \tag{83}$$

and from the central-limit theorem, cf. [13], [24], we get

$$H(\hat{P}_{S^k}) \stackrel{d}{=} kH(\mathsf{S}) + \sqrt{\frac{V(\mathsf{S})}{k}}U + O(\log k) \qquad U \sim \mathcal{N}(0,1)\,. \tag{84}$$

Thus, putting together (82), (83), (84) and after some algebra (74) follows.

E. Discussion

Theorem 4 exhibits an unusual phenomenon in which the dispersion term improves the achievable average rate. As illustrated in Fig. 2, a nonzero error probability ϵ decreases the average achievable rate as the source outcomes falling into the shaded area are assigned length 0. The total reduction in average length is composed of the reduction in asymptotically achievable average length due to nonzero ϵ and the reduction due to finite blocklength. The asymptotic average length is reduced because the center of probabilistic mass Fig. 2 shifts to the left when the ϵ-tail of the distribution is chopped off. Moreover, for a fixed ϵ the wider the distribution the bigger is this shift, thus shorter blocklengths and larger dispersions help to achieve a lower average rate.

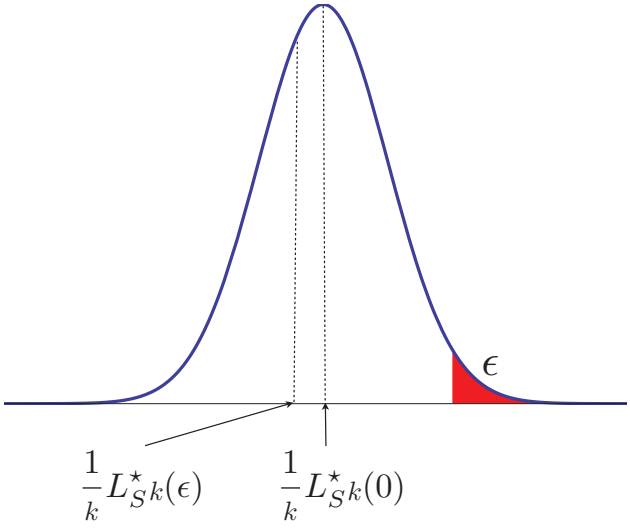

Fig. 2. The benefit of nonzero ϵ and dispersion. The bell-shaped curve depicts an idealized form of the pmf of $\frac{1}{k}\ell\left(f^\star(S^k)\right)$.

For a source of biased coin flips, Fig. 4 depicts the exact average rate of the optimal code as well as the approximation in (74). Both curves are monotonically increasing in k.

The dispersion term in (74) vanishes quickly with ϵ. More precisely, as $\epsilon \to 0$, we have (Appendix B)

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{\left(Q^{-1}(\epsilon)\right)^2}{2}} = \epsilon\sqrt{2\log_2\frac{1}{\epsilon}} + o\left(\epsilon\right) \tag{85}$$

Therefore, a refined analysis of higher order terms in the expansion (74) is desirable in order to obtain an approximation that is accurate even at short blocklengths. Inspired by [25], in Fig. 4

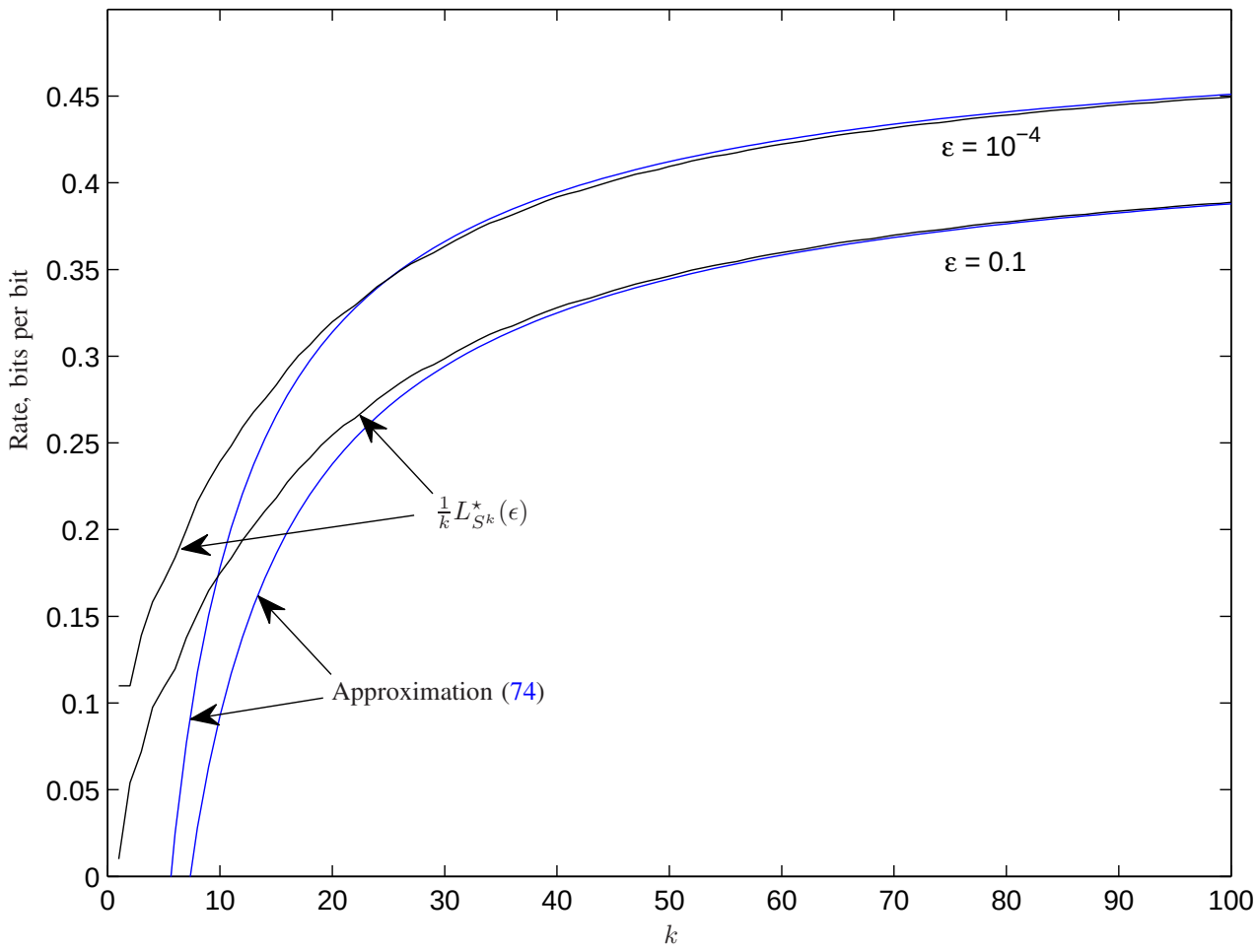

Fig. 3. Average rate achievable for variable-rate almost lossless encoding of a memoryless binary source with bias $p = 0.11$ and two values of ϵ. For $\epsilon < 10^{-4}$, the resulting curves are almost indistinguishable from the $\epsilon = 10^{-4}$ curve.

we adopt the following value for the remainder in (74):

$$\theta(k) = (1-\epsilon)\left(\frac{\log_2 k}{2} - \frac{1}{2}\log_2(4e^3\pi) + \frac{p}{1-2p} + \log_2\frac{1}{1-2p} + \frac{1}{2(1-2p)}\log_2\frac{1-p}{p}\right) \tag{86}$$

where p is the coin bias, which proves to yield a remarkably good approximation, accurate for blocklengths as short as 20.

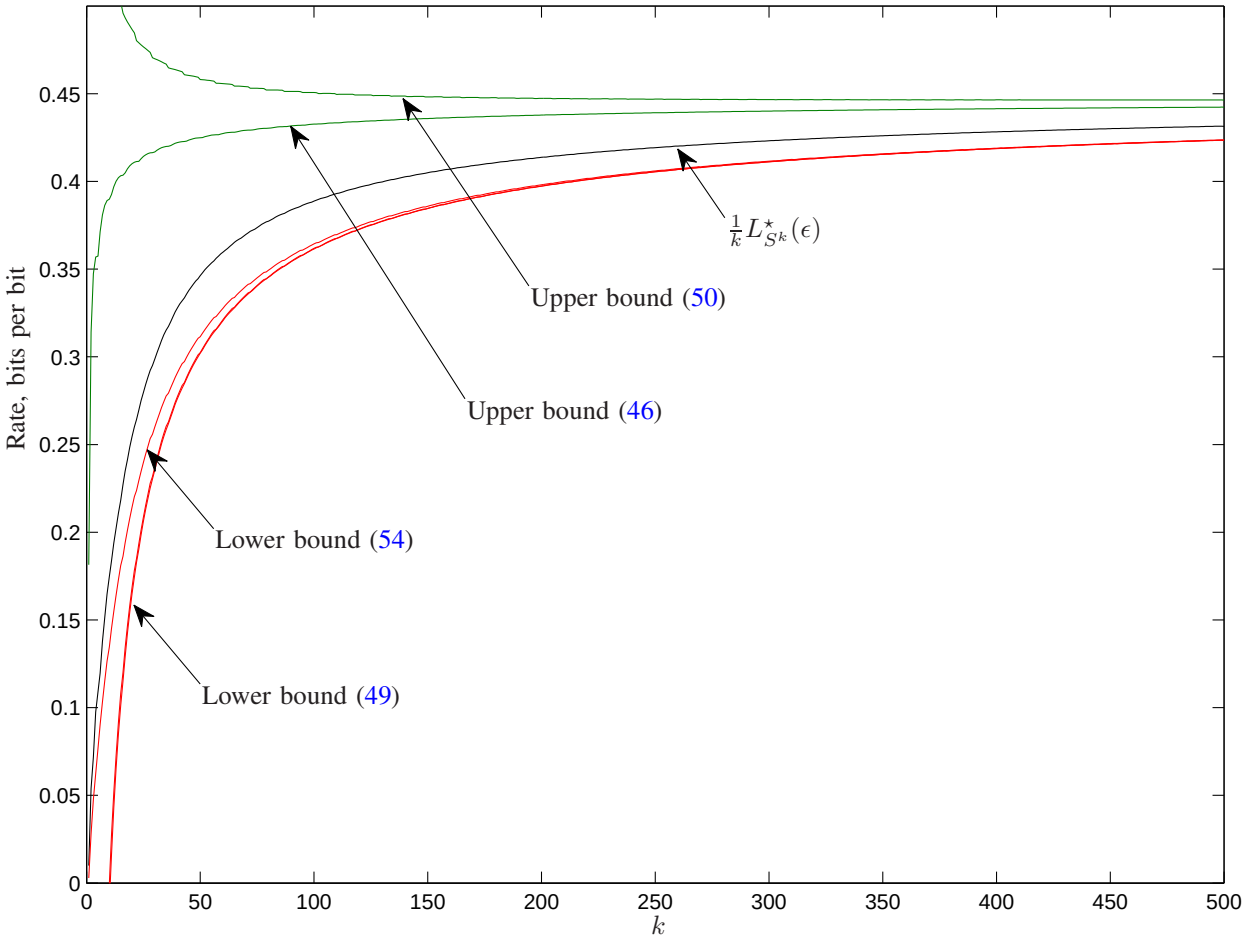

Fig. 4. Bounds to the average rate achievable for variable-rate almost lossless encoding of a memoryless binary source with bias $p = 0.11$ and $\epsilon = 0.1$. The lower bound in (49) is virtually indistinguishable from a weakening of (45) using (4).

III. Lossy variable-length compression

A. The setup

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* $P_{Z|W}\colon \{0,1\}^\star \mapsto \widehat{\mathcal{M}}$ *such that*

$$\mathbb{P}\left[\mathsf{d}\left(S,Z\right) > d\right] \le \epsilon \tag{87}$$

$$\mathbb{E}\left[\ell(W)\right] \le L \tag{88}$$

The goal of this section is to characterize the minimum achievable average length compatible with the given tolerable error ϵ:

$$L_S^\star(d,\epsilon) \triangleq \{\min\ \ L\colon \exists \text{ an } (L,d,\epsilon) \text{ code}\} \tag{89}$$

Section III-B discusses the properties of the optimal code. Section III-C reviews some background facts from rate-distortion theory. Section III-D presents single-shot results, and Section III-E focuses on the asymptotics.

B. Optimal code

Unlike the lossless setup in Section II, the optimal encoding and decoding mappings do not admit, in general, explicit descriptions. We can however point out several properties of the optimal code.

We first focus on the case $\epsilon = 0$. The optimal $(d,0)$ code satisfies the following properties.

1) The optimal encoder $\mathsf{f}^\star$ and decoder $\mathsf{g}^\star$ are deterministic mappings.
2) The output $W^\star = \mathsf{f}^\star(S)$ of the optimal encoder satisfies $P_{W^\star}(\varnothing) \ge P_{W^\star}(0) \ge P_{W^\star}(1) \ge P_{W^\star}(00) \ge \ldots$
3) For each $w \in \{0,1\}^\star$

$$\mathsf{f}^{\star -1}(w) = B_{\mathsf{g}^\star(w)} \setminus \cup_{v \prec w} B_{\mathsf{g}^\star(v)} \tag{90}$$

where $\prec$ is lexicographic ordering, and

$$B_z \triangleq \{s\colon \mathsf{d}(s,z) \le d\} \tag{91}$$

Let $z_1, z_2, \ldots$ be a d-covering of $\mathcal{M}$. First, we will show that the foregoing claims hold for decoders whose image is constrained to the given d-covering $z_1, z_2, \ldots$. Then, we will conclude that since the claims hold for all d-coverings, they hold for the one that results in the minimum average length as well.

To show 1), let $(P_{W|S}, P_{Z|W})$ be a $(d,0)$ code. The optimal encoder is deterministic because if there exist $s \in \mathcal{M}$ and $w \prec v \in \{0,1\}^\star$ such that $P_{W|S=s}(w) > 0$ and $P_{W|S=s}(v) > 0$ we may decrease the average length by setting $P_{W|S=s}(w) = 1$. The optimal decoder is deterministic because if for some $w \in \{0,1\}^\star$ there exist $z', z'' \in \{z_1, z_2, \ldots\}$ such that $P_{Z|W=w}(z') > 0$ and $P_{Z|W=w}(z'') > 0$, then nothing changes by setting $P_{Z|W=w}(z') = 1$.

To show 2), observe that if there exist $w \prec v \in \{0,1\}^\star$ such that $P_W(w) < P_W(v)$, then the average length is shortened by swapping w and v.

To show 3), notice that the average length decreases as $P_W(\varnothing)$ increases, and the latter is maximized by setting $\mathsf{f}^{-1}(\varnothing) = B_{\mathsf{g}(\varnothing)}$. Further, $P_W(0)$ is maximized without affecting $P_W(\varnothing)$ by setting $\mathsf{f}^{-1}(0) = B_{\mathsf{g}(0)} \backslash B_{\mathsf{g}(\varnothing)}$ and so forth.

We now consider the case $\epsilon > 0$. The optimal (d,ϵ) code satisfies the following properties.

1) The optimal decoder $\mathsf{g}^\star$ is deterministic, and the optimal encoder $P_{W^\star|S}$ satisfies $P_{W^\star|S=s}(w) = 1 - P_{W^\star|S=s}(\varnothing)$ for all $s \in \mathcal{M}$ and all $w \in \{0,1\}^\star \backslash \varnothing$.
2) The output of the optimal encoder satisfies $P_{W^\star}(\varnothing) \geq P_{W^\star}(0) \geq P_{W^\star}(1) \geq P_{W^\star}(00) \geq \ldots$
3) There exist $\eta \in \mathbb{R}^+$ such that $\mathbb{P}\left[\ell(W^\star) > \eta\right] = 0$ and $0 \leq \alpha < 1$ such that for each $w \in \{0,1\}^\star \backslash \varnothing$

$$P_{W^\star|S=s}(w) = \begin{cases} 1, \ s \in B_{\mathsf{g}^\star(w)} \backslash \cup_{v \prec w} B_{\mathsf{g}^\star(v)} \ \& \ \ell(w) < \eta \\ 1-\alpha, \ s \in B_{\mathsf{g}^\star(w)} \backslash \cup_{v \prec w} B_{\mathsf{g}^\star(v)} \ \& \ \ell(w) = \eta \end{cases} \tag{92}$$

and

$$P_{W^\star|S=s}(\varnothing) = \begin{cases} 1, \ s \notin \cup_w B_{\mathsf{g}^\star(w)} \\ \alpha, \ s \in \cup_w B_{\mathsf{g}^\star(w)} \ \& \ \ell(w) = \eta \end{cases} \tag{93}$$

Property 3) implies in particular that $\ell(\mathsf{f}^\star(s)) = 0$ as long as $\mathsf{d}(s, \mathsf{g}^\star(\mathsf{f}^\star(s))) > d$.

We say that $\mathcal{F} \subseteq \widehat{\mathcal{M}}$ is a (d,ϵ)-covering of $\mathcal{M}$ if $\mathbb{P}\left[\min_{z \in \mathcal{F}} \mathsf{d}(S,z) > d\right] \leq \epsilon$. Note that a finite (d,ϵ)-covering always exists as long as a d-covering exists [20]: indeed, given a d-covering $z_1, z_2, \ldots$, let M satisfy $\sum_{m>M} \mathbb{P}\left[S \in B_{z_m} \backslash \cup_{i<m} B_{z_i}\right] \leq \epsilon$ and just drop all $z_m \colon m > \eta$ to obtain a finite (d,ϵ)-covering. Let $z_1, z_2, \ldots, z_M$ be a (d,ϵ)-covering of $\mathcal{M}$. Observing that an infinite (d,ϵ)-covering $z_1, z_2, \ldots$ can only result in a longer average length, we will first show that the foregoing claims hold for decoders whose image is constrained to a given (d,ϵ)-covering

$z_1, z_2, \ldots, z_M$. Then, we will conclude that since the claims hold for all finite (d, ϵ)-coverings, they hold for the one that results in the minimum average length as well.

To show 1), notice that for a given encoder $P_{W|S}$, the optimal decoder is always deterministic. Indeed, if for some $w \in \{0,1\}^\star$ there exist $z', z'' \in \{z_1, z_2, \ldots, z_M\}$ such that $P_{Z|W=w}(z') > 0$, $P_{Z|W=w}(z'') > 0$ and $P_{S|W=w}(B_{z'}) \geq P_{S|W=w}(B_{z''})$ then the excess distortion can only be reduced by setting $P_{Z|W=w}(z') = 1$, without affecting the average length. Denote that deterministic decoder by g. As for the encoder, consider $w \in \{0,1\}^\star \setminus \varnothing$ and source realization s with $P_{W|S=s}(w) > 0$. If $\mathsf{d}(s, \mathsf{g}(w)) > d$, the average length can be decreased, without increasing the excess distortion probability, by setting $P_{W|S=s}(w) = 0$ and adjusting $P_{W|S=s}(\varnothing) = 1$ accordingly. This argument implies that the optimal encoder satisfies $P_{S|W=w}(B_{\mathsf{g}(w)}) = 1$ for each $w \neq \varnothing$. Now, if there exist s and $w \prec v \in \{0,1\}^\star \backslash \varnothing$ such that $P_{W|S=s}(w) > 0$ and $P_{W|S=s}(v) > 0$, we may decrease the average length with no impact on the probability of excess distortion by setting $P_{W|S=s}(w) = 1$.

To show 2), notice that if there exist $w \prec v \in \{0,1\}^\star \backslash \varnothing$ such that $P_W(w) < P_W(v)$, then the average length is shortened by swapping w and v. If there exist $w \in \{0,1\}^\star \backslash \varnothing$ with $P_W(w) > P_W(\varnothing)$ then the average length is shortened by swapping w and $\varnothing$ and setting $P_{W|S=s}(w) = 0$ while adjusting $P_{W|S=s}(\varnothing) = 1$ accordingly for each $s \notin B_{g(w)}$.

To show 3), we argue as in the case $\epsilon = 0$ that setting

$$P_{W|S=s}(w) = 1, \;\; s \in B_{\mathsf{g}(w)} \backslash \cup_{v \prec w} B_{\mathsf{g}(v)} \tag{94}$$

$$P_{W|S=s}(\varnothing) = 1, \;\; s \notin \cup_w B_{\mathsf{g}(w)} \tag{95}$$

yields the minimum average length among all (d, ϵ') codes with codebook $z_1, z_2, \ldots$ satisfying 1) and 2) where $\epsilon' \triangleq \mathbb{P}\left[\min_m \mathsf{d}(S, z_m) > d\right]$. If $\epsilon' = \epsilon$, there is nothing else to prove. If $\epsilon' < \epsilon$, let $\eta \in \mathbb{R}^+$ and $0 < \alpha < 1$ solve

$$\mathbb{P}\left[\ell(W) > \eta\right] + \alpha \mathbb{P}\left[\ell(W) = \eta\right] = \epsilon - \epsilon' \tag{96}$$

and observe that dropping all $w \colon \ell(w) > \eta$ reduces the average length while keeping the excess distortion probability below ϵ. Now, letting $P_{W|S=s}(w) = 1 - \alpha$ for each $s \in B_{\mathsf{g}(w)} \backslash \cup_{v \prec w} B_{\mathsf{g}(v)}$ and each $w \colon \ell(w) = \eta$ and adjusting $P_{W|S=s}(\varnothing)$ accordingly further reduces the average length while making the excess distortion probability exactly ϵ.

Property 3) implies that randomization is not essential as almost the same average length can be achieved with deterministic encoding and decoding operations. Precisely, denoting by $L^{\star}_{S,\mathtt{det}}(d,\epsilon)$ the minimum average length achievable with deterministic codes, we have

$$L^{\star}_{S}(d,\epsilon) \leq L^{\star}_{S,\mathtt{det}}(d,\epsilon) \tag{97}$$

$$\leq L^{\star}_{S}(d,\epsilon) + \phi(\min\{\epsilon, e^{-1}\}) \tag{98}$$

where (98) is obtained in the same way as (31), and $0 \leq \phi(\cdot) \leq 0.531$ is defined in (32).

C. A bit of rate-distortion theory

The minimal mutual information function

$$\mathbb{R}_S(d) \triangleq \inf_{\substack{P_{Z|S}: \\ \mathbb{E}[\mathsf{d}(S,Z)]\leq d}} I(S;Z) \tag{99}$$

characterizes the minimum asymptotically achievable rate in both fixed-length compression under the average or excess distortion constraint and variable-length lossy compression under the almost sure distortion constraint [26], [27].

We assume throughout that the following basic assumptions are met.

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

$$d_{\min} \triangleq \inf\{d\colon\ \mathbb{R}_S(d) < \infty\} \tag{100}$$

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

$$\mathbb{E}\left[\min_{z\in E} \mathsf{d}(S,z)\right] < \infty \tag{101}$$

The following characterization of $\mathbb{R}_S(d)$ due to Csiszár [28] will be instrumental.

Theorem 5 (Characterization of $\mathbb{R}_S(d)$ [28, Theorem 2.3])**.** *For each $d > d_{\min}$ it holds that*

$$\mathbb{R}_S(d) = \max_{J(s),\ \lambda} \{\mathbb{E}\left[J(S)\right] - \lambda d\} \tag{102}$$

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

$$\mathbb{E}\left[\exp\left\{J(S) - \lambda \mathsf{d}(S,z)\right\}\right] \leq 1 \ \forall z \in \widehat{\mathcal{M}} \tag{103}$$

Let $(J_S(s), \lambda_S)$ attain the maximum in the right side of (102). If there exists a transition probability kernel $P_{Z^\star|S}$ that actually achieves the infimum in the right side of (99), then [28]

$$J_S(s) = \imath_{S;Z^\star}(s;z) + \lambda_S \mathsf{d}(s,z) \tag{104}$$

$$= -\log_2 \mathbb{E}\left[\exp\left(-\lambda_S \mathsf{d}(s,Z^\star)\right)\right] \tag{105}$$

where (104) holds for $P_{Z^\star}$-a.e. z, the expectation in (105) is with respect to the unconditional distribution of $Z^\star$, and the usual information density is denoted by

$$\imath_{S;Z}(s;z) \triangleq \log_2 \frac{dP_{Z|S=s}}{dP_Z}(z) \tag{106}$$

Note from (105) that by the concavity of logarithm

$$0 \le J_S(s) \le \mathbb{E}\left[\mathsf{d}(s,Z^\star)\right] \tag{107}$$

The random variable that plays the key role in characterizing the nonasymptotic fundamental limit of lossy data compression is the d-tilted information in $s \in \mathcal{M}$ [15]:

$$\jmath_S(s,d) \triangleq J_S(s) - \lambda_S d \tag{108}$$

It follows from (102) that

$$\mathbb{R}_S(d) = \mathbb{E}\left[\jmath_S(S,d)\right] \tag{109}$$

Much like information in $s \in \mathcal{M}$ which quantifies the number of bits necessary to represent s losslessly, d-tilted information in s quantifies the number of bits necessary to represent s within distortion d, in a sense that goes beyond average as in (109) [15], [17]. Particularizing (103), we observe that the d-tilted information satisfies

$$\mathbb{E}\left[\exp(\jmath_S(S,d) + \lambda_S d - \lambda_S \mathsf{d}(S,z))\right] \le 1 \tag{110}$$

Using Markov's inequality and (105), it is easy to see that the d-tilted information is linked to the probability that $Z^\star$ falls within distortion d from $s \in \mathcal{M}$:

$$\jmath_S(s,d) \le \log_2 \frac{1}{P_{Z^\star}(B_d(s))} \tag{111}$$

where

$$B_d(s) \triangleq \left\{z \in \widehat{\mathcal{M}} \colon \mathsf{d}(s,z) \le d\right\} \tag{112}$$

Moreover, under regularity conditions the reverse inequality in (111) can be closely approached [17, Proposition 3].

D. Nonasymptotic bounds

We begin with a simple generalization of basic bounds (4) and (5) to an arbitrary distortion measure and nonzero ϵ, in which the role of entropy is assumed by the (ϵ, δ)-entropy of the source S, defined as [20]:

$$H_{\epsilon,\delta}(S) \triangleq \min_{\substack{\mathsf{f}\colon \mathcal{M}\mapsto\widehat{\mathcal{M}}:\\ \mathbb{P}[\mathsf{d}(S,\mathsf{f}(S))>\epsilon]\le\delta}} H(\mathsf{f}(S)). \tag{113}$$

Theorem 6 (Bounds to $L^{\star}_{S,\mathtt{det}}(d,\epsilon)$)**.** *The minimal average length achievable with deterministic codes under an excess-distortion constraint satisfies*

$$H_{d,\epsilon}(S) - \log_2(H_{d,\epsilon}(S)+1) - \log_2 e \le L^{\star}_{S,det}(d,\epsilon) \tag{114}$$

$$\le H_{d,\epsilon}(S) \tag{115}$$

Proof. The converse bound in (114) follows by applying (4) and minimizing over all possible output entropies. The achievability bound in (115) is implied by Wyner's bound (5) recalling (Section III-B) that the codewords of the optimal code are ordered in decreasing probabilities. □

Note that $L^{\star}(d,\epsilon)$ is also bounded in terms of $H_{d,\epsilon}(S)$, in view of Theorem 6 and (98).

Particularizing Theorem 6 to $\epsilon = 0$ and using $L^{\star}_{S}(d,0) = L^{\star}_{S,\mathtt{det}}(d,0)$ (as shown in Section III-B), we see that the minimum average length of d-semifaithful codes is bounded by

$$H_d(S) - \log_2(H_d(S)+1) - \log_2 e \le L^{\star}_{S}(d,0) \tag{116}$$

$$\le H_d(S)\,, \tag{117}$$

where $H_\epsilon(S)$ is the ϵ-entropy of the source S [20]:

$$H_{\epsilon}(S) \triangleq \min_{\substack{\mathsf{f}\colon \mathcal{M}\mapsto\widehat{\mathcal{M}}:\\ \mathsf{d}(S,\mathsf{f}(S))\le\epsilon \text{ a.s.}}} H(\mathsf{f}(S)), \tag{118}$$

which is bounded as follows:

$$\mathbb{R}_S(d,0) \le H_d(S) \tag{119}$$

$$\le \mathbb{R}_S(d,0) + \log_2\left(\mathbb{R}_S(d,0)+1\right) + C, \tag{120}$$

where C is a universal constant, and (120) holds whenever d is a metric by [29, Theorem 2].

Theorem 6 applies to the almost-lossless setting of Section II, in which case the (ϵ, δ)-entropy particularizes to $\epsilon = 0$ and Hamming distortion as

$$H_{0,\delta}(S) = \min_{\substack{\mathsf{f}\colon \mathcal{M}\mapsto\widehat{\mathcal{M}}\colon \\ \mathbb{P}[S\neq\mathsf{f}(S)]\leq\delta}} H(\mathsf{f}(S)). \tag{121}$$

The (ϵ, δ)-entropy is difficult to compute and analyze directly. We proceed to give bounds on $L_S^\star(d,\epsilon)$ and $H_{d,\epsilon}(S)$ that will essentially show that all the functions $L_S^\star(d,\epsilon)$, $H_{d,\epsilon}(S)$, $\mathbb{R}_S(d,\epsilon)$ (defined in (19)), are within $O\left(\log_2 \mathbb{R}_S(d)\right)$ bits from the easy-to-analyze function $\mathbb{E}\left[\langle \jmath_S(S,d)\rangle_\epsilon\right]$. We will show that the same is true for the function

$$\mathbb{R}_S^+(d,\epsilon) \triangleq \inf_{P_Z} \mathbb{E}\left[\langle -\log_2 P_Z(B_d(S))\rangle_\epsilon\right], \tag{122}$$

where $B_d(s)$ is the distortion d-ball around s (formally defined in (112)) and the infimum is over all distributions on $\widehat{\mathcal{M}}$,

The next result provides nonasymptotic bounds to the minimum achievable average length when randomized encoding and decoding operations are allowed.

Theorem 7 (Bounds to $L_S^\star(d,\epsilon)$)**.** *The minimal average length achievable under an excess-distortion constraint satisfies*

$$\mathbb{R}_S(d,\epsilon) - \log_2\left(\mathbb{R}_S(d,\epsilon)+1\right) - \log_2 e \leq L_S^\star(d,\epsilon) \tag{123}$$

$$\leq \mathbb{R}_S^+(d,\epsilon) \tag{124}$$

where $\mathbb{R}_S(d,\epsilon)$ is the minimal information quantity defined in (19)*, and $\mathbb{R}_S^+(d,\epsilon)$ is defined in* (122)*.*

Proof. The converse bound in (123) is shown in the same way as (114). To show the achievability bound in (124), consider the (d,ϵ) code that, given an infinite list of codewords $z_1, z_2, \ldots$, outputs the first d-close match to s as long as s is not too atypical. Specifically, the encoder outputs the lexicographic binary encoding (including the empty string) of

$$W \triangleq \begin{cases} \min\left\{m\colon \mathsf{d}(S,z_m)\leq d\right\} & \left\langle -\log_2 P_Z(B_d(S))\right\rangle_\epsilon > 0 \\ 1 & \text{otherwise} \end{cases} \tag{125}$$

The encoded length averaged over both the source and all codebooks with codewords $Z_1, Z_2, \ldots$ drawn i.i.d. from P_Z is upper bounded by

$$\mathbb{E}\left[\lfloor \log_2 W \rfloor\right] \le \mathbb{E}\left[\log_2 W \; 1\left\{\langle -\log_2 P_Z(B_d(S))\rangle_\epsilon > 0\right\}\right] \tag{126}$$

$$= \mathbb{E}\left[1\left\{\langle -\log_2 P_Z(B_d(S))\rangle_\epsilon > 0\right\} \mathbb{E}\left[\log_2 W | S\right]\right] \tag{127}$$

$$\le \mathbb{E}\left[1\left\{\langle -\log_2 P_Z(B_d(S))\rangle_\epsilon > 0\right\} \log_2 \mathbb{E}\left[W|S\right]\right] \tag{128}$$

$$= \mathbb{E}\left[\langle -\log_2 P_Z(B_d(S))\rangle_\epsilon\right] \tag{129}$$

where

- (128) is by Jensen's inequality;
- (129) holds because conditioned on $S = s$ and averaged over codebooks, W has geometric distribution with success probability $P_Z(B_d(s))$.

It follows that there is at least one codebook that yields the encoded length not exceeding the expectation in (129).

□

Remark 4. Both (114) and (123) can be strengthened as in Remark 1.

Theorem 8 (Bounds to $\mathbb{R}_S(d,\epsilon)$ and to $H_{d,\epsilon}(S)$)**.** *For all $d > d_{\min}$ we have*

$$\mathbb{E}\left[\langle \jmath_S(S,d)\rangle_\epsilon\right] - \log_2\left(\mathbb{R}_S(d) - \mathbb{R}'_S(d)d + 1\right) - \log_2 e - h(\epsilon) \le \mathbb{R}_S(d,\epsilon) \tag{130}$$

$$\le \mathbb{R}^+_S(d,\epsilon) \tag{131}$$

and for all $d \ge d_{\min}$ we have

$$\mathbb{R}^+_S(d,\epsilon) - \phi(\max\left\{1-\epsilon, e^{-1}\right\}) \le H_{d,\epsilon}(S) \tag{132}$$

$$\le \mathbb{R}^+_S(d,\epsilon) + \log_2\left(\mathbb{R}^+_S(d,\epsilon) + 1 + \phi\left(\min\left\{\epsilon, e^{-1}\right\}\right)\right)$$

$$+ 1 + \phi\left(\min\left\{\epsilon, e^{-1}\right\}\right) \tag{133}$$

where $0 \le \phi(\cdot) \le e^{-1}\log_2 e$ is defined in (32)*.*

Proof. Appendix C. □

Trivially, $\mathbb{R}_S(d,\epsilon) \le H_{d,\epsilon}(S)$.

Remark 5. In the almost-lossless setting (Hamming distortion and $d = 0$), the following bounds hold (Appendix D).

$$\mathbb{E}\left[\langle \imath_S(S)\rangle_\epsilon\right] - \phi\left(\max\left\{1-\epsilon, e^{-1}\right\}\right) \le H_{0,\epsilon}(S) \tag{134}$$

$$\le \mathbb{E}\left[\langle \imath_S(S)\rangle_\epsilon\right] + \phi\left(\min\left\{\epsilon, e^{-1}\right\}\right) \tag{135}$$

Remark 6. Particularizing (132) to the case $\epsilon = 0$, we recover the lower bound on ϵ-entropy in [20, Lemma 9]:

$$\inf_{P_Z} \mathbb{E}\left[-\log_2 P_Z(B_d(S))\right] \le H_d(S) \tag{136}$$

Remark 7. As follows from Lemma 3 in Appendix C, in the special case where

$$\jmath_S(S,d) = \mathbb{R}_S(d) \text{ a.s.} \tag{137}$$

which in particular includes the equiprobable source under a permutation distortion measure (e.g. symbol error rate) [30], the lower bound in (130) can be tightened as

$$\mathbb{R}_S(d,\epsilon) \ge (1-\epsilon)\mathbb{R}_S(d) - h(\epsilon) \tag{138}$$

Remark 8. Applying (37) to the random variable $\jmath_S(S,d)$, we have the variational characterization:

$$\mathbb{E}\left[\langle \jmath_S(S,d)\rangle_\epsilon\right] = \mathbb{R}_S(d) - \max_{\substack{\varepsilon:\ \mathcal{M}\mapsto[0,1]\\ \mathbb{E}[\varepsilon(S)]\le\epsilon}} \mathbb{E}\left[\varepsilon(S)\jmath_S(S,d)\right] \tag{139}$$

from where it follows, via (111), that

$$\mathbb{E}\left[\langle \jmath_S(S,d)\rangle_\epsilon\right] \le \mathbb{E}\left[\langle -\log_2 P_{Z^\star}(B_d(S))\rangle_\epsilon\right] \tag{140}$$

$$\le \mathbb{E}\left[\langle \jmath_S(S,d)\rangle_\epsilon\right] + \mathbb{E}\left[-\log_2 P_{Z^\star}(B_d(S))\right] - \mathbb{R}_S(d) \tag{141}$$

where $P_{Z^\star}$ is the output distribution that achieves $\mathbb{R}_S(d)$.

E. Asymptotic analysis

In this section we assume that the following conditions are satisfied.

(i) The source $\{S_i\}$ is stationary and memoryless, $P_{S^k} = P_{\mathsf{S}} \times \ldots \times P_{\mathsf{S}}$.

(ii) The distortion measure is separable, $\mathsf{d}(s^k, z^k) = \frac{1}{k}\sum_{i=1}^k \mathsf{d}(s_i, z_i)$.

(iii) The distortion level satisfies $d_{\min} < d < d_{\max}$, where $d_{\min}$ is defined in (100), 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 $P_{\mathsf{S}} \times P_{\mathsf{Z}^\star}$, and $\mathsf{Z}^\star$ achieves the rate-distortion function $\mathbb{R}_{\mathsf{S}}(d)$.

If conditions (i)–(iii) are satisfied, then $\lambda_{S^k} = k\lambda_{\mathsf{S}}$ and $P_{Z^{k\star}|S^k} = P_{\mathsf{Z}^\star|\mathsf{S}} \times \ldots \times P_{\mathsf{Z}^\star|\mathsf{S}}$, where $P_{\mathsf{Z}^\star|\mathsf{S}}$ achieves $\mathbb{R}_{\mathsf{S}}(d)$. Moreover, even if $\mathbb{R}_{\mathsf{S}}(d)$ is not achieved by any conditional distribution

$$\jmath_{S^k}(s^k, d) = \sum_{i=1}^{k} \jmath_{\mathsf{S}}(s_i, d) \tag{142}$$

Finiteness of the twelfth moment of $\mathsf{d}(\mathsf{S}, \mathsf{Z}^\star)$ in restriction (iv) is required for the achievability part of the asymptotic expansion in Theorem 9.

Theorem 9. *Under assumptions* (i)–(iv)*, for any* $0 \leq \epsilon \leq 1$

$$\left.\begin{array}{l} L^\star_{S^k}(d, \epsilon) \\ \mathbb{R}_{S^k}(d, \epsilon) \\ \mathbb{R}^+_{S^k}(d, \epsilon) \\ H_{d,\epsilon}(S^k) \\ \mathbb{E}\left[\left\langle \jmath_{S^k}(S^k, d)\right\rangle_\epsilon\right] \end{array}\right\} = (1-\epsilon)kR(d) - \sqrt{\frac{k\mathcal{V}(d)}{2\pi}} e^{-\frac{(Q^{-1}(\epsilon))^2}{2}} + \theta(k) \tag{143}$$

where

$$\mathcal{V}(d) = \mathrm{Var}\left[\jmath_{\mathsf{S}}(\mathsf{S}, d)\right] \tag{144}$$

is the rate-dispersion function, and the remainder term in the expansion satisfies

$$-2\log_2 k + O(1) \leq \theta(k) \leq \frac{3}{2}\log_2 k + O(1) \tag{145}$$

Proof. Due to (107), the assumption (iv) implies that the twelfth (and thus the third) moment of $\jmath_{\mathsf{S}}(\mathsf{S}, d)$ is finite, and the expansion for $\mathbb{E}\left[\left\langle \jmath_{S^k}(S^k, d)\right\rangle_\epsilon\right]$ follows from (142) and Lemma 1. The converse direction is now immediate from Theorems 7 and 8. The achievability direction follows by an application of Lemma 2 below to weaken the upper bounds in Theorems 7 and 8. □

Lemma 2. *Let* $0 \leq \epsilon \leq 1$. *Under assumptions* (i)–(iv)

$$\mathbb{E}\left[\left\langle -\log_2 P_{Z^{k\star}}(B_d(S^k))\right\rangle_\epsilon\right] = (1-\epsilon)kR(d) - \sqrt{\frac{k\mathcal{V}(d)}{2\pi}}e^{-\frac{(Q^{-1}(\epsilon))^2}{2}} + \theta(k) \tag{146}$$

where

$$O(1) \leq \theta(k) \leq \frac{1}{2}\log_2 k + O(1) \tag{147}$$

Proof. Appendix E. □

Appendix A
Proof of Lemma 1

The following non-uniform strengthening of the Berry-Esseén inequality is instrumental in the proof of Lemma 1.

Theorem 10 (Bikelis (1966), e.g. [31])**.** *Fix a positive integer* k. *Let* X_i, $i = 1, \ldots, k$ *be independent,* $\mathbb{E}[X_i] = 0$, $\mathbb{E}[|X_i|^3] < \infty$. *Then, for any real* t

$$\left|\mathbb{P}\left[\sum_{i=1}^k X_i > t\sqrt{kV_k}\right] - Q(t)\right| \leq \frac{B_k}{\sqrt{k}(1+|t|^3)}, \tag{148}$$

where

$$V_k = \frac{1}{k}\sum_{i=1}^k \mathbb{E}\left[|X_i|^2\right] \tag{149}$$

$$T_k = \frac{1}{k}\sum_{i=1}^k \mathbb{E}\left[|X_i|^3\right] \tag{150}$$

$$B_k = \frac{c_0 T_k}{V_k^{3/2}} \tag{151}$$

and c_0 *is a positive constant.*

Denote for brevity

$$Y_k \triangleq \sum_{i=1}^k X_i \tag{152}$$

If $\mathrm{Var}[\mathsf{X}] = 0$

$$\mathbb{E}[\langle Y_k\rangle_\epsilon] = (1-\epsilon)k\mathbb{E}[\mathsf{X}], \tag{153}$$

and (76) holds.

If $\mathrm{Var}\left[\mathsf{X}\right] > 0$ notice that

$$(1-\epsilon)k\mathbb{E}\left[\mathsf{X}\right] - \mathbb{E}\left[\langle Y_k \rangle_\epsilon\right] = \mathbb{E}\left[\left(Y_k - k\mathbb{E}\left[\mathsf{X}\right]\right)1\left\{Y_k > \eta\right\}\right] + \alpha\left(\eta - k\mathbb{E}\left[\mathsf{X}\right]\right)\mathbb{P}\left[Y_k = \eta\right] \tag{154}$$

$$= \int_\eta^\infty \mathbb{P}\left[Y_k > t\right] dt + \epsilon\left(\eta - k\mathbb{E}\left[\mathsf{X}\right]\right), \tag{155}$$

where η and α are those in (14), and to write (155) we used

$$\mathbb{E}\left[Y_k 1\left\{Y_k > \eta\right\}\right] = \int_\eta^\infty \mathbb{P}\left[Y_k > t\right] dt + \eta\mathbb{P}\left[Y_k > \eta\right]. \tag{156}$$

We proceed to evaluate the right side of (155). Using Theorem 10, we observe that η that satisfies (14) has the form

$$\eta = k\mathbb{E}\left[\mathsf{X}\right] + \sqrt{k\mathrm{Var}\left[\mathsf{X}\right]}Q^{-1}\left(\epsilon\right) + b_k \tag{157}$$

where $b_k = O\left(1\right)$. Using (157), we may write

$$\begin{aligned}
&\int_\eta^\infty \mathbb{P}\left[Y_k > t\right] dt \\
&= \int_0^\infty \mathbb{P}\left[Y_k > \eta + t\right] dt && (158)\\
&= \int_{b_k}^\infty \mathbb{P}\left[Y_k > k\mathbb{E}\left[\mathsf{X}\right] + \sqrt{k\mathrm{Var}\left[\mathsf{X}\right]}Q^{-1}\left(\epsilon\right) + t\right] dt && (159)\\
&= \int_0^\infty \mathbb{P}\left[Y_k > k\mathbb{E}\left[\mathsf{X}\right] + \sqrt{k\mathrm{Var}\left[\mathsf{X}\right]}Q^{-1}\left(\epsilon\right) + t\right] dt + O\left(1\right) && (160)\\
&= \sqrt{k\mathrm{Var}\left[\mathsf{X}\right]}\int_0^\infty \mathbb{P}\left[Y_k > k\mathbb{E}\left[\mathsf{X}\right] + \sqrt{k\mathrm{Var}\left[\mathsf{X}\right]}(Q^{-1}\left(\epsilon\right) + r)\right] dr + O\left(1\right) && (161)\\
&= \sqrt{k\mathrm{Var}\left[\mathsf{X}\right]}\int_0^\infty Q\left(Q^{-1}\left(\epsilon\right) + r\right) dr + O\left(1\right) && (162)\\
&= \sqrt{k\mathrm{Var}\left[\mathsf{X}\right]}\int_{Q^{-1}(\epsilon)}^\infty Q\left(r\right) dr + O\left(1\right) && (163)\\
&= \sqrt{k\mathrm{Var}\left[\mathsf{X}\right]}\left[\int_{Q^{-1}(\epsilon)}^\infty \frac{1}{\sqrt{2\pi}} x e^{-\frac{x^2}{2}} dx - \epsilon Q^{-1}\left(\epsilon\right)\right] + O\left(1\right) && (164)\\
&= \sqrt{k\mathrm{Var}\left[\mathsf{X}\right]}\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{(Q^{-1}(\epsilon))^2}{2}} - \epsilon Q^{-1}\left(\epsilon\right)\right) + O\left(1\right) && (165)
\end{aligned}$$

where (162) follows by applying Theorem 10 to the integrand in the left side and observing that

$$\int_0^\infty \frac{dr}{1 + \left(Q^{-1}\left(\epsilon\right) + r\right)^3} < \infty \tag{166}$$

Applying (157) and (165) to (155), we conclude that

$$(1-\epsilon)k\mathbb{E}\left[\mathsf{X}\right] - \mathbb{E}\left[\langle Y_k \rangle_\epsilon\right] = \frac{\sqrt{k\mathrm{Var}\left[\mathsf{X}\right]}}{\sqrt{2\pi}} e^{-\frac{(Q^{-1}(\epsilon))^2}{2}} + O\left(1\right), \tag{167}$$

which is exactly (76).

APPENDIX B
PROOF OF (85)

Denote for brevity

$$f(\epsilon) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\left(Q^{-1}(\epsilon)\right)^2}{2}} \tag{168}$$

Direct computation yields

$$f(\epsilon) = -\frac{1}{\left(Q^{-1}\right)'(\epsilon)} \tag{169}$$

$$f'(\epsilon) = Q^{-1}(\epsilon) \tag{170}$$

$$f''(\epsilon) = -\frac{1}{f(\epsilon)} \tag{171}$$

Furthermore, using the bounds

$$\frac{x}{\sqrt{2\pi}(1+x^2)} e^{-\frac{x^2}{2}} < Q(x) < \frac{1}{\sqrt{2\pi}x} e^{-\frac{x^2}{2}},\ x > 0 \tag{172}$$

we infer that as $\epsilon \to 0$

$$Q^{-1}(\epsilon) = \sqrt{2\log_e \frac{1}{\epsilon}} + O\left(\log_e \log_e \frac{1}{\epsilon}\right) \tag{173}$$

Finally

$$\lim_{\epsilon\to 0} \frac{f(\epsilon) - \epsilon\sqrt{2\log_e \frac{1}{\epsilon}}}{\epsilon} = \lim_{\epsilon\to 0} \frac{f(\epsilon) - \epsilon f'(\epsilon)}{\epsilon} \tag{174}$$

$$= \lim_{\epsilon\to 0} f''(\epsilon)\epsilon \tag{175}$$

$$= \lim_{\epsilon\to 0} \frac{-\epsilon}{f(\epsilon)} \tag{176}$$

$$= \lim_{\epsilon\to 0} \frac{1}{Q^{-1}(\epsilon)} \tag{177}$$

$$= 0 \tag{178}$$

where

- (174) is due to (170) and (173);
- (175) is by the l'Hôpital rule;
- (176) applies (171);
- (177) is by the l'Hôpital rule and (170).

APPENDIX C
PROOF OF THEOREM 8

Given P_S, d, denote for measurable $\mathcal{F} \subseteq \mathcal{M}$

$$\mathbb{R}_{S|\mathcal{F}}(d,\epsilon) \triangleq \min_{\substack{P_{Z|S}: \\ \mathbb{P}[\mathsf{d}(S,Z)>d|S\in\mathcal{F}]\le\epsilon}} I(S;Z|S\in\mathcal{F}) \tag{179}$$

In the proof of the converse bound in (130), the following result is instrumental.

Lemma 3. *Suppose P_S, d, $d > d_{\min}$ and $\mathcal{F} \subseteq \mathcal{M}$ are such that for all $s \in \mathcal{F}$*

$$\jmath_S(S,d) \ge r \text{ a.s.} \tag{180}$$

for some real r. Then

$$\mathbb{R}_{S|\mathcal{F}}(d,\epsilon) \ge \left|(1-\epsilon)r + (1-\epsilon)\log_2 \mathbb{P}\left[S\in\mathcal{F}\right] - h(\epsilon)\right|^+ \tag{181}$$

Proof. Denote

$$p_S(z) \triangleq \mathbb{P}\left[\mathsf{d}(S,z) \le d | S\in\mathcal{F}\right] \tag{182}$$

$$p \triangleq \sup_{z\in\widehat{\mathcal{M}}} p_S(z) \tag{183}$$

If $\epsilon > 1-p$, $\mathbb{R}_S(d,\epsilon) = 0$, so in the sequel we focus on the nontrivial case

$$\epsilon \le 1-p \tag{184}$$

To lower-bound the left side of (181), we weaken the supremum in (102) by selecting a suitable pair $(J(s),\lambda)$ satisfying the constraint in (103). Specifically, we choose

$$\exp(-\lambda) = \frac{\epsilon p}{(1-\epsilon)(1-p)} \tag{185}$$

$$\exp(J(s)) = \exp(J) \triangleq \frac{1-\epsilon}{p},\ s\in\mathcal{F} \tag{186}$$

To verify that the condition (103) is satisfied, we substitute the choice in (185) and (186) into the left side of (103) to obtain

$$\epsilon\frac{1-p_S(z)}{1-p} + (1-\epsilon)\frac{p_S(z)}{p} \le (1-p)\left[\frac{1-p_S(z)}{1-p} - \frac{p_S(z)}{p}\right] + \frac{p_S(z)}{p} \tag{187}$$

$$= 1 \tag{188}$$

where (187) is due to (184) and the observation that the expression in square brackets in the right side of (187) is nonnegative. Plugging (185) and (186) into (102), we conclude that

$$\mathbb{R}_{S|\mathcal{F}}(d,\epsilon) \geq J - \lambda\epsilon \tag{189}$$

$$= d(\epsilon\|1-p) - h(\epsilon) \tag{190}$$

$$\geq (1-\epsilon)\log_2\frac{1}{p} - h(\epsilon) \tag{191}$$

$$\geq (1-\epsilon)r + (1-\epsilon)\log_2\mathbb{P}\left[S\in\mathcal{F}\right] - h(\epsilon) \tag{192}$$

where (192) is due to

$$p_S(z) \leq \mathbb{E}\left[\exp(\lambda_S d - \lambda_S \mathsf{d}(S,z))|S\in\mathcal{F}\right] \tag{193}$$

$$\leq \mathbb{E}\left[\exp(\jmath_S(S,d) + \lambda_S d - \lambda_S \mathsf{d}(S,z) - r)|S\in\mathcal{F}\right] \tag{194}$$

$$\leq \frac{\exp(-r)}{\mathbb{P}\left[S\in\mathcal{F}\right]}\mathbb{E}\left[\exp(\jmath_S(S,d) + \lambda_S d - \lambda_S \mathsf{d}(S,z))\right] \tag{195}$$

$$\leq \frac{\exp(-r)}{\mathbb{P}\left[S\in\mathcal{F}\right]} \tag{196}$$

where $\lambda_S \triangleq -\mathbb{R}_S(d)$, and

- (193) is Markov's inequality;
- (194) applies (180);
- (196) is equivalent to (110).

□

Proof of Theorem 8. We start with the converse bound in (130). Note first that, similar to (38), the constraint in (19) is achieved with equality. Denoting the random variable

$$F \triangleq \lfloor \jmath_S(S,d)\rfloor + 1 \tag{197}$$

and the sets

$$\mathcal{F}_j \triangleq \{s\in\mathcal{M}\colon F = j\}, \tag{198}$$

we may write

$$I(S;Z) = I(S,F;Z) \tag{199}$$

$$= I(S;Z|F) + I(F;Z) \tag{200}$$

so

$$\mathbb{R}_S(d,\epsilon) \geq \min_{\substack{P_{Z|S}:\\ \mathbb{P}[\mathsf{d}(S,Z)>d]\leq\epsilon}} I(S;Z|F) \tag{201}$$

$$= \min_{\varepsilon(\cdot):\, \mathbb{E}[\varepsilon(F)]\leq\epsilon} \sum_{j=-\infty}^{\infty} P_F(j)\mathbb{R}_{S|\mathcal{F}_j}(d,\epsilon(j)) \tag{202}$$

We apply Lemma 3 to lower bound each term of the sum by

$$\mathbb{R}_{S|\mathcal{F}_j}(d,\epsilon(j)) \geq \left|(1-\epsilon(j))j + (1-\epsilon)\log_2 P_F(j) - h(\epsilon(j))\right|^+ \tag{203}$$

to obtain

$$\mathbb{R}_S(d,\epsilon) \geq \min_{\varepsilon(\cdot):\, \mathbb{E}[\varepsilon(F)]\leq\epsilon} \left\{\mathbb{E}\left[(1-\epsilon(F))\jmath_S(S,d)\right] - \mathbb{E}\left[h(\epsilon(F))\right]\right\} - H(F) \tag{204}$$

$$= \min_{\varepsilon(\cdot):\, \mathbb{E}[\varepsilon(F)]\leq\epsilon} \left\{\mathbb{E}\left[(1-\epsilon(F))\jmath_S(S,d)\right]\right\} - H(F) - h(\epsilon) \tag{205}$$

$$\geq \mathbb{E}\left[\langle \jmath_S(S,d)\rangle_\epsilon\right] - H(F) - h(\epsilon) \tag{206}$$

$$\geq \mathbb{E}\left[\langle \jmath_S(S,d)\rangle_\epsilon\right] - \log_2\left(\mathbb{E}\left[J_S(S)\right]+1\right) - \log_2 e - h(\epsilon) \tag{207}$$

where (204) uses (111), (205) is by concavity of $h(\cdot)$, (206) is due to (139), and (207) holds because $F+\lambda_S d \geq J_S(S) \geq 0$, and the entropy of a random variable on $\mathbb{Z}_+$ with a given mean is maximized by that of the geometric distribution.

To show the upper bound in (131), fix an arbitrary distribution $P_{\bar{Z}}$ and define the conditional probability distribution $P_{Z|S}$ through[6]

$$\frac{dP_{Z|S=s}(z)}{dP_{\bar{Z}}(z)} = \begin{cases} \frac{1\{\mathsf{d}(s,z)\leq d\}}{P_{\bar{Z}}(B_d(s))} & \langle -\log_2 P_{\bar{Z}}(B_d(s))\rangle_\epsilon > 0 \\ 1 & \text{otherwise} \end{cases} \tag{208}$$

By the definition of $P_{Z|S}$

$$\mathbb{P}\left[\mathsf{d}(S,Z) > d\right] \leq \epsilon \tag{209}$$

Upper-bounding the minimum in (19) with the choice of $P_{Z|S}$ in (208), we obtain the following

[6]Note that in general $P_S \to P_{Z|S} \nrightarrow P_{\bar{Z}}$.

nonasymptotic bound:

$$\mathbb{R}_S(d,\epsilon) \le I(S;Z) \tag{210}$$
$$= D\left(P_{Z|S}\|P_{\bar{Z}}|P_S\right) - D(P_Z\|P_{\bar{Z}}) \tag{211}$$
$$\le D\left(P_{Z|S}\|P_{\bar{Z}}|P_S\right) \tag{212}$$
$$= \mathbb{E}\left[\langle -\log_2 P_{\bar{Z}}(B_d(S))\rangle_\epsilon\right] \tag{213}$$

which leads to (131) after minimizing the right side over all $P_{\bar{Z}}$.

To show the lower bound on (ϵ,δ)-entropy in (132), fix f satisfying the constraint in (113), denote

$$Z \triangleq \mathsf{f}(S) \tag{214}$$
$$\varepsilon(s) \triangleq 1\left\{\mathsf{d}(s,\mathsf{f}(s)) > d\right\} \tag{215}$$

and write

$$H(Z) \ge H(Z|\varepsilon(S)) \tag{216}$$
$$\ge P_{\varepsilon(S)}(0) H(Z|\varepsilon(S)=0) \tag{217}$$
$$= \mathbb{E}\left[\imath_{Z,\varepsilon(S)=0}(Z)(1-\varepsilon(S))\right] + P_{\varepsilon(S)}(0)\log_2 P_{\varepsilon(S)}(0) \tag{218}$$
$$\ge \mathbb{E}\left[\langle -\log_2 P_Z(B_d(S))\rangle_\epsilon\right] - \phi(\min\{\epsilon, e^{-1}\}) \tag{219}$$

where the second term is bounded by maximizing $p\log_2\frac{1}{p}$ over $[1-\epsilon,1]$, and the first term is bounded via the following chain.

$$\mathbb{E}\left[\imath_{Z,\varepsilon(S)=0}(Z)(1-\varepsilon(S))\right] \ge \mathbb{E}\left[-\log_2 P_Z(B_d(S))(1-\varepsilon(S))\right] \tag{220}$$
$$\ge \min_{\varepsilon(\cdot)\colon \mathbb{E}[\varepsilon(S)]\le\epsilon} \mathbb{E}\left[-\log_2 P_Z(B_d(S))(1-\varepsilon(S))\right] \tag{221}$$
$$= \mathbb{E}\left[\langle -\log_2 P_Z(B_d(S))\rangle_\epsilon\right] \tag{222}$$

where (220) holds because due to $\{s\in\mathcal{M}\colon \mathsf{f}(s)=z, \epsilon(s)=0\} \subseteq B_d(s)$ we have for all $s\in\mathcal{M}$

$$\mathbb{P}\left[Z=\mathsf{f}(s), \varepsilon(S)=0\right] \le P_Z(B_d(s)) \tag{223}$$

and (222) is due to (37).

To show the upper bound on (ϵ,δ)-entropy in (133), fix P_Z such

$$P_Z(B_d(s)) > 0 \tag{224}$$

for P_S-a.s. $s \in \mathcal{M}$, let $Z^\infty \sim P_Z \times P_Z \times \ldots$, and define W as

$$W \triangleq \begin{cases} \min\{m\colon \mathsf{d}(S, Z_m) \le d\} & \langle -\log_2 P_Z(B_d(S))\rangle_{\epsilon'} > 0 \\ 1 & \text{otherwise} \end{cases} \tag{225}$$

where ϵ' is the maximum of $\epsilon' \le \epsilon$ such that the randomization on the boundary of $\langle -\log_2 P_Z(B_d(S))\rangle_{\epsilon'}$ can be implemented without the actual randomization (see Section II-A for an explanation of this phenomenon).

If $z_1, z_2, \ldots$ is a realization of Z^∞, $\mathsf{f}(s) = z_w$ is a deterministic mapping that satisfies the constraint in (113), so, since $w \mapsto z_w$ is injective, we have

$$H_{d,\epsilon}(S) \le H(W|Z^\infty = z^\infty) \tag{226}$$

We proceed to show that $H(W|Z^\infty)$ is upper bounded by the right side of (133). Via the random coding argument this will imply that there exists at least one codebook z^∞ such that $H(W|Z^\infty = z^\infty)$ is also upper bounded by the right side of (133), and the proof will be complete.

Let

$$G \triangleq \lfloor \log_2 W \rfloor \, \langle -\log_2 P_Z(B_d(S))\rangle_{\epsilon'} > 0 \tag{227}$$

and consider the chain

$$\begin{aligned} H(W|Z^\infty) &\le H(W) && (228)\\ &= H(W|G) + I(W;G) && (229)\\ &\le \mathbb{E}\,[G] + H(G) && (230)\\ &\le \mathbb{E}\,[G] + \log_2(1 + \mathbb{E}\,[G]) + \log_2 e && (231) \end{aligned}$$

where

- (228) holds because conditioning decreases entropy;
- (230) holds because conditioned on $G = i$, W can have at most i values;
- (231) holds because the entropy of a positive integer-valued random variable with a given mean is maximized by the geometric distribution.

Finally, it was shown in (129) that

$$\begin{aligned} \mathbb{E}\,[G] &= \mathbb{E}\,[\langle -\log_2 P_Z(B_d(S))\rangle_{\epsilon'}] && (232)\\ &\le \mathbb{E}\,[\langle -\log_2 P_Z(B_d(S))\rangle_{\epsilon}] + \phi(\min\{\epsilon, e^{-1}\}) && (233) \end{aligned}$$

where $\phi(\cdot)$ is the no-randomization penalty as explained in the proof of (31).

□

APPENDIX D
PROOF OF THE BOUNDS (134) AND (135) ON $H_{0,\epsilon}(S)$ (HAMMING DISTORTION)

The upper bound in (135) is obtained by a suboptimal choice (in (121)) of $\mathsf{f}(s) = s$ for all $s \leq m_0$, where m_0 is that in (33), and $\mathsf{f}(s) = m_0 + 1$ otherwise.

To show the lower bound in (134), fix f satisfying the constraint in (121), put

$$\varepsilon(S) \triangleq 1\{S \neq \mathsf{f}(S)\} \tag{234}$$

and write

$$H(\mathsf{f}(S)) \geq H(\mathsf{f}(S)|\varepsilon(S)=0)P_{\varepsilon(S)}(0) \tag{235}$$

$$= \mathbb{E}\left[\log_2 \frac{1}{P_{\mathsf{f}(S)|\varepsilon(S)=0}(S)}|\varepsilon(S)=0\right] P_{\varepsilon(S)}(0) \tag{236}$$

$$\geq H\left(S|\varepsilon(S)=0\right) P_{\varepsilon(S)}(0) \tag{237}$$

$$= \mathbb{E}\left[\imath_S(S)1\left\{\varepsilon(S)=0\right\}\right] + P_{\varepsilon(S)}(0)\log_2 P_{\varepsilon(S)}(0) \tag{238}$$

$$\geq \mathbb{E}\left[\langle \imath_S(S)\rangle_\epsilon\right] - \phi\left(\max\left\{1-\epsilon, e^{-1}\right\}\right) \tag{239}$$

where

- (235) is because conditioning decreases entropy;
- (236) is due to

$$\min_{P_Y} \mathbb{E}\left[\imath_Y(X)\right] = H(X) \tag{240}$$

- in (239), the first term is bounded using (37), and the second term is bounded by maximizing $p\log_2\frac{1}{p}$ over $[1-\epsilon, 1]$.

APPENDIX E
PROOF OF LEMMA 2

The following refinement of the lossy AEP is essentially contained in [19].

Lemma 4. *Under restrictions* (i)–(iv)*, there exist constants* C_1, C_2 *such that eventually, almost surely*

$$\log_2 \frac{1}{P_{Z^{k\star}}(B_d(S^k))} \le \sum_{i=1}^{k} \jmath_{\mathsf{S}}(S_i, d) + \frac{1}{2}\log_2 k - k\lambda_{\mathsf{S}}(d - \bar{\mathsf{d}}(S^k)) + kC_1(d - \bar{\mathsf{d}}(S^k))^2 + C_2 \tag{241}$$

where

$$\bar{\mathsf{d}}(s^k) \triangleq \frac{1}{k}\sum_{i=1}^{k} \mathbb{E}\left[\mathsf{d}(s_i, \mathsf{Z}^\star)|\mathsf{S} = s_i\right] \tag{242}$$

Proof. It follows from [19, (4.6), (5.5)] that the probability of violating (241) is $O\left(\frac{1}{k^2}\right)$. Since $\sum_{k=1}^{\infty}\frac{1}{k^2}$ is summable, by the Borel-Cantelli lemma (241) holds w. p. 1 for k large enough. □

Noting that $\bar{\mathsf{d}}(s^k)$ is a normalized sum of independent random variables with mean d, we conclude using Lemma 4 that for k large enough

$$\mathbb{E}\left[\log_2 \frac{1}{P_{Z^{k\star}}(B_d(S^k))}\right] \le kR(d) + \frac{1}{2}\log_2 k + O(1) \tag{243}$$

Lemma 2 is now immediate from (140) and (141) and the expansion for $\mathbb{E}\left[\left\langle \jmath_{S^k}(S^k, d)\right\rangle_\epsilon\right]$ in (143).

REFERENCES

[1] V. Kostina, Y. Polyanskiy, and S. Verdú, "Variable-length compression allowing errors," in *Proceedings 2014 IEEE International Symposium on Information Theory*, Honolulu, HI, July 2014, pp. 2679–2683.

[2] 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.

[3] I. Kontoyiannis and S. Verdú, "Optimal lossless data compression: Non-asymptotics and asymptotics," *IEEE Transactions on Information Theory*, vol. 60, no. 2, pp. 777–795, Feb. 2014.

[4] 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.

[5] A. D. Wyner, "An upper bound on the entropy series," *Inf. Contr.*, vol. 20, no. 2, pp. 176–181, 1972.

[6] T. S. Han, "Weak variable-length source coding," *IEEE Transactions on Information Theory*, vol. 46, no. 4, pp. 1217–1226, 2000.

[7] ——, *Information-Spectrum Methods in Information Theory*. Springer, Berlin, 2003.

[8] H. Koga and H. Yamamoto, "Asymptotic properties on codeword lengths of an optimal FV code for general sources," *IEEE Transactions on Information Theory*, vol. 51, no. 4, pp. 1546–1555, 2005.

[9] 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.

[10] 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, Aug. 2011.

[11] H. Koga, "Source coding using families of universal hash functions," *IEEE Transactions on Information Theory*, vol. 53, no. 9, pp. 3226–3233, Sep. 2007.

[12] V. Erokhin, "Epsilon-entropy of a discrete random variable," *Theory of Probability and Applications*, vol. 3, pp. 97–100, 1958.

[13] V. Strassen, "Asymptotische abschätzungen in Shannon's informationstheorie," in *Proceedings 3rd Prague Conference on Information Theory*, Prague, Czechoslovakia, 1962, pp. 689–723.

[14] 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.

[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] O. Kosut and L. Sankar, "Asymptotics and non-asymptotics for universal fixed-to-variable source coding," *arXiv preprint arXiv:1412.4444*, 2014.

[17] 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.

[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] 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.

[20] E. C. Posner, E. R. Rodemich, and H. Rumsey, "Epsilon-entropy of stochastic processes," *The Annals of Mathematical Statistics*, vol. 38, no. 4, pp. 1000–1020, 1967.

[21] 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.

[22] F. Jelinek, *Probabilistic information theory: discrete and memoryless models*. McGraw-Hill, 1968.

[23] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed. Cambridge Univ Press, 2011.

[24] A. A. Yushkevich, "On limit theorems connected with the concept of entropy of Markov chains," *Uspekhi Matematicheskikh Nauk*, vol. 8, no. 5, pp. 177–180, 1953.

[25] W. Szpankowski, "A one-to-one code and its anti-redundancy," *IEEE Transactions on Information Theory*, vol. 54, no. 10, pp. 4762–4766, 2008.

[26] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRE Int. Conv. Rec.*, vol. 7, 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.

[27] 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.

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

[29] E. C. Posner and E. R. Rodemich, "Epsilon-entropy and data compression," *The Annals of Mathematical Statistics*, vol. 42, pp. 2079–2125, 1971.

[30] A. Dembo and I. Kontoyiannis, "Critical behavior in lossy source coding," *IEEE Transactions on Information Theory*, vol. 47, no. 3, pp. 1230–1236, 2001.

[31] V. V. Petrov, *Limit Theorems of Probability Theory*. Oxford Science Publications, 1995.