Universal Quadratic Lower Bounds on Source Coding Error Exponents

Cheng Chang and Anant Sahai

***Abstract*— We consider the problem of block-size selection to achieve a desired probability of error for universal source coding. While Baron, *et al* in [1], [9] studied this question for rates in the vicinity of entropy for *known* distributions using central-limit-theorem techniques, we are interested in all rates for *unknown* distributions and use error-exponent techniques. By adapting a technique of Gallager from the exercises of [7], we derive a universal lower bound to the source-coding error exponent that depends only on the alphabet size and is quadratic in the gap to entropy.**

I. Introduction

In [10], the lossless source coding with decoder side-information problem, as shown in Figure 1, is introduced. The source and decoder side-information sequence (x_1^n, y_1^n) are drawn iid from a joint distribution p_{xy} on a finite alphabet $\mathcal{X} \times \mathcal{Y}$. If the decoder knows y_1^n, the error probability $Pr(\hat{x}_1^n \neq x_1^n)$, goes to 0, as the code length n goes to infinity, for any rate $R > H(p_{x|y})$, where $H(p_{x|y})$ is the conditional entropy of x given y.

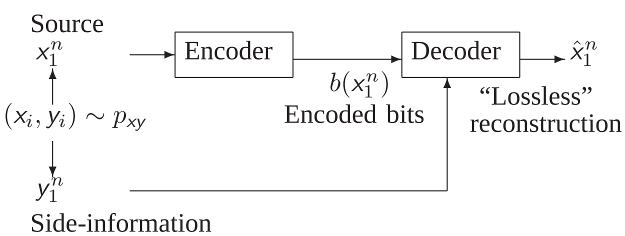

Fig. 1. Lossless source coding with decoder side-information

The performance of the coding system, i.e. how fast the error probability converges to 0 with block length n, when the coding rate is above the minimum required rate, is studied in [5], [6], [7]. We summarize the relevant error exponent results from the literature in the following.

Theorem 1: [6] Assume a decoder with access to the side information, where the memoryless source and side information are generated from a distribution p_{xy}. A random binning encoder and jointly ML decoding system, shown in Figure 1, has error probability $Pr(\hat{x}_1^n \neq x_1^n) \leq e^{-nE_r(R)}$, where

$$E_r(R) = \max_{0 \leq \rho \leq 1} \rho R - \bar{E}_0(\rho) \quad (1)$$

$$\text{where } \bar{E}_0(\rho) = \ln(\sum_y (\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}})^{1+\rho})$$

Without decoder side information, the Gallager function $\bar{E}_0$ simplifies[1] to:

$$E_0(\rho) = (1+\rho)\ln(\sum_{x \in \mathcal{X}} p_x(x)^{\frac{1}{1+\rho}})$$

In Theorem 1, the random binning scheme at the encoder is uniform, and thus universal in nature [6]. However for the ML decoding rule, the decoder needs to know the statistics of the source. In [5], a universal system based on minimum entropy decoding is shown to achieve the same error exponent asymptotically. For the universal decoder,

$$Pr(\hat{x}_1^n \neq x_1^n) \leq e^{-n(E_r(R) - \phi(n))} \quad (2)$$

Where $\phi(n)$ is the vanishing term $\frac{|\mathcal{X}| \ln n}{n}$ for the case without side-information and $\phi(n) = \frac{|\mathcal{X}||\mathcal{Y}| \ln n}{n}$ for the case with decoder side-information.

A. Motivation and related work

For fixed block source coding systems, block length is an important parameter as it is related to both system delay and complexity. Suppose there is a system-level requirement that the block error probability $Pr(x_1^n \neq \hat{x}_1^n)$ be below some constant $P_e > 0$. If the distributions are known, the minimum block length can be calculated from Theorem 1. However, the exact distribution need not be available to the encoder since that knowledge is not needed to do uniform binning. Thus, a universal estimate to the error exponent is desirable.

A related problem is studied in [1], [9]. They turn the question around and ask: for non-asymptotic length source coding with side-information, what is the minimum rate required to achieve block error P_e assuming that the distribution is in fact known? A more quantitative discussion of the relation between the problem in [1] and our work here is deferred to Section II-A.

C. Chang and A. Sahai are with the Department of Electrical Engineering and Computer Science, University of California at Berkeley. Email: cchang@eecs.berkeley.edu, sahai@eecs.berkeley.edu

[1]This is the source coding counterpart of the channel coding result in Theorem 5.6.4 [7].

B. A universal bound on channel coding exponents

In Exercise 5.23 [7], Gallager gives a quadratic lower bound on the random channel coding error exponent for a discrete memoryless channel $P(\cdot|\cdot)$ with output alphabet size J. If Q is the distribution that achieves the channel capacity C, then the random coding error exponent $E_r^c(R, Q)$, defined in Theorem 5.6.4 of [7] is lower bounded by the following quadratic function of the gap to capacity $(C - R)$ for all $R < C$.

$$E_r^c(R, Q) \geq \frac{(C-R)^2}{8/e^2 + 4[\ln J]^2} \tag{3}$$

This bound can be further tightened and we give the new result as a corollary to Lemma 1. This bound is universal in the sense that it only depends on the size of output alphabet and the gap to capacity and not on the detailed channel statistics.

Following Gallager's techniques, we derive universal quadratic bounds on the random source coding error exponent with and without decoder side-information. For both cases, the quadratic bounds are only determined by the gap to entropy and the size of the source alphabet $|\mathcal{X}|$. The results are summarized in Theorem 2. The proof details are in Section III.

II. MAIN RESULTS AND DISCUSSION

Theorem 2: For a memoryless source x and decoder side information y, jointly generated iid from $p_{\mathsf{x}\mathsf{y}}$ with conditional entropy $H(p_{\mathsf{x}|\mathsf{y}}) = h$ on finite alphabet $\mathcal{X} \times \mathcal{Y}$, the random coding error exponent $E_r(R)$, defined in (1), is lower bounded by a quadratic function, $\forall R \in [H(p_{\mathsf{x}|\mathsf{y}}), \ln|\mathcal{X}|)$:

$$E_r(R) \geq G_h(R)$$

$$\text{where} \quad G_h(R) = \begin{cases} \frac{(R-h)^2}{2(\ln|\mathcal{X}|)^2} & \text{if } |\mathcal{X}| \geq 3 \\ \frac{(R-h)^2}{2\ln 2} & \text{if } |\mathcal{X}| = 2 \end{cases} \tag{4}$$

Because the bound depends only on the gap to entropy, we can also write it as $G(R-h)$. Furthermore, if there is no side-information, the source is from p_{x}, s.t. $H(p_{\mathsf{x}}) = h$ and the same bound applies.

It is interesting to note that this quadratic bound on the error exponent $E_r(R)$ has no dependence on the size of the side-information alphabet $|\mathcal{Y}|$.

A. Discussion and Examples

For sources with $H(p_{\mathsf{x}}) = h$, it is easy to see that $R - h$ is always an upper bound to $E_r(R)$ and hence Theorem 2 implies:

$$G_h(R) \leq \min_{p_{\mathsf{x}}:H(p_{\mathsf{x}})=h} E_r(R) \leq \max_{p_{\mathsf{x}}:H(p_{\mathsf{x}})=h} E_r(R) \leq R - h$$

To illustrate the looseness of our bounds, consider $|\mathcal{X}| = 3$, and distributions p_{x}, s.t. $H(p_{\mathsf{x}}) = h = 0.394$. Since the alphabet size is so small, we can use brute-force optimization to obtain the upper and lower contours of possible $E_r(R)$. These are plotted along with the universal quadratic lower bound $G(R - h) = \frac{(R-h)^2}{2(\ln 3)^2}$ and the linear upper bound $R - h$ in Figure 2.

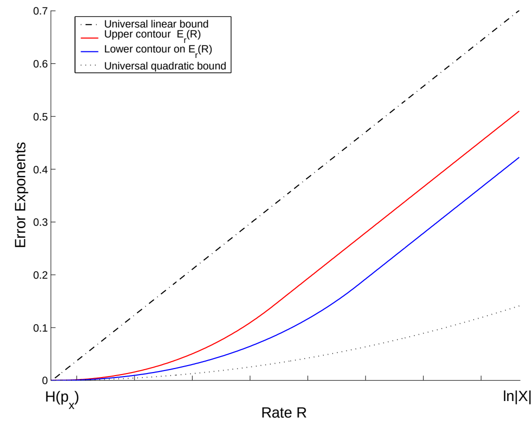

Fig. 2. Plot of the error exponents bounds for a three letter alphabet. In order from top to bottom: $R - H(p_{\mathsf{x}})$, $\max_{p_{\mathsf{x}}:H(p_{\mathsf{x}})=h} E_r(R)$, $\min_{p_{\mathsf{x}}:H(p_{\mathsf{x}})=h} E_r(R)$, and $G_h(R)$

For ML decoding where the decoder knows the distribution, the encoder can pick a block length sufficient to achieve block error probability $Pr(x_1^n \neq \hat{x}_1^n) \leq P_e$, knowing only the gap to entropy $R - h$ and the source alphabet size $|\mathcal{X}|$ by choosing

$$n \geq -\ln P_e / G(R - h)$$

The size of the side-information alphabet $|\mathcal{Y}|$ is not needed at all! If the decoder is also ignorant of the joint distribution, then $\phi(n)$ in (2) must be taken into consideration. n can be chosen by solving:

$$nG(R - h) - n\phi(n) \geq -\ln P_e$$

Since $n\phi(n) = |\mathcal{X}||\mathcal{Y}| \ln n$, this implies that determining n from our bound requires the encoder to know the side-information alphabet size. At low probabilities of error, the dependence is relatively weak however since the $\ln n$ is dominated by the $nG(R - h)$ term.

In [1], it is shown that, for source coding with side-information, the required rate is $H(p_{\mathsf{x}|\mathsf{y}}) + K(P_e)\sqrt{-\ln P_e/n} + o(\sqrt{-\ln P_e/n})$ to achieve block error probability P_e with fixed block length n — where the $O(\sqrt{-\ln P_e/n})$ is called the redundancy rate. The exact constant $K(P_e)$ is also computed and clearly depends on the probability distribution of the source. The converse, proved in [1] for binary symmetric sources, is not universal and moreover, cannot be made so. A simple counter-example is that given $\mathsf{y} = y$, suppose x is uniform on some subset of $S_y \subset \mathcal{X}$, where $|S_y| = K' < |\mathcal{X}|$ for all $y \in \mathcal{Y}$. In this case, the random coding error exponent $E_r(R)$ is a straight line $E_r(R) = R - \ln |K|$. For this example, the redundancy rate needs to be $\ln P_e/n$ when using random coding, and could potentially be zero with some other scheme.

Theorem 2 tells us that for block length n, rate $R = H(p_{\mathsf{x}|\mathsf{y}}) + K''(|\mathcal{X}|)\sqrt{-\ln P_e/n}$ then the block error is smaller than P_e, no matter what distribution is encountered. While not tight, it does show that a redundancy of $O(\sqrt{-\ln P_e/n})$ suffices for universality.

III. Proof of Theorem 2

In this section we prove Theorem 2. First, we need a technical lemma and definitions of tilted distributions[6].

A. Lemmas and Definitions

In this paper we use the following lemma to upper bound a non-concave function $f_E(\cdot)$.

Lemma 1: For constant $E \geq 0$, write

$$f_E(\vec{\omega}) = \sum_{j=1}^{J} \omega_j (\ln \omega_j - E)^2 \tag{5}$$

$\vec{\omega} \in \mathcal{S}_J$, where $\mathcal{S}_J = \{\vec{\omega} \in \mathcal{R}^J | \sum_k \omega_k = 1, \text{ and } \omega_j \geq 0, \forall j\}$ is the probability simplex of dimension J. Then for any distribution $\vec{\omega} \in \mathcal{S}_J$,

$$f_E(\vec{\omega}) \leq \begin{cases} E^2 + 2E(\ln J) + (\ln J)^2 & \text{if } J \geq 3 \\ E^2 + 2E(\ln 2) + T & \text{if } J = 2 \end{cases} \tag{6}$$

$$\text{where } T = t_1(\ln t_1)^2 + t_2(\ln t_2)^2 \tag{7}$$

$$\text{and } t_1 = \frac{1 + \sqrt{1 - 4e^{-2}}}{2}; \quad t_2 = \frac{1 + \sqrt{1 + 4e^{-2}}}{2}$$

$T \approx 0.563 > (\ln 2)^2$ and $T < \ln 2$.

The proof is in the appendix. The challenge in the proof lies in the non-concavity of $f_E(\vec{\omega})$. In Figure 3, for $J = 2$ thus $\vec{\omega} = (x, 1-x)$ and $E = 0$, we plot $f_0((x, 1-x))$. The maximum occurs at $x = t_1$ or t_2 which are defined above.

Definition 1: Tilted distributions: For a distribution p_{x} on a finite alphabet $\mathcal{X}$, $\rho \in (-1, \infty)$, we denote the $\rho-$tilted distribution by p_{x}^{ρ}, where

$$p_{\mathsf{x}}^{\rho}(x) = \frac{p_{\mathsf{x}}(x)^{\frac{1}{1+\rho}}}{\sum\limits_{s \in \mathcal{X}} p_{\mathsf{x}}(s)^{\frac{1}{1+\rho}}}$$

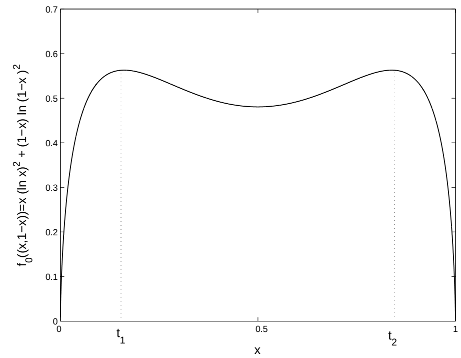

Fig. 3. Non concaveness of $f_E(\vec{\omega})$ defined in (5) , $E = 0$, $J = 2$.

For a distribution p_{xy} on a finite alphabet $\mathcal{X} \times \mathcal{Y}$, we denote $\mathsf{x} - \mathsf{y}$ tilted distribution of p_{xy} by $\bar{p}_{\mathsf{xy}}^{\rho}$,

$$\bar{p}_{\mathsf{xy}}^{\rho}(x, y) = \frac{[\sum\limits_{s \in \mathcal{X}} p_{\mathsf{xy}}(s, y)^{\frac{1}{1+\rho}}]^{1+\rho}}{\sum\limits_{t \in \mathcal{Y}} [\sum\limits_{s \in \mathcal{X}} p_{\mathsf{xy}}(s, t)^{\frac{1}{1+\rho}}]^{1+\rho}} \times \frac{p_{\mathsf{xy}}(x, y)^{\frac{1}{1+\rho}}}{\sum\limits_{s \in \mathcal{X}} p_{\mathsf{xy}}(s, y)^{\frac{1}{1+\rho}}}$$

Obviously $p_{\mathsf{x}}^0 = p_{\mathsf{x}}$ and $\bar{p}_{\mathsf{xy}}^0 = p_{\mathsf{xy}}$. Write the marginal distribution of y under distribution $\bar{p}_{\mathsf{xy}}^{\rho}$ as $\bar{p}_{\mathsf{y}}^{\rho}$ and the conditional distribution of x given y under distribution $\bar{p}_{\mathsf{xy}}^{\rho}$ as $\bar{p}_{\mathsf{x}|\mathsf{y}}^{\rho}$, then from the definition:

$$\begin{aligned} \bar{p}_{\mathsf{xy}}^{\rho}(x, y) &= \bar{p}_{\mathsf{y}}^{\rho}(y)\bar{p}_{\mathsf{x}|\mathsf{y}}^{\rho}(x|y) \\ \bar{p}_{\mathsf{y}}^{\rho}(y) &= \frac{[\sum\limits_{s \in \mathcal{X}} p_{\mathsf{xy}}(s, y)^{\frac{1}{1+\rho}}]^{1+\rho}}{\sum\limits_{t \in \mathcal{Y}} [\sum\limits_{s \in \mathcal{X}} p_{\mathsf{xy}}(s, t)^{\frac{1}{1+\rho}}]^{1+\rho}} \\ \bar{p}_{\mathsf{x}|\mathsf{y}}^{\rho}(x|y) &= \frac{p_{\mathsf{xy}}(x, y)^{\frac{1}{1+\rho}}}{\sum\limits_{s \in \mathcal{X}} p_{\mathsf{xy}}(s, y)^{\frac{1}{1+\rho}}} \end{aligned}$$

Denote the entropy of p_{x}^{ρ} by $H(p_{\mathsf{x}}^{\rho})$ and the conditional entropy of x given y under distribution $\bar{p}_{\mathsf{xy}}^{\rho}$ by $H(\bar{p}_{\mathsf{x}|\mathsf{y}}^{\rho})$, then $H(\bar{p}_{\mathsf{x}|\mathsf{y}}^{0}) = H(p_{\mathsf{x}|\mathsf{y}})$. Write $H(\bar{p}_{\mathsf{x}|\mathsf{y}=y}^{\rho})$ as the conditional entropy of x given $\mathsf{y} = y$, then: $H(\bar{p}_{\mathsf{x}|\mathsf{y}=y}^{\rho}) = -\sum_x \bar{p}_{\mathsf{x}|\mathsf{y}}^{\rho}(x|y) \ln \bar{p}_{\mathsf{x}|\mathsf{y}}^{\rho}(x|y)$.

B. Proof of the case without side-information

As in the solution to Gallager's problem 5.23 in [7], we use the Taylor expansion for $E_0(R)$ to find a quadratic bound on $E_r(R)$.

Proof: From the mean value theorem, we expand $E_0(\rho)$ at 0, where $\rho \leq 1$, $\exists \rho_1 \in [0, \rho]$ s.t.

$$E_0(\rho) = E_0(0) + \rho E_0'(0) + \frac{\rho^2}{2} E_0''(\rho_1)$$

From basic calculus, as shown in the appendix of [3], it can be shown that

$$E_0^{'}(\rho) = \frac{dE_0(\rho)}{d\rho} = H(p_x^\rho) \quad (8)$$

and hence $E_0^{'}(0) = H(p_x)$. Note that $E_0(0) = 0$, and so

$$E_0(\rho) \le \rho H(p_x) + \frac{\rho^2}{2}\alpha \quad (9)$$

where $\alpha > 0$ is any upper bound of $E_0^{''}(\rho_1)$ that holds, $\forall \rho_1 \in [0,1]$. Substitute (9) into the definition of $E_r(R)$ to get $E_r(R)$

$$\begin{aligned}
&= \max_{0\le\rho\le1} \rho R - E_0(\rho) \\
&\ge \max_{0\le\rho\le1} \rho R - \rho H(p_x) - \frac{\rho^2}{2}\alpha \\
&= \max_{0\le\rho\le1} -\frac{\alpha}{2}(\rho - \frac{R - H(p_x)}{\alpha})^2 + \frac{(R-H(p_x))^2}{2\alpha} \\
&= \frac{(R-H(p_x))^2}{2\alpha} \quad (10)
\end{aligned}$$

for $R - H(p_x) \le \alpha$. In the last step, we note that $\rho = \frac{R-H(p_x)}{\alpha}$ is the maximizer, which is within $[0,1]$. To find α, an upper bound on $E_0^{''}(\rho)$, $\forall \rho \in [0,1]$, we expand $E_0^{''}(\rho) = \frac{dH(p_x^\rho)}{d\rho}$

$$\begin{aligned}
&= -\sum_x (1 + \ln p_x^\rho(x)) \frac{dp_x^\rho(x)}{d\rho} \\
&= \sum_x (1 + \ln p_x^\rho(x)) \frac{p_x^\rho(x)}{1+\rho} (\ln p_x^\rho(x) + H(p_x^\rho)) \\
&= \frac{1}{1+\rho} \sum_x \{ p_x^\rho(x)(\ln p_x^\rho(x))^2 + p_x^\rho(x) \ln p_x^\rho(x) \\
&\quad + p_x^\rho(x) H(p_x^\rho) + p_x^\rho(x) \big(\ln p_x^\rho(x) \big) H(p_x^\rho) \} \\
&= \frac{1}{1+\rho} \sum_x p_x^\rho(x) (\ln p_x^\rho(x))^2 - \frac{H(p_x^\rho)^2}{1+\rho} \quad (11)
\end{aligned}$$

Since the last term in (11) is negative[2] and $\rho > 0$, by the definition of $f_E(\cdot)$ in (5),

$$E_0^{''}(\rho) \le \sum_x p_x^\rho(x)(\ln p_x^\rho(x))^2 = f_0(p_x^\rho)$$

Lemma 1 tells us that $E_0^{''}(\rho) \le \alpha$, where

$$\alpha = \begin{cases} (\ln|\mathcal{X}|)^2 & \text{if } |\mathcal{X}| \ge 3 \\ \ln 2 & \text{if } |\mathcal{X}| = 2 \end{cases} \quad (12)$$

Here we replace the T from Lemma 1 with a looser upper bound $\ln 2$. Since $(\ln|\mathcal{X}|)^2 > \ln|\mathcal{X}|$ for $|\mathcal{X}| \ge 3$, we have $R - H(p_x) \le \alpha$, $\forall R \in [H(p_x), \ln|\mathcal{X}|)$. Combining (10) and (12), for the case without side-information the theorem is proved.[3] ■

[2] This is a loose analysis. For $|\mathcal{X}| \ge 3$, the upper bound on the first term is achieved when p_x^ρ is uniform on $\mathcal{X}$, giving the maximum at $(\ln|\mathcal{X}|)^2$ as shown in (12). The actual value of (11) is 0 for the uniform distribution.

C. Proof in General

The general proof is parallel:

Proof: Once again, we expand $\bar{E}_0(\rho)$ and basic calculus as shown in the appendix of [3], reveals that

$$\bar{E}_0^{'}(\rho) = \frac{d\bar{E}_0(\rho)}{d\rho} = H(\bar{p}_{x|y}^\rho) \quad (13)$$

and hence $\bar{E}_0^{'}(0) = H(p_{x|y})$. Thus:

$$\bar{E}_0(\rho) \le \rho H(p_{x|y}) + \frac{\rho^2}{2}\alpha \quad (14)$$

where $\alpha > 0$ is any upper bound of $\bar{E}_0^{''}(\rho_1)$ that holds $\forall \rho_1 \in [0,1]$. Substituting as before shows that

$$\begin{aligned}
E_r(R) &\ge \max_{0\le\rho\le1} \rho R - \rho H(p_{x|y}) - \frac{\rho^2}{2}\alpha \\
&= \frac{(R - H(p_{x|y}))^2}{2\alpha} \quad (15)
\end{aligned}$$

for $R - H(p_{x|y}) \le \alpha$. To find α, an upper bound on $\bar{E}_0^{''}(\rho)$, $\forall \rho \in [0,1]$, we expand $\bar{E}_0^{''}(\rho) = \frac{dH(p_{x|y}^\rho)}{d\rho}$

$$\begin{aligned}
&= \frac{d}{d\rho} \sum_{y\in\mathcal{Y}} \bar{p}_y^\rho(y) H(\bar{p}_{x|y=y}^\rho) \\
&= \sum_{y\in\mathcal{Y}} \bar{p}_y^\rho(y) \frac{dH(\bar{p}_{x|y=y}^\rho)}{d\rho} + \sum_{y\in\mathcal{Y}} \frac{d\bar{p}_y^\rho(y)}{d\rho} H(\bar{p}_{x|y=y}^\rho)
\end{aligned} \quad (16)$$

By basic calculus[4], we have:

$$\begin{aligned}
\frac{dH(\bar{p}_{x|y=y}^\rho)}{d\rho} &= \frac{1}{1+\rho} \sum_x \bar{p}_{x|y}^\rho(x|y)(\ln \bar{p}_{x|y}^\rho(x|y))^2 \\
&\quad - \frac{1}{1+\rho} \big(H(\bar{p}_{x|y=y}^\rho) \big)^2 \quad (17)
\end{aligned}$$

and,

$$\begin{aligned}
&\sum_{y\in\mathcal{Y}} \frac{d\bar{p}_y^\rho(y)}{d\rho} H(\bar{p}_{x|y=y}^\rho) \\
&\quad = \sum_y \bar{p}_y^\rho(y) H(\bar{p}_{x|y=y}^\rho)^2 - H(\bar{p}_{x|y}^\rho)^2 \quad (18)
\end{aligned}$$

[3] Although the upper bound on $E_0^{''}(\rho)$ is not tight as we drop the negative term in (11), it has the right order on $|\mathcal{X}|$. For a distribution $p = \{\frac{1}{2}, \frac{1}{2(|\mathcal{X}|-1)}, \ldots, \frac{1}{2(|\mathcal{X}|-1)}\}$, the evaluation of (11) is $\sim \frac{1}{4}(\ln|\mathcal{X}|)^2$ for large $|\mathcal{X}|$, thus the upper bound in (12) of $(\ln|\mathcal{X}|)^2$ has the right order.

[4] The tedious details of the derivation are in the proofs of Lemma 10 and Lemma 11, in the appendix of [3].

Substituting (17) and (18) in (16), we have $\bar{E}_0^{''}(\rho)$

$$= \frac{1}{1+\rho}\sum_y \bar{p}_y^\rho(y)[\sum_x \bar{p}_{x|y}^\rho(x|y)(\ln \bar{p}_{x|y}^\rho(x|y))^2]$$
$$-\frac{1}{1+\rho}\sum_y \bar{p}_y^\rho(y)H(\bar{p}_{x|y=y}^\rho)^2$$
$$+\sum_y \bar{p}_y^\rho(y)H(\bar{p}_{x|y=y}^\rho)^2 - H(\bar{p}_{x|y}^\rho)^2$$
$$= \frac{1}{1+\rho}\sum_y \bar{p}_y^\rho(y)[\sum_x \bar{p}_{x|y}^\rho(x|y)(\ln \bar{p}_{x|y}^\rho(x|y))^2]$$
$$+\frac{\rho}{1+\rho}\sum_y \bar{p}_y^\rho(y)H(\bar{p}_{x|y=y}^\rho)^2 - H(\bar{p}_{x|y}^\rho)^2 \quad (19)$$

Since $\sum_x \bar{p}_{x|y}^\rho(x|y) = 1$ for any $y \in \mathcal{Y}$, Lemma 1 tells us,

$$\sum_x \bar{p}_{x|y}^\rho(x|y)(\ln \bar{p}_{x|y}^\rho(x|y))^2 \le \alpha \quad (20)$$

$$\text{where } \alpha = \begin{cases} (\ln|\mathcal{X}|)^2 & \text{if } |\mathcal{X}| \ge 3 \\ \ln 2 & \text{if } |\mathcal{X}| = 2 \end{cases}$$

It is clear that:

$$H(\bar{p}_{x|y=y}^\rho)^2 \le (\ln|\mathcal{X}|)^2 \le \alpha, \quad \forall y \quad (21)$$

Substituting (20) and (21) in (19) and dropping the last term in (19) which is negative, we have

$$\bar{E}_0^{''}(\rho) \le \frac{1}{1+\rho}\sum_y \bar{p}_y^\rho(y)\alpha + \frac{\rho}{1+\rho}\sum_y \bar{p}_y^\rho(y)\alpha = \alpha \quad (22)$$

Since $(\ln|\mathcal{X}|)^2 > \ln|\mathcal{X}|$ for $|\mathcal{X}| \ge 3$, we have $R - H(p_{x|y}) \le \alpha$, $\forall R \in [H(p_{x|y}), \ln|\mathcal{X}|)$. Combining (15) and (22), the general theorem is proved. ■

IV. Conclusions and Future Work

In this paper we have derived a universal lower bound to random source coding error exponents. This bound has the quadratic form $a(R-h)^2$, where a, determining the shape of the quadratic function, is determined by the size of the source alphabet, and $R-h$ is the excess rate beyond the relevant entropy. It quantifies the intuitive idea that driving the probability of error to zero comes at the cost of either greater rate or longer block-lengths. These results are the source coding counterparts to the quadratic bounds on channel coding error exponents in Exercise 5.23 of [7], which can also be tightened slightly by using Lemma 1 as shown in [4]. Interestingly, the side-information alphabet size plays no role in the bound.

Numerical investigation reveals that this bound is loose and so it remains an open problem to see if it can be tightened while still maintaining an easy closed-form expression. This will involve solving the non-concave maximization problem in (11) exactly instead of dropping the negative term. We also suspect that similar universal bounds exist for all sorts of error exponents. It would be interesting to find a unified treatment that could also give a universal bound on the error exponent for lossy source coding investigated in [8].

Appendix

A. Proof of Lemma 1

Proof: We prove Lemma 1 by solving the following maximization problem for $f_E(\vec{\omega})$ with constraint $\vec{\omega} \in \mathcal{S}_J$.

$$\max_{\vec{\omega}\in\mathcal{S}_J} f_E(\vec{\omega}) = \max_{\vec{\omega}\in\mathcal{S}_J}\sum_{j=1}^J \omega_j(\ln\omega_j - E)^2$$

We have one equality constraint $\sum_{j=1}^J \omega_j = 1$ and J inequality constraints, $\omega_j \ge 0$, $\forall j = 1, 2, ..., J$, for the maximization problem. Note that $f_E(\vec{\omega})$ is a bounded differentiable function and $\mathcal{S}_J$ is a compact set in $\mathcal{R}^J$. Thus, there exists a point in $\mathcal{S}_J$, to maximize it. We examine the necessary conditions for a point $\vec{\omega}^*$, in $\mathcal{S}_J$, to maximize $f_E(\vec{\omega})$. By the Karush-Kuhn-Tucker necessary conditions [2], there exist $\gamma_j \ge 0$, $j = 1, 2, ..., J$ and $\lambda \ge 0$, s.t.

$$\nabla f_E(\vec{\omega}^*) + \sum_{j=1}^J \gamma_j \nabla\omega_j^* + \lambda\nabla\sum_{j=1}^J \omega_j^* = 0$$

$$\gamma_j\omega_j^* = 0, \quad \forall j = 1, 2, ..., J; \text{ and } \sum_{j=1}^J \omega_j^* = 1$$

That is,

$$(\ln\omega_j^*)^2 + 2(1-E)\ln\omega_j^* + \gamma_j + \lambda - 2E = 0$$

$$\gamma_j\omega_j^* = 0, \quad \forall j = 1, 2, ..., J; \text{ and } \sum_{j=1}^J \omega_j^* = 1$$

Note that

$$\frac{\partial f_E(\vec{\omega})}{\partial\omega_j}|_{\omega_j=0} = (\ln\omega_j)^2 + 2(1-E)\ln\omega_j|_{\omega_j=0} = \infty$$
$$\frac{\partial f_E(\vec{\omega})}{\partial\omega_j}|_{\omega_j\neq0} = (\ln\omega_j)^2 + 2(1-E)\ln\omega_j|_{\omega_j\neq0} < \infty$$

and thus to maximize $f_E(\vec{\omega})$, the ω_j^* are strictly positive. Hence $\gamma_j = 0$,

$$(\ln\omega_j^*)^2 + 2(1-E)\ln\omega_j^* + \lambda - 2E = 0, \quad \forall j$$

$$\text{and } \sum_{j=1}^J \omega_j^* = 1$$

Since $\ln\omega_j^*$ is a root of a quadratic equation $x^2 + 2(1-E)x + \lambda - 2E = 0$, this implies ω_j^* can only be either

$t_1 = e^{\beta_1}$ or $t_2 = e^{\beta_2}$, where β_1 and β_2 are the two roots of the quadratic equation. Because $\beta_1 + \beta_2 = -2 + 2E$, we know $t_1 t_2 = e^{-2+2E}$. Now, either all the ω_j^*'s have the same value $1/J$ and we are essentially done, or there are K of ω_j^*'s are t_1, $J-K$ of ω_j^*'s are t_2, where K is an integer and $0 < K < J$. In that case, we would have the following equations:

$$t_1 t_2 = e^{-2+2E}$$
$$K t_1 + (J-K) t_2 = 1$$

Solve for t_2 using the first equation and substitute into the linear one:

$$\frac{K}{J-K} t_1^2 - \frac{1}{J-K} t_1 + e^{-2+2E} = 0$$

This quadratic equation has real roots if and only if the following is true[5]

$$\left(\frac{1}{J-K}\right)^2 - \frac{4K e^{-2+2E}}{J-K} \geq 0$$

Because of the assumption that $0 < K < J$, this can be simplified to

$$\frac{e^{2-2E}}{4} \geq K(J-K)$$

Since $E \geq 0$, we have,

$$1.847 \approx \frac{e^2}{4} \geq \frac{e^{2-2E}}{4} \geq K(J-K)$$

With $0 < K < J$ and both K and J being positive integers, the above condition can only possibly be satisfied if $J = 2$, $K = 1$. Otherwise for all $J \geq 3$, all the ω_j^* have to be the same to optimize the desired function. Since $\sum_{j=1}^{J} \omega_j^* = 1$, we know that the optimal point is $\omega_j^* = 1/J$, $\forall j$. Substituting into $f_E(\vec{\omega})$, we get $\forall J \geq 3$,

$$\begin{aligned} f_E(\vec{\omega}) &\leq f_E(\vec{\omega}^*) \\ &= \sum_{j=1}^{J} \frac{1}{J}\left(\ln \frac{1}{J} - E\right)^2 \\ &= E^2 + 2E(\ln J) + (\ln J)^2 \end{aligned} \tag{23}$$

For $J = 2$, $K = 1$, the quadratic equation:

$$t^2 - t + e^{-2+2E} = 0 \tag{24}$$

has real-value roots if and only if $\frac{e^{2-2E}}{4} \geq 1$, or $E \leq \frac{1}{2} \ln \frac{e^2}{4} \approx 0.307$. For this case:

$$\begin{aligned} f_E(\vec{\omega}) &= \sum_{j=1}^{2} \omega_j (\ln \omega_j - E)^2 \\ &= \sum_{j=1}^{2} \omega_j (\ln \omega_j)^2 + 2H(\vec{\omega})E + E^2 \\ &\leq \max_{\vec{\omega} \in \mathcal{S}_2} \sum_{j=1}^{2} \omega_j (\ln \omega_j - 0)^2 + 2(\ln 2)E + E^2 \\ &\leq t_1 (\ln t_1)^2 + t_2 (\ln t_2)^2 + 2(\ln 2)E + E^2 \end{aligned}$$

Where t_1 and t_2 are the two different roots of (24) with $E = 0$: $t^2 - t + e^{-2} = 0$. Hence for $E \leq \frac{1}{2} \ln \frac{e^2}{4}$,

$$f_E(\vec{\omega}) \leq T + 2(\ln 2)E + E^2 \tag{25}$$

where T is defined in (7). Numerically, $T \approx 0.563 > (\ln 2)^2 \approx 0.480$.

For $E > \frac{1}{2} \ln \frac{e^2}{4}$, the optimal point is at $(0.5, 0.5)$:

$$f_E(\vec{\omega}) \leq (\ln 2)^2 + 2(\ln 2)E + E^2 \tag{26}$$

Combining (25) and (26), and because $T > (\ln 2)^2$, we conclude that, for all $E \geq 0$,

$$f_E(\vec{\omega}) \leq T + 2(\ln 2)E + E^2 \tag{27}$$

With (23) and (27) we prove the lemma. □

[5]Note that a quadratic equation $ax^2 + bx + c = 0$ has real roots if and only if $b^2 - 4ac \geq 0$

REFERENCES

[1] D. Baron, M. A. Khojastepour, and R. G. Baraniuk. Redundancy rates of Slepian-Wolf coding. *42nd Allerton Conference*, 2004.

[2] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[3] Cheng Chang, Stark Draper, and Anant Sahai. Lossless coding for distributed streaming sources. *IEEE Trans. Inform. Theory*, submitted.

[4] Cheng Chang and Anant Sahai. Solution to gallager's exercise 5.23. *http://www.eecs.berkeley.edu/ cchang/5_23.pdf*, 2006.

[5] Imre Csiszár and János Körner. *Information Theory*. Akadémiai Kiadó, Budapest, 1986.

[6] Robert Gallager. Source coding with side information and universal coding. *LIDS-P-937*, 1979.

[7] Robert G. Gallager. *Information Theory and Reliable Communication*. John Wiley, New York, NY, 1971.

[8] Katalin Marton. Error exponent for source coding with a fidelity criterion. *IEEE Transactions on Information Theory*, 20:197–199, 1974.

[9] S. Sarvotham, D. Baron, and R. G. Baraniuk. Non-asymptotic performance of symmetric Slepian-Wolf coding. *39th CISS*, 2005.

[10] D. Slepian and J.K. Wolf. Noiseless coding of correlated information sources. *IEEE Transactions on Information Theory*, 19, 1973.