Source Coding with Side Information: A Dual-Domain Expurgated Error Exponent

Mehdi Dabirnia
CTTC
mehdi.dabirnia@cttc.es

Hamdi Joudeh
Eindhoven University of Technology
h.joudeh@tue.nl

Albert Guillén i Fàbregas
University of Cambridge
Universitat Politècnica de Catalunya
guillen@ieee.org

***Abstract*—We introduce an expurgation technique for source coding with side information that demonstrates existence of a code in the expurgated ensemble for which every source sequence satisfies a desired upper bound on error probability. Applying the developed technique, directly in dual-domain, we derive an expurgated error exponent for standard random coding ensemble with a possibly mismatched decoding metric.**

I. Introduction

Consider a pair of discrete memoryless correlated sources with finite alphabets $\mathcal{X}$ and $\mathcal{Y}$, and joint distribution P_{XY}. A standard block source code $\mathcal{C}(n, R)$ is defined as a mapping ϕ of source sequences $\boldsymbol{x} = x_1, x_2, \ldots, x_n \in \mathcal{X}^n$ to a set of codewords $\mathcal{M} = \{1, \ldots, M\}$ where $R = \frac{\log M}{n}$ is the code rate. At the receiver end, the decoder ψ maps each codeword $m \in \mathcal{M}$ and side information sequence $\boldsymbol{y} = y_1, y_2, \ldots, y_n \in \mathcal{Y}^n$ back into a source sequence $\hat{\boldsymbol{x}}$. The decoder makes an error whenever $\psi(\phi(\boldsymbol{x}), \boldsymbol{y}) \neq \boldsymbol{x}$ and the probability of error is given by

$$p_e = \mathbb{P}[\psi(\phi(\boldsymbol{X}), \boldsymbol{Y}) \neq \boldsymbol{X}]. \tag{1}$$

This setting is known as block source coding with decoder side information, or Slepian-Wolf coding [1], [2]. A dual-domain achievable random coding error exponent for the above setting with maximum a posteriori probability (MAP) decoding was derived by Gallager in [2] as

$$E_{\mathrm{r}}(R) = \max_{\rho \in [0,1]} \rho R - E_0(\rho), \tag{2}$$

where the function $E_0(\rho)$ is defined as

$$E_0(\rho) \triangleq \log \sum_{y \in \mathcal{Y}} P_Y(y) \bigg(\sum_{x \in \mathcal{X}} P_{X|Y}(x|y)^{\frac{1}{1+\rho}} \bigg)^{1+\rho}. \tag{3}$$

Gallager [2] also derived a corresponding upper bound $E_{\mathrm{sp}}(R)$ by providing the side information sequence to the encoder, which takes a similar form as (2) with optimization over $\rho \geq 0$.

Relating the Slepian-Wolf source coding problem to a counterpart channel coding problem, a tighter achievable error exponent for high code rate regime (referred to as the *expurgated exponent*) was derived by Csiszár and Körner in [3] for a generic (possibly mismatched) decoder. For the case of using optimum ML decoder this exponent is given by

$$E_{\mathrm{ex}}(R) = \min_{P_{\tilde{X}}} \bigg\{ D(P_{\tilde{X}} \| P_X) + \min_{P_{\hat{X}\tilde{X}} : P_{\hat{X}} = P_{\tilde{X}}, H(\hat{X}|\tilde{X}) \geq R} \Big\{ \mathbb{E}\big[d(\hat{X}, \tilde{X})\big] + R - H(\hat{X}|\tilde{X}) \Big\} \bigg\} \tag{4}$$

where $d(\hat{x}, \tilde{x})$ is the Bhattacharyya distance, defined as

$$d(\hat{x}, \tilde{x}) \triangleq -\log \sum_{y \in \mathcal{Y}} \sqrt{P_{Y|X}(y|\hat{x}) P_{Y|X}(y|\tilde{x})}. \tag{5}$$

Exploiting the connection to counterpart channel coding problem and using permutation codes, the exponent in (4) was also derived later by Ahlswede and Dueck in [4]. Both proofs in [3], [4] rely on type analysis and combinatorial arguments and do not use random coding or expurgation.

In this work we develop an expurgation technique for source coding with side information, and then employ it to directly derive a dual-domain expurgated exponent under generic, and possibly mismatched, decoding metrics. Specifically we consider a maximum metric decoder of the form

$$\psi(m, \boldsymbol{y}) = \underset{\boldsymbol{x} \in \mathcal{X}^n : \phi(\boldsymbol{x}) = m}{\arg\max} \; q(\boldsymbol{x}, \boldsymbol{y}), \tag{6}$$

where q is an arbitrary non-negative decoding metric.

II. Main result

We develop an expurgation method for source coding that is valid for any source and side information model with arbitrary decoding metric. The method is based on repeated expurgation, starting with all sequences in $\mathcal{X}^n$, in each repetition we show existence of a code in the random coding ensemble for remaining source sequences, such that half of those meet a desired upper bound on the error probability. In order to exhaust all sequences, the expurgation process is repeated for at most $n \log_2 |\mathcal{X}|$ times. Combining the selected codes from all repetitions, we obtain a code in which all the source sequences satisfy the desired upper bound as given by the following lemma. We first define the ensemble.

This work was supported in part by the European Research Council under Grants 101116550, 101142747 and 101158232, and in part by the Spanish Government under Grants PID2020-116683GB-C22 and PID2021-128373OB-I00 and partially by the WAVE-XL project (PID2024-160457OB-I00) funded by MCIN/AEI/10.13039/501100011033 and by FEDER, UE.

Definition 1. *The standard random coding ensemble is the set of all (n, R) standard block codes for source alphabet $\mathcal{X}$ where each source sequence is mapped independently and with equal probability $\frac{1}{M}$ into one of the M codewords.*

Lemma 1. *There exists a code $\mathcal{C}_{\text{ex}}$ in the standard random coding ensemble such that for every source sequence $\boldsymbol{x} \in \mathcal{X}^n$ and $\rho \geq 0$*

$$p_e(\boldsymbol{x}, \mathcal{C}_{\text{ex}}) \leq \left(2\mathbb{E}\left[p_e(\boldsymbol{x}, \mathsf{C})^{\frac{1}{\rho}}\right]\right)^{\rho}, \tag{7}$$

where the expectation is over the standard random coding ensemble with $\frac{M}{k}$ codewords with $k = n \log_2 |\mathcal{X}|$.

Considering the standard random coding ensemble and applying Lemma 1 we obtain the following achievable error exponent for memoryless sources employing a memoryless decoding metric $q(\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^{n} q(x_i, y_i)$.

Theorem 1. *For every $R > 0$ and every distribution $P_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ there exists a standard block source code with maximum metric decoder* (6) *employing decoding metric $q(x, y)$ that achieves the exponent*

$$E_q(R) = \max\{E_{q,\text{r}}(R), E_{q,\text{ex}}(R)\}, \tag{8}$$

where

$$E_{q,\text{r}}(R) = \sup_{\substack{\rho \in [0,1] \\ s \geq 0}} \rho R - E_{\text{s}}(\rho, s), \tag{9}$$

$$E_{q,\text{ex}}(R) = \sup_{\substack{\rho \geq 1 \\ s \geq 0}} \rho R - E_{\text{x}}(\rho, s), \tag{10}$$

with

$$E_{\text{s}}(\rho, s) = \log \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P_{XY}(x, y) \left(\sum_{\bar{x} \in \mathcal{X}} \left(\frac{q(\bar{x}, y)}{q(x, y)} \right)^{s} \right)^{\rho}, \tag{11}$$

$$E_{\text{x}}(\rho, s) = \log \sum_{x \in \mathcal{X}} \left(\sum_{\bar{x} \in \mathcal{X}} \left(\sum_{y \in \mathcal{Y}} P_{XY}(x, y) \left(\frac{q(\bar{x}, y)}{q(x, y)} \right)^{s} \right)^{\frac{1}{\rho}} \right)^{\rho}. \tag{12}$$

Proofs of Lemma 1 and Theorem 1 can be found in [5]. Noticing that exponent is a convex function of the rate and the maximizing ρ is the slope of the exponent curve, similar to [6, Ch. 5] by evaluating the partial derivative of the exponent to find the rate at which $\rho = 0$ maximizes the exponent, we obtain the following rate achieved by standard random coding:

$$H_q(X|Y) = \inf_{s \geq 0} - \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P_{XY}(x, y) \log \frac{q(x, y)^s}{\sum_{\bar{x} \in \mathcal{X}} q(\bar{x}, y)^s} \tag{13}$$

$$= H(X|Y) + \inf_{s \geq 0} D(P_{X|Y} \| Q_{X|Y}^{(s)}), \tag{14}$$

where $Q_{X|Y}^{(s)}(x|y) = \frac{q(x,y)^s}{\sum_{\bar{x} \in \mathcal{X}} q(\bar{x},y)^s}$.

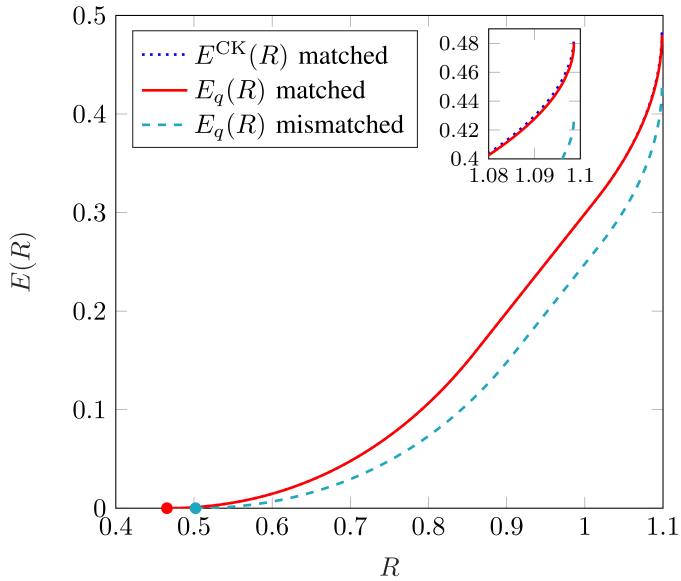

Fig. 1: Error exponents for the source X and side information Y with joint distribution given in (15). The mismatched decoder uses the minimum Hamming distance metric in (16).

III. Example

The joint distribution of the source X with side information Y is defined by the entries of the $|\mathcal{X}| \times |\mathcal{Y}|$ matrix

$$P_{XY} = \begin{bmatrix} 0.49 & 0.005 & 0.005 \\ 0.015 & 0.27 & 0.015 \\ 0.05 & 0.05 & 0.1 \end{bmatrix}. \tag{15}$$

We consider using a mismatched decoder with a memoryless metric given by the matrix

$$q(x, y) = \begin{bmatrix} 1 - 2\delta & \delta & \delta \\ \delta & 1 - 2\delta & \delta \\ \delta & \delta & 1 - 2\delta \end{bmatrix} \tag{16}$$

with $\delta \in (0, \frac{1}{3})$. This is equivalent to a minimum Hamming distance decoding metric.

Figure 1 illustrates the exponents for the standard ensemble with both matched (MAP) and mismatched decoders and Csiszár and Körner's exponent E^{CK} from [3] with a matched decoder. We observe that the derived expurgated exponent in (8) is higher for the matched decoding and it is weaker than the exponent of [3] for high rates.

References

[1] D. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inf. Theory*, vol. 19, no. 4, pp. 471–480, 1973.

[2] R. G. Gallager, "Source coding with side information and universal coding," M.I.T. Technical Report LIDS-P-937, Tech. Rep., 1979.

[3] I. Csiszár and J. Körner, "Graph decomposition: A new key to coding theorems," *IEEE Trans. Inf. Theory*, vol. 27, no. 1, pp. 5–12, Jan. 1981.

[4] R. Ahlswede and G. Dueck, "Good codes can be produced by a few permutations," *IEEE Trans. Inf. Theory*, vol. 28, no. 3, pp. 430–443, 1982.

[5] M. Dabirnia, H. Joudeh, and A. Guillén i Fàbregas, "Dual domain expurgated error exponents for source coding with side information," Aug. 2025. [Online]. Available: http://arxiv.org/abs/2508.03467

[6] R. G. Gallager, *Information Theory and Reliable Communication*. John Wiley & Sons, Inc. New York, NY, USA, 1968.