A Recursive Cost-Constrained Construction that Attains the Expurgated Exponent

Anelia Somekh-Baruch
Bar-Ilan University
somekha@biu.ac.il

Jonathan Scarlett
National University of Singapore
scarlett@comp.nus.edu.sg

Albert Guillén i Fàbregas
ICREA & Universitat Pompeu Fabra
University of Cambridge
guillen@ieee.org

***Abstract*—We show that a recursive cost-constrained random coding scheme attains an error exponent that is at least as high as both the random-coding exponent and the expurgated exponent. The random coding scheme enforces that every pair of codewords in the codebook meets a minimum distance condition, and is reminiscent of the Gilbert-Varshamov construction, but with the notable feature of permitting continuous-alphabet channels. The distance function is initially arbitrary, and it is shown that the Chernoff/Bhattacharyya distance suffices to attain the random coding and expurgated exponents.**

I. Problem Setup

We consider transmission over a memoryless channel described by a conditional probability rule $W(y|x)$, with input $x \in \mathcal{X}$ and output $y \in \mathcal{Y}$ for arbitrary alphabets $\mathcal{X}$ and $\mathcal{Y}$; in particular, $W(y|x)$ is a conditional probability mass function (PMF) in the discrete case, and a conditional probability density function (PDF) in the continuous case. We define $W^n(\boldsymbol{y}|\boldsymbol{x}) = \prod_{k=1}^n W(y_k|x_k)$ for input/output sequences $\boldsymbol{x} = (x_1, \dots, x_n) \in \mathcal{X}^n$ and $\boldsymbol{y} = (y_1, \dots, y_n) \in \mathcal{Y}^n$. The corresponding random variables are denoted by $\boldsymbol{X}$ and $\boldsymbol{Y}$.

An encoder maps a message $m \in \{1, \dots, M_n\}$ to a channel input sequence $\boldsymbol{x}_m \in \mathcal{X}$, where the number of messages is denoted by M_n. The message, represented by the random variable S, is assumed to take values on $\{1, \dots, M_n\}$ equiprobably. This mapping induces an (n, M_n)-codebook $\mathcal{M}_n = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_{M_n}\}$ with rate $R_n = \frac{1}{n} \log M_n$.

Upon observing the channel output $\boldsymbol{y}$, the decoder produces an estimate of the transmitted message $\hat{m} \in \{1, \dots, M_n\}$. We consider the family of type-dependent maximum-metric decoders, for which the transmitted message is estimated as

$$\hat{m} = \operatorname*{arg\,max}_{\boldsymbol{x}_i \in \mathcal{M}_n} q(\boldsymbol{x}_i, \boldsymbol{y}), \tag{1}$$

where $q(\boldsymbol{x}_i, \boldsymbol{y})$ is a certain decoding metric depending only on the joint type [1, Ch. 2] of $(\boldsymbol{x}_i, \boldsymbol{y})$. Maximum-likelihood (ML) decoding is a special case of (1), but more generally, the decoder may be *mismatched* [2]–[4].

Denoting the random variable corresponding to the decoded message by $\hat{S}$, we define the probability of error as $P_e = \Pr\big[\hat{S} \neq S\big]$. A rate-exponent pair (R, E) is said to be achievable for channel W if, for all $\epsilon > 0$, there exists a sequence of $(n, \mathrm{e}^{n(R-\epsilon)})$-codebooks such that

$$\liminf_{n \to \infty} -\frac{1}{n} \log \Pr\big[\hat{S} \neq S\big] \geq E - \epsilon. \tag{2}$$

Equivalently, we say that E is an achievable error exponent at rate R if (R, E) is an achievable rate-exponent pair.

In the case of discrete memoryless channels (i.e., $\mathcal{X}$ and $\mathcal{Y}$ are finite), using random selection and graph decomposition techniques, Csiszár and Körner [5] studied the error exponents of constant-composition codes under a decoder that uses a type-dependent decoding metric, and derived the following achievable exponent for an arbitrary input distribution P:

$$E_q(R, P, W) = \min_{V \in \mathcal{T}_I} D(V_{Y|X} \| W | P) + \big| I(\widetilde{X}; Y, X) - R \big|_+, \tag{3}$$

where

$$\mathcal{T}_I \triangleq \Big\{ V_{X\widetilde{X}Y} \in \mathcal{P}(\mathcal{X} \times \mathcal{X} \times \mathcal{Y}) : V_X = V_{\widetilde{X}} = P, \; q(V_{\widetilde{X}Y}) \geq q(V_{XY}), I(X; \widetilde{X}) \leq R \Big\} \tag{4}$$

with $\mathcal{P}(\cdot)$ denoting the set of probability distributions on the given alphabet. This error exponent was shown to be at least as high as both the expurgated and random coding exponents.

In this paper, we consider a generalization of (4) that we recently gave in [6] (outlined below), with our main goal being to provide an equivalent form and an alternative proof that extends immediately to continuous-alphabet channels.

II. Discrete Memoryless Channels: Exponent and Lagrange Duality

In this section, we restrict ourselves to discrete memoryless channels (DMCs). We overview a recursive construction that we recently proposed [6] for attaining the Csiszár-Körner exponent (3) based on analogies with the Gilbert-Varshamov construction in Hamming spaces [7], [8].

The set of sequences with type P_n is denoted by $\mathcal{T}(P_n)$. For $i < j$, we let $\boldsymbol{x}_i^j$ denote $(\boldsymbol{x}_i, \dots, \boldsymbol{x}_j)$. Fixing n, M_n, an input distribution $P \in \mathcal{P}(\mathcal{X})$, a symmetric type-dependent "distance" function $d(\cdot, \cdot)$,[1] and constants $\delta > 0, \Delta \in \mathbb{R}$, the construction is described by the following steps:

This work was supported in part by the Israel Science Foundation under grant 631/17, the European Research Council under Grant 725411, by the Spanish Ministry of Economy and Competitiveness under Grant TEC2016-78434-C3-1-R, and by an NUS Early Career Research Award.

[1] We use the terminology *distance* even though d need not be a distance function in the topological sense.

1) The first codeword, $\boldsymbol{x}_1$, is drawn uniformly from $\mathcal{T}(P_n)$;
2) The second codeword $\boldsymbol{x}_2$ is drawn uniformly from

$$\mathcal{T}(P_n, \boldsymbol{x}_1) \triangleq \{\bar{\boldsymbol{x}} \in \mathcal{T}(P_n) \,:\, d(\bar{\boldsymbol{x}}, \boldsymbol{x}_1) > \Delta\} \tag{5}$$
$$= \mathcal{T}(P_n) \backslash \{\bar{\boldsymbol{x}} \in \mathcal{T}(P_n) \,:\, d(\bar{\boldsymbol{x}}, \boldsymbol{x}_1) \le \Delta\}, \tag{6}$$

the set of sequences of composition P_n whose distance to $\boldsymbol{x}_1$ exceeds Δ;
3) The i-th codeword $\boldsymbol{x}_i$ is drawn uniformly from

$$\mathcal{T}(P_n, \boldsymbol{x}_1^{i-1})$$
$$= \{\bar{\boldsymbol{x}} \in \mathcal{T}(P_n) \,:\, d(\bar{\boldsymbol{x}}, \boldsymbol{x}_j) > \Delta,\, j = 1 \ldots, i-1\} \tag{7}$$
$$= \mathcal{T}(P_n, \boldsymbol{x}_1^{i-2}) \backslash \left\{\bar{\boldsymbol{x}} \in \mathcal{T}(P_n, \boldsymbol{x}_1^{i-2}) : d(\bar{\boldsymbol{x}}, \boldsymbol{x}_{i-1}) \le \Delta\right\}. \tag{8}$$

In order to ensure that the above procedure generates the desired number of codewords $M_n = \mathrm{e}^{nR_n}$ (i.e., the sets $\mathcal{T}(P_n, \boldsymbol{x}_1^{i-1})$ are non-empty for all $i = 1, \ldots, M_n$), it suffices to choose Δ and δ such that [6]

$$\mathrm{e}^{n(R_n+\delta)} \mathrm{vol}_{\boldsymbol{x}}(\Delta) \le |\mathcal{T}(P_n)| \tag{9}$$

where $\mathrm{vol}_{\boldsymbol{x}}(\Delta) = |\{\bar{\boldsymbol{x}} \in \mathcal{T}(P_n) \,:\, d(\bar{\boldsymbol{x}}, \boldsymbol{x}) \le \Delta\}|$ is the volume of a ball of radius Δ according to distance $d(\cdot,\cdot)$ centered at some $\boldsymbol{x} \in \mathcal{T}(P_n)$. Since $d \in \Omega$ is symmetric and type-dependent, $\mathrm{vol}_{\boldsymbol{x}}(\Delta)$ does not depend on the specific choice of $\boldsymbol{x} \in \mathcal{T}(P_n)$.

Let $d(P_{X\widetilde{X}})$ be a shorthand for $d(\boldsymbol{x}.\widetilde{\boldsymbol{x}})$ when $(\boldsymbol{x}.\widetilde{\boldsymbol{x}}) \in \mathcal{T}(P_{X\widetilde{X}})$. It was shown in [6] that under the rate condition

$$R \le \min_{P_{X\widetilde{X}} \,:\, d(P_{X\widetilde{X}}) \le \Delta,\, P_X = P_{\widetilde{X}} = P} I(X; \widetilde{X}) - 2\delta \tag{10}$$

(which ensures that (9) holds), the ensemble average error probability of the above recursive random code construction attains the following exponent:

$$E_{\mathrm{RGV}}(R, P, W, q, d, \Delta)$$
$$= \min_{V_{X\widetilde{X}Y} \in \mathcal{T}_{\mathrm{RGV}}} D(V_{Y|X} \| W | P) + \left| I(\widetilde{X}; Y, X) - R \right|_+, \tag{11}$$

where

$$\mathcal{T}_{\mathrm{RGV}} \triangleq \Big\{ V_{X\widetilde{X}Y} \in \mathcal{P}(\mathcal{X} \times \mathcal{X} \times \mathcal{Y}) \,:\, V_X = V_{\widetilde{X}} = P,$$
$$q(V_{\widetilde{X}Y}) \ge q(V_{XY}),\, d(P_{X\widetilde{X}}) \ge \Delta \Big\}. \tag{12}$$

While we have stated the error exponent, the central part of the analysis is in arriving at the following asymptotic expression for the ensemble average probability of error (where $\doteq$ denotes equality to first order in the exponent and P_n is a type converging to P as $n \to \infty$):

$$\bar{P}_{\mathrm{e}}^{(n)} \doteq \sum_{\boldsymbol{x} \in \mathcal{T}(P_n), \boldsymbol{y}} \frac{1}{|\mathcal{T}(P_n)|} W^n(\boldsymbol{y}|\boldsymbol{x})$$
$$\cdot \min\left\{1, (M_n - 1) \sum_{\substack{\boldsymbol{x}' \in \mathcal{T}(P_n) \,:\, q^n(\boldsymbol{x}', \boldsymbol{y}) \ge q^n(\boldsymbol{x}, \boldsymbol{y}) \\ d(\boldsymbol{x}', \boldsymbol{x}) \ge \Delta}} \frac{1}{|\mathcal{T}(P_n)|} \right\}. \tag{13}$$

This can be interpreted as a stronger (albeit asymptotic) analog of the *random coding union* bound [9] that achieves not only the random coding exponent, but also the low-rate improvements of the expurgated exponent.

It is also shown in [6] that $E_{\mathrm{RGV}}(R, P, W, q, d, \Delta) = E_q(R, P, W)$ for certain choices of the distance function $d(\cdot,\cdot)$, notably including the negative empirical mutual information. Note that the above construction is only valid for finite input alphabets, and the analysis of [5] relies on both the input and output alphabets being finite.

Towards handling general alphabets, we begin by stating a Lagrange dual form of the RGV exponent and rate condition in (10)–(11). To facilitate this, we henceforth restrict our attention to *additive* distances and decoding metrics:

$$d(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{n} \sum_{k=1}^{n} d(x_k, x_k') \tag{14}$$

for some single-letter function $d(x, x')$, and similarly

$$q(\hat{P}_{\boldsymbol{x},\boldsymbol{y}}) = \frac{1}{n} \sum_{i=1}^{n} q(x_i, y_i), \tag{15}$$

for some single-letter function $q(x, y)$ (abusing notation slightly). While the additivity assumption is more restrictive than the assumption of being type-dependent, there are many interesting examples of additive distances and metrics, such as Hamming distance, Bhattacharyya distance, maximum-likelihood decoding, and (single-letter) mismatched decoding.

Theorem 1. *Under the preceding setup with an additive distance function d and additive decoding metric q, the error exponent* (11) *can be written as*

$$E_{\mathrm{RGV}}(R, P, W, q, d, \Delta) = \sup_{\rho \in [0,1]} E_0(\rho) - \rho R, \tag{16}$$

where

$$E_0(\rho) = \sup_{r \ge 0, s \ge 0, a(\cdot)} -\sum_x P(x)$$
$$\cdot \log \sum_y W(y|x) \left(\frac{\sum_{x'} P(x') e^{sq(x',y)} e^{a(x')} e^{r(d(x,x')-\Delta)}}{e^{sq(x,y)} e^{a(x)}} \right)^{\rho} \tag{17}$$

and rate condition (10) *can be written as*

$$R \le E_v(P, d, \Delta) - 2\delta, \tag{18}$$

where, defining $\phi_a = \mathbb{E}_P[a(X)]$, we have

$$E_v(P, d, \Delta) \triangleq \sup_{r \ge 0, a(\cdot)} \sum_x P(x)$$
$$\cdot \log \sum_{x'} P(x') e^{a(x') - \phi_a} e^{-r(d(x,x')-\Delta)}. \tag{19}$$

Proof outline: The proof uses Lagrange duality analogously to the corresponding statements for the random coding and expurgated exponents [10], [11]. Specifically, the Lagrange dual of the rate condition (10) yields (18), and the Lagrange dual of the exponent (11) yields (17). The optimization parameters r, s, and $a(\cdot)$ correspond to Lagrange multipliers.

For the latter of these, the exponent can alternatively be attained directly by bounding the inner term in (13) as

$$\sum_{\substack{\bar{\boldsymbol{x}}\in\mathcal{T}(P_n)\\ q(\bar{\boldsymbol{x}},\boldsymbol{y})\geq q(\boldsymbol{x}_m,\boldsymbol{y})\\ d(\bar{\boldsymbol{x}},\boldsymbol{x}_m)\geq\Delta}} \frac{1}{|\mathcal{T}(P_n)|}$$

$$= \sum_{\bar{\boldsymbol{x}}\in\mathcal{T}(P_n)} \frac{1}{|\mathcal{T}(P_n)|}\mathbb{1}\{q(\bar{\boldsymbol{x}},\boldsymbol{y})\geq q(\boldsymbol{x}_m,\boldsymbol{y})\}\mathbb{1}\{d(\bar{\boldsymbol{x}},\boldsymbol{x}_m)\geq\Delta\} \tag{20}$$

$$\leq \sum_{\bar{\boldsymbol{y}}\in\mathcal{T}(P_n)} \frac{1}{|\mathcal{T}(P_n)|}\frac{\mathrm{e}^{sq(\bar{\boldsymbol{x}},\boldsymbol{y})+a(\bar{\boldsymbol{x}})}}{\mathrm{e}^{sq(\boldsymbol{x}_m,\boldsymbol{y})+a(\boldsymbol{x})}}\mathrm{e}^{r(d(\bar{\boldsymbol{x}},\boldsymbol{x}_m)-\Delta)} \tag{21}$$

and then following the steps of Poltyrev [12]. ∎

The expression in (16) bears a strong resemblance to the mismatched random coding exponent for constant-composition coding [10]; in fact, the only difference is the presence of additional term $e^{r(d(x,x')-\Delta)}$. Indeed, in order to prove the achievability of the random coding exponent from (16), one simply sets $r=0$.

To show that (16) also achieves the expurgated exponent, we first claim that the achievability of (11) (and hence its equivalent expression (16)) for symmetric d implies the same for non-symmetric d; the proof is deferred to the full version due to space constraints [13, Cor. 1]. This fact allows us to set $\rho=1$ and choose the distance to be the Chernoff distance:

$$d(x,x')=d_s(x,x')=-\log\sum_y W(y|x)\left(\frac{\mathrm{e}^{q(x',y)}}{\mathrm{e}^{q(x,y)}}\right)^s, \tag{22}$$

which reduces to the Bhattacharyya distance in the matched case $q(x,y)=\log W(y|x)$ when $s=\frac{1}{2}$. The preceding choice $d=d_s$ yields for any s, r, and $a(\cdot)$ that

$$E_{\mathrm{RGV}}(R,P,W,q,d,\Delta)$$
$$\geq -\sum_x P(x)\log\sum_{x'}P(x')$$
$$\cdot\sum_y W(y|x)\frac{\mathrm{e}^{sq(x',y)+a(x')}}{\mathrm{e}^{sq(x,y)+a(x)}}e^{r(d_s(x,x')-\Delta)}-R \tag{23}$$

$$=-\sum_x P(x)\log\sum_{x'}P(x')\mathrm{e}^{-d_s(x,x')}\frac{\mathrm{e}^{a(x')}}{\mathrm{e}^{a(x)}}\mathrm{e}^{r(d_s(x,x')-\Delta)}-R. \tag{24}$$

Setting $r=\frac{\rho'}{1+\rho'}$ for some $\rho'\geq 0$ gives

$$E_{\mathrm{RGV}}(R,P,W,q,d,\Delta)\geq -\sum_x P(x)$$
$$\cdot\log\sum_{x'}P(x')\mathrm{e}^{-\frac{d_s(x,x')}{1+\rho'}}\frac{\mathrm{e}^{a(x')}}{\mathrm{e}^{a(x)}}+\Delta\frac{\rho'}{1+\rho'}-R. \tag{25}$$

Then, choosing

$$\Delta=-(1+\rho')\left(\sum_x P(x)\log\sum_{x'}P(x')\mathrm{e}^{-\frac{d_s(x,x')}{1+\rho'}}\frac{\mathrm{e}^{a(x')}}{\mathrm{e}^{a(x)}}+R+2\delta\right) \tag{26}$$

we obtain from (25) that

$$E_{\mathrm{RGV}}(R,P,W,q,d,\Delta)$$
$$\geq -\sum_x P(x)\log\sum_{x'}P(x')\mathrm{e}^{-\frac{d_s(x,x')}{1+\rho'}}\frac{\mathrm{e}^{a(x')}}{\mathrm{e}^{a(x)}}$$
$$-\rho'\left(\sum_x P(x)\log\sum_{x'}P(x')\mathrm{e}^{-\frac{d_s(x,x')}{1+\rho'}}\frac{\mathrm{e}^{a(x')}}{\mathrm{e}^{a(x)}}+R+2\delta\right)-R \tag{27}$$

$$=-(1+\rho')\left(\sum_x P(x)\log\sum_{x'}P(x')\mathrm{e}^{-\frac{d_s(x,x')}{1+\rho'}}\frac{\mathrm{e}^{a(x')}}{\mathrm{e}^{a(x)}}\right)$$
$$-(1+\rho'+2\delta\rho')R. \tag{28}$$

Upon taking $\delta\to 0$ and optimizing $\rho'\geq 0$, $s\geq 0$, and $a(\cdot)$, this exponent is identical to the dual form of the mismatched decoding expurgated exponent given in [11], which is known to be equivalent to the primal form given in [5].

We also need to check that the choice of Δ in (26) complies with the rate condition in (18). We choose the same $a(\cdot)$ as in the exponent, but a different value of r (note that the two need not be the same). We simplify the condition as follows:

$$R\leq -\sum_x P(x)\log\sum_{x'}P(x')\mathrm{e}^{a(x')-\phi_a}\mathrm{e}^{-r(d_s(x,x')-\Delta)}-2\delta \tag{29}$$

$$=-\sum_x P(x)\log\sum_{x'}P(x')\mathrm{e}^{a(x')-\phi_a}\mathrm{e}^{-rd_s(x,x')}-r\Delta-2\delta \tag{30}$$

$$=-\sum_x P(x)\log\sum_{x'}P(x')\mathrm{e}^{a(x')-\phi_a}\mathrm{e}^{-rd_s(x,x')}$$
$$+r(1+\rho')\left(\sum_x P(x)\log\sum_{x'}P(x')\mathrm{e}^{-\frac{d_s(x,x')}{1+\rho'}}\frac{\mathrm{e}^{a(x')}}{\mathrm{e}^{a(x)}}+R+2\delta\right)-2\delta, \tag{31}$$

where we have substituted (26).

By setting $r=\frac{1}{1+\rho'}$ and noting that $-\sum_x P(x)\log\sum_{x'}P(x')\mathrm{e}^{a(x')-\phi_a}\mathrm{e}^{-rd_s(x,x')}$ is identical to $-\sum_x P(x)\log\sum_{x'}P(x')\frac{\mathrm{e}^{a(x')}}{\mathrm{e}^{a(x)}}\mathrm{e}^{-rd_s(x,x')}$ (by expanding the logarithms and using $\phi_a=\sum_x P(x)a(x)$), we observe that (31) reduces to $R\leq R$, which is trivially satisfied.

III. Cost-Constrained Codebook and Exponent

The preceding analysis is restricted to finite input alphabets. In the following, we describe how to construct a cost-constrained random code $\mathcal{M}_n$ with M_n codewords of length n over an arbitrary input alphabet, while guaranteeing that the minimum distance of the codebook exceeds Δ, in order to attain the exponent stated in Theorem 1 more generally.

As outlined in the proof of Theorem 1, one way of understanding (16) is by noting that it is the exponent that one obtains upon applying Gallager-type bounding techniques, e.g., Markov's inequality and $\min\{1,\alpha\}\leq\min_{\rho\in[0,1]}\alpha^\rho$, to the asymptotic multi-letter random coding union bound expression in (13) for constant-composition random coding. To our knowledge, the dual-domain analysis of constant-composition coding was initiated by Poltyrev [12].

It turns out that we can attain an analog of (13) for a cost-constrained coding scheme in which the input may also be continuous. In this section, we describe the changes needed in the code construction and analysis for this purpose. To simplify the presentation, we still use summations to denote averaging, but these can directly be replaced by integrals in continuous-alphabet settings. A disadvantage of cost-constrained coding, however, is that it is difficult to claim ensemble tightness; we provide only achievability results.

A. Code construction

Fix an input distribution P and four auxiliary costs $a_1(x), \ldots, a_4(x)$ (see, for example, [4], [10] for previous uses of auxiliary costs) Let P^n be the n-fold product of P, let $a_j(\boldsymbol{x}) = \frac{1}{n}\sum_{k=1}^n a_j(x_k)$ be the normalized additive extension of a_j, and define the cost-constrained distribution

$$P_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{\mu} P^n(\boldsymbol{x}) \mathbb{1}\Big\{\big|a_j(\boldsymbol{x}) - \phi_j\big| \le \epsilon, \quad j = 1,2,3,4\Big\}, \tag{32}$$

where $P^n(\boldsymbol{x}) = \prod_{k=1}^n P(x_k)$, $\phi_j = \mathbb{E}_P[a_j(X)]$, $\epsilon > 0$ is a parameter, and μ is a normalizing constant. Note that the auxiliary costs are intentionally introduced to improve the performance (in terms of the error exponent) of the random-coding ensemble. One can incorporate a *system cost* (e.g., a power constraint) in exactly the same way to ensure a per-codeword constraint of the form $\frac{1}{n}\sum_{k=1}^n c(x_k) \le \Gamma$ for some cost function c and threshold Γ; in such cases (which are crucial for continuous-alphabet settings), all of the subsequent analysis remains unchanged as long as P is chosen to satisfy $\mathbb{E}_P[c(X)] < \Gamma$.

By definition, $P_{\boldsymbol{X}}$ is i.i.d. conditioned on each a_j being close to its mean. Moreover, μ is the probability (under P^n) of the event in the indicator function of (32), and we immediately obtain $\lim_{n\to\infty}\mu = 1$ by the law of large numbers.

In the following, $\Pr(\boldsymbol{x}_m)$ is a shorthand for $\Pr(\boldsymbol{X}_m = \boldsymbol{x}_m)$, and similarly for conditional and joint probabilities. With the definition of $P_{\boldsymbol{X}}$ in place, we recursively generate the codewords in a similar manner to Section II:

$$\Pr(\boldsymbol{x}_1) = P_{\boldsymbol{X}}(\boldsymbol{x}_1) \tag{33}$$

$$\Pr(\boldsymbol{x}_2|\boldsymbol{x}_1) = \frac{1}{\mu_2(\boldsymbol{x}_1)} P_{\boldsymbol{X}}(\boldsymbol{x}_2)\mathbb{1}\big\{d(\boldsymbol{x}_1,\boldsymbol{x}_2) > \Delta\big\} \tag{34}$$

$$\vdots$$

$$\Pr(\boldsymbol{x}_m|\boldsymbol{x}_1^{m-1}) = \frac{1}{\mu_m(\boldsymbol{x}_1^{m-1})} \cdot P_{\boldsymbol{X}}(\boldsymbol{x}_m)\mathbb{1}\big\{d(\boldsymbol{x}_i,\boldsymbol{x}_m) > \Delta, i < m\big\}, \tag{35}$$

where each $\mu_m(\cdot)$ is a normalizing constant depending on all of the previous codewords. Note that in the case of continuous alphabets, each probability $\Pr(\boldsymbol{x}_i\,|\,\cdot)$ should be replaced by a conditional density function $f(\boldsymbol{x}_i\,|\,\cdot)$.

B. Codebook Properties

Here we provide two lemmas that characterizes the key properties of the recursive cost-constrained construction.

Lemma 1. *For any $\boldsymbol{x}$ with $P_{\boldsymbol{X}}(\boldsymbol{x}) > 0$, $\boldsymbol{X}' \sim P_{\boldsymbol{X}}$, and sufficiently large n, we have*

$$-\frac{1}{n}\log\Pr(d(\boldsymbol{x},\boldsymbol{X}') \le \Delta) \ge E_v(P,d,\Delta) - \delta \tag{36}$$

under suitable choices of the auxiliary costs $a_1(\cdot)$ and $a_2(\cdot)$.

Proof: We bound the probability as follows:

$$\begin{aligned}
&\Pr\big(d(\boldsymbol{x},\boldsymbol{X}') \le \Delta\big) \\
&= \sum_{\boldsymbol{x}'} P_{\boldsymbol{X}}(\boldsymbol{x}')\mathbb{1}\big\{d(\boldsymbol{x},\boldsymbol{x}') \le \Delta\big\} && (37)\\
&\le \sum_{\boldsymbol{x}'} P_{\boldsymbol{X}}(\boldsymbol{x}')\mathrm{e}^{-nr(d(\boldsymbol{x},\boldsymbol{x}')-\Delta)} && (38)\\
&\le \sum_{\boldsymbol{x}'} P_{\boldsymbol{X}}(\boldsymbol{x}')\mathrm{e}^{-nr(d(\boldsymbol{x},\boldsymbol{x}')-\Delta)}\mathrm{e}^{n(a_1(\boldsymbol{x}')-\phi_1+\epsilon)} && (39)\\
&\le \sum_{\boldsymbol{x}'} P^n(\boldsymbol{x}')\mathrm{e}^{-nr(d(\boldsymbol{x},\boldsymbol{x}')-\Delta)}\mathrm{e}^{n(a_1(\boldsymbol{x}')-\phi_1+2\epsilon)}, && (40)
\end{aligned}$$

where (38) uses Markov inequality with an arbitrary parameter $r > 0$, (39) uses the fact that $a_1(\boldsymbol{x}') \ge \phi_1 - \epsilon$ by construction, and (40) holds for sufficiently large n because $\mu \to 1$ in (32). Taking the logarithm and applying Gallager's single-letterization argument [14], we get

$$\begin{aligned}
&-\log\Pr\big(d(\boldsymbol{x},\boldsymbol{X}') \le \Delta\big) \\
&\ge -\sum_{k=1}^n \log\sum_{x'} P(x')\mathrm{e}^{-r(d(x_k,x')-\Delta)}\mathrm{e}^{a_1(x')-\phi_1} - 2n\epsilon.
\end{aligned} \tag{41}$$

We now choose the second auxiliary cost as $a_2(x) = -\log\sum_{x'} P(x')\mathrm{e}^{r(d(x,x')-\Delta)}\mathrm{e}^{a_1(x')-\phi_1}$, which ensures that the leading term on the right-hand side of (41) is equal to $na_2^n(\boldsymbol{x})$. Hence, substituting the definition $\phi_2 = \mathbb{E}_P[a_2(X)]$ and using $a_2(\boldsymbol{x}) \ge \phi_2 - \epsilon$ by construction, we obtain

$$\begin{aligned}
&-\frac{1}{n}\log\Pr\big(d(\boldsymbol{x},\boldsymbol{X}') \le \Delta\big) \\
&\ge -\sum_{x} P(x)\log\sum_{x'} P(x')\mathrm{e}^{-r(d(x,x')-\Delta)}\mathrm{e}^{a_1(x')-\phi_1} - 3\epsilon.
\end{aligned} \tag{42}$$

The proof is completed by choosing $\epsilon = \frac{\delta}{3}$, and optimizing r and $a_1(\cdot)$. ∎

Lemma 2. *Under the above cost-constrained random coding scheme with a given $\delta > 0$, if* (18) *holds then*

$$\Pr(\boldsymbol{x}_m) \doteq P_{\boldsymbol{X}}(\boldsymbol{x}_m), \tag{43}$$

$$\Pr(\boldsymbol{x}_k,\boldsymbol{x}_m) \le \frac{1}{(1-\mathrm{e}^{-n\delta})^2} \cdot P_{\boldsymbol{X}}(\boldsymbol{x}_k)P_{\boldsymbol{X}}(\boldsymbol{x}_m)\mathbb{1}\{d(\boldsymbol{x}_k,\boldsymbol{x}_m) > \Delta\}. \tag{44}$$

Proof: Letting $\boldsymbol{X}' \sim P_{\boldsymbol{X}}$, we have $\mu_m(\boldsymbol{x}_1^{m-1}) = \Pr(d(\boldsymbol{x}_i, \boldsymbol{X}') > \Delta, \forall i < m)$, and the union bound gives

$$1 - \mu_m(\boldsymbol{x}_1^{m-1}) \le \sum_{i<m} \Pr\left(d(\boldsymbol{x}_i, \boldsymbol{X}') \le \Delta\right) \tag{45}$$

$$\le e^{nR_n} \Pr\left(d(\boldsymbol{x}_i, \boldsymbol{X}') \le \Delta\right) \tag{46}$$

$$\le e^{-n\delta}, \tag{47}$$

where (47) follows from Lemma 1 and the rate condition (18).

Upper bounding the indicator functions in (33)–(35) by one gives $\Pr(\boldsymbol{x}_m) \dot{\le} P_{\boldsymbol{X}}(\boldsymbol{x}_m)$, thus proving one direction of the dot-equality in the first property of the lemma. The other direction requires more effort, and is deferred to [13].

For the second property of the lemma, we use (35) and the fact that $\mu_m(\boldsymbol{x}_1^{m-1}) \ge 1 - e^{-n\delta}$ to obtain

$$\begin{aligned}
&\Pr(\boldsymbol{x}_k, \boldsymbol{x}_m) \\
&= \sum_{\boldsymbol{x}_1^{k-1}, \boldsymbol{x}_{k+1}^{m-1}} \Pr(\boldsymbol{x}_1^{k-1}) \Pr(\boldsymbol{x}_k | \boldsymbol{x}_1^{k-1}) \\
&\quad \cdot \Pr(\boldsymbol{x}_{k+1}^{m-1} | \boldsymbol{x}_1^{k}) \Pr(\boldsymbol{x}_m | \boldsymbol{x}_1^{m-1}) \qquad (48) \\
&\le \sum_{\boldsymbol{x}_1^{k-1}, \boldsymbol{x}_{k+1}^{m-1}} \Pr(\boldsymbol{x}_1^{k-1}) \cdot \frac{P_{\boldsymbol{X}}(\boldsymbol{x}_k)}{1 - e^{-n\delta}} \cdot \Pr(\boldsymbol{x}_{k+1}^{m-1} | \boldsymbol{x}_1^{k}) \\
&\quad \cdot \frac{P_{\boldsymbol{X}}(\boldsymbol{x}_m) \mathbb{1}\{d(\boldsymbol{x}_k, \boldsymbol{x}_m) > \Delta\}}{1 - e^{-n\delta}} \qquad (49) \\
&= \frac{1}{(1 - e^{-n\delta})^2} P_{\boldsymbol{X}}(\boldsymbol{x}_k) P_{\boldsymbol{X}}(\boldsymbol{x}_m) \mathbb{1}\{d(\boldsymbol{x}_k, \boldsymbol{x}_m) > \Delta\}. \qquad (50)
\end{aligned}$$

■

C. Error Probability and Exponent

For a fixed transmitted codeword $\boldsymbol{x}_m$ and received sequence $\boldsymbol{y}$, we write that

$$\begin{aligned}
&\Pr\left(\bigcup_{\substack{i=1 \\ i \ne m}}^{M_n} \mathcal{E}_i \,\middle|\, \boldsymbol{X}_m = \boldsymbol{x}_m, \boldsymbol{Y} = \boldsymbol{y} \right) \\
&\le \sum_{\substack{i=1 \\ i \ne m}}^{M_n} \Pr\left(\mathcal{E}_i \,\middle|\, \boldsymbol{X}_m = \boldsymbol{x}_m, \boldsymbol{Y} = \boldsymbol{y}\right) \qquad (51) \\
&= \sum_{\substack{i=1 \\ i \ne m}}^{M_n} \sum_{\substack{\boldsymbol{x}_i \,:\, q(\boldsymbol{x}_i, \boldsymbol{y}) \ge q(\boldsymbol{x}_m, \boldsymbol{y}) \\ d(\boldsymbol{x}_i, \boldsymbol{x}_m) > \Delta}} \Pr(\boldsymbol{x}_i | \boldsymbol{x}_m, \boldsymbol{y}) \qquad (52) \\
&= \sum_{\substack{i=1 \\ i \ne m}}^{M_n} \sum_{\substack{\boldsymbol{x}_i \,:\, q(\boldsymbol{x}_i, \boldsymbol{y}) \ge q(\boldsymbol{x}_m, \boldsymbol{y}) \\ d(\boldsymbol{x}_i, \boldsymbol{x}_m) > \Delta}} \Pr(\boldsymbol{x}_i | \boldsymbol{x}_m) \qquad (53) \\
&= \sum_{\substack{i=1 \\ i \ne m}}^{M_n} \sum_{\substack{\boldsymbol{x}_i \,:\, q(\boldsymbol{x}_i, \boldsymbol{y}) \ge q(\boldsymbol{x}_m, \boldsymbol{y}) \\ d(\boldsymbol{x}_i, \boldsymbol{x}_m) > \Delta}} \frac{\Pr(\boldsymbol{x}_i, \boldsymbol{x}_m)}{\Pr(\boldsymbol{x}_m)} \qquad (54) \\
&\dot{\le} \frac{M_n - 1}{(1 - \mathrm{e}^{-n\delta})^2} \sum_{\substack{\bar{\boldsymbol{x}} \,:\, q(\bar{\boldsymbol{x}}, \boldsymbol{y}) \ge q(\boldsymbol{x}_m, \boldsymbol{y}) \\ d(\bar{\boldsymbol{x}}, \boldsymbol{x}_m) \ge \Delta}} P_{\boldsymbol{X}}(\bar{\boldsymbol{x}}) \qquad (55)
\end{aligned}$$

where (53) follows since $\boldsymbol{X}_i - \boldsymbol{X}_m - \boldsymbol{Y}$ forms a Markov chain, and (55) follows from Lemma 2. Taking the minimum of this union bound and one, and averaging over all choices of transmitted codeword $\boldsymbol{x}_m$ and received sequence $\boldsymbol{y}$, we obtain an asymptotic upper bound that matches (13):

$$\begin{aligned}
\bar{P}_{\mathrm{e}}^{(n)} &\dot{\le} \sum_{\boldsymbol{x}, \boldsymbol{y}} P_{\boldsymbol{X}}(\boldsymbol{x}) W^n(\boldsymbol{y} | \boldsymbol{x}) \\
&\quad \cdot \min\left\{1, (M_n - 1) \sum_{\substack{\boldsymbol{x}' \,:\, q(\boldsymbol{x}', \boldsymbol{y}) \ge q(\boldsymbol{x}, \boldsymbol{y}) \\ d(\boldsymbol{x}', \boldsymbol{x}) \ge \Delta}} P_{\boldsymbol{X}}(\boldsymbol{x}')\right\}. \qquad (56)
\end{aligned}$$

Once (56) is established, the steps in deriving (16) are standard. Such an analysis requires two additional auxiliary costs, and these are given by a_3 and a_4 in (32). In particular, we set $a_3(x) = a(x)$ in (16) and

$$\begin{aligned}
a_4(x) &= -\log \sum_y W(y|x) \\
&\quad \cdot \left(\frac{\sum_{x'} P(x') e^{sq(x', y)} e^{a(x')} e^{r(d(x, x') - \Delta)}}{e^{sq(x, y)} e^{a(x)}} \right)^{\rho}. \qquad (57)
\end{aligned}$$

In fact, removing the constraint $d(\boldsymbol{x}, \boldsymbol{x}') > \Delta$ from the pairwise error probability term in (56) recovers the standard random-coding union bound, which was already used in [10] to establish the exponent in (16) without the term $\mathrm{e}^{r(d(x,x')-\Delta)}$. Hence, the change in the analysis compared to [10] only amounts to an application of $\mathbb{1}\{d(\boldsymbol{x}, \boldsymbol{x}') \ge \Delta\} \le \mathrm{e}^{nr(d(\boldsymbol{x}, \boldsymbol{x}') - \Delta)}$. Due to this similarity, the details are omitted.

REFERENCES

[1] I. Csiszár and J. Körner, *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.

[2] N. Merhav, G. Kaplan, A. Lapidoth, and S. Shamai (Shitz), "On information rates for mismatched decoders," *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 1953–1967, Nov. 1994.

[3] I. Csiszár and P. Narayan, "Channel capacity for a given decoding metric," *IEEE Trans. Inf. Theory*, vol. 41, no. 1, pp. 35–43, Jan. 1995.

[4] A. Ganti, A. Lapidoth, and I. Telatar, "Mismatched decoding revisited: general alphabets, channels with memory, and the wide-band limit," *IEEE Trans. Inf. Theory*, vol. 46, no. 7, pp. 2315–2328, Nov. 2000.

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

[6] A. Somekh-Baruch, J. Scarlett, and A. Guillén i Fàbregas, "The error exponent of generalized random Gilbert-Varshamov codes," in *IEEE Int. Symp. on Inf. Theory (ISIT)*, 2018.

[7] E. N. Gilbert, "A comparison of signalling alphabets," *Bell Labs Tech. J.*, vol. 31, no. 3, pp. 504–522, 1952.

[8] R. R. Varshamov, "Estimate of the number of signals in error correcting codes," in *Dokl. Akad. Nauk SSSR*, vol. 117, no. 5, 1957, pp. 739–741.

[9] Y. Polyanskiy, V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[10] J. Scarlett, A. Martinez, and A. Guillén i Fàbregas, "Mismatched decoding: Error exponents, second-order rates and saddlepoint approximations," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2647–2666, May 2014.

[11] J. Scarlett, L. Peng, N. Merhav, A. Martinez, and A. Guillén i Fàbregas, "Expurgated random-coding ensembles: Exponents, refinements, and connections," *IEEE Trans. Inf. Theory*, vol. 60, no. 8, pp. 4449–4462, Aug 2014.

[12] G. Poltyrev, "Random coding bounds for discrete memoryless channels," *Prob. Inf. Transm.*, vol. 18, no. 1, pp. 9–21, 1982.

[13] A. Somekh-Baruch, J. Scarlett, and A. Guillén i Fàbregas, "Generalized random Gilbert-Varshamov codes," 2019, accepted to *IEEE Transactions on Information Theory*.

[14] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.