Generalized Random Gilbert-Varshamov Codes

Anelia Somekh-Baruch, Jonathan Scarlett, and Albert Guillén i Fàbregas

***Abstract*—We introduce a random coding technique for transmission over discrete memoryless channels, reminiscent of the basic construction attaining the Gilbert-Varshamov bound for codes in Hamming spaces. The code construction is based on drawing codewords recursively from a fixed type class, in such a way that a newly generated codeword must be at a certain minimum distance from all previously chosen codewords, according to some generic distance function. We derive an achievable error exponent for this construction, and prove its tightness with respect to the ensemble average. We show that the exponent recovers the Csiszár and Körner exponent as a special case, which is known to be at least as high as both the random-coding and expurgated exponents, and we establish the optimality of certain choices of the distance function. In addition, for additive distances and decoding metrics, we present an equivalent dual expression, along with a generalization to non-finite alphabets via cost-constrained random coding.**

I. Introduction

The problem of characterizing the error exponents of channel coding has been studied extensively since the early days of information theory. The goal is to establish bounds on the rate of decay of the error probability for fixed rates below capacity. While the random coding exponent and sphere packing exponent establish the exact error exponent at rates sufficiently close to capacity, the optimal exponent at low rates has generally remained open, except in the limit of zero rate.

For discrete memoryless channels (DMC), improvements over the random-coding exponent at low rates are provided by the expurgated exponent. The idea of the original derivation of this exponent is simple [1]: After generating the codewords independently at random, remove a fraction of the worst codewords (i.e., those with the highest error probability) while keeping enough so that the loss in the rate is negligible. Alternative derivations have since appeared based on the method of types and random selection [2], graph decomposition techniques [3], and type class enumeration [4]. For other related works, see [5], [6], [7] and references therein.

A. Somekh-Baruch is with the Faculty of Engineering, Bar-Ilan University, Ramat Gan 52900, Israel (e-mail: somekha@biu.ac.il).

J. Scarlett is with the Department of Computer Science and Department of Mathematics, National University of Singapore, Singapore (e-mail: scarlett@comp.nus.edu.sg).

A. Guillén i Fàbregas is with the Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona 08018, Spain, also with the Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona 08010, Spain, and also with the Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, U.K. (e-mail: guillen@ieee.org).

This work was supported in part by the Israel Science Foundation under grant 631/17, by 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. This paper was presented in part at the 2018 International Zurich Seminar, the 2018 Conference on Information Sciences and Systems, Princeton University and the 2018 IEEE International Symposium on Information Theory.

In the literature, many of the most commonly-studied error exponents admit (at least) two equivalent forms:

- A *primal* expression is written as a minimization over joint distributions subject to suitable constraints, and is typically derived using the method of types [2]. Such derivations often have the advantage of immediately proving tightness with respect to the random-coding ensemble under consideration.
- A *dual* expression is written as a maximization over auxiliary parameters, and is typically derived using Gallager-type techniques [1] such as Markov's inequality and $\min\{1,\alpha\} \leq \alpha^{\rho}$ for $\rho \in [0,1]$. Such derivations often have the advantage of extending to continuous-alphabet memoryless channels. In addition, dual expressions provide achievable exponents for arbitrary *fixed* choices of the auxiliary parameters.

This naming convention arises from the fact that the equivalence of the expressions is proved using Lagrange duality. In the setting of the present paper with a general additive decoding metric, such equivalences were given for achievable rates in [8], for random coding error exponents in [9], [10], and for expurgated exponents in [4].

In this paper, we introduce a recursive random coding construction that achieves the exponent of Csiszár and Körner [3], thus achieving the maximum of the random-coding and expurgated exponents. The code construction is based on drawing codewords recursively from a fixed type class, in such a way that a newly generated codeword must be at a certain minimum distance from all previously chosen codewords, according to some generic distance function. This construction is reminiscent to those in the binary Hamming space dating back to the 1950s [11]–[13] (see also [14]–[17]), known to achieve the Gilbert-Varshamov bound. We therefore adopt the name generalized *Random Gilbert-Varshamov (RGV) code* for our randomized construction with a general distance function and constant-composition codewords. A related work by Blahut [18] studied properties of the Bhattacharyya and the equivocation distance functions and derived generalized bounds similar to the Gilbert-Varshamov and Elias bounds. These bounds are used to derive an upper bound on the reliability function. Connections with the expurgated exponent are also explored. Another related work is that of Barg and Forney [19], who showed that for the binary symmetric channel (BSC), typical linear codes, whose minimum distance attains the Gilbert-Varshamov bound, achieve the expurgated exponent.

A. Contributions

The main contributions of this work are as follows:

- As outlined above, we introduce the generalized RGV construction, and analyze its error exponent for a given

DMC, decoding metric, and distance function. Similarly to the Gilbert-Varshamov bound, our construction induces a tradeoff between the rate and the minimum distance of the code. As well as establishing an achievable exponent, we derive an *ensemble tightness* result implying that one cannot do better with such a construction. Proving this is non-trivial compared to previous ensemble tightness results (e.g., for random coding exponents [9], [20] and achievable rates [8]). Among other things, the distribution of the drawn codeword depends on its index in the recursive construction, and on all of the previous codewords, so one cannot use a symmetry argument to focus on a single message.

- We show that when the distance function is optimized, the generalized RGV construction achieves the exponent of Csiszár and Körner [3], which is at least as high as both the random-coding and expurgated exponents. While the analysis of [3] establishes the existence of codes attaining the exponent using a decomposition lemma, our scheme provides a specific randomized construction that spreads the codewords according to a generic distance function, and whose ensemble average directly achieves the exponent.
- In the case of an additive distance measure (e.g., Hamming or Bhattacharyya distance) and decoding metric (e.g., maximum-likelihood), we give an equivalent dual expression for our error exponent, as well as providing a direct derivation of the dual form using cost-constrained random coding [4], [10]. This alternative derivation allows us to extend the achievability part to memoryless channels with infinite or continuous alphabets.
- We prove that the distance function that measures closeness according to the joint empirical mutual information (equivalent to the equivocation distance [18]) maximizes the exponent of our construction, at least among symmetric distance functions depending only on the joint type. This optimality is universal, in the sense that it holds for every channel and every type-dependent decoding metric. In addition, we provide an alternative non-universal distance function yielding the same error exponent, and we show that an additive Chernoff-based distance measure (which reduces to the Bhattacharrya distance in the case of maximum-likelihood decoding) recovers both the random coding and expurgated exponents.

B. Notation

The set of probability mass functions on a finite alphabet $\mathcal{X}$ is denoted by $\mathcal{P}(\mathcal{X})$. We use standard notations for entropy, mutual information, and so on (e.g., $I(X;Y)$, $H(X|Y)$), sometimes using a subscript to indicate the underlying distribution (e.g., $I_V(X;Y)$ for some joint distribution V_{XY}). These are all taken to be in units of nats, and the function $\log$ has the natural base. We denote sequences (vectors) in boldfaced font, e.g., $\boldsymbol{x}$. For $i < j$, we let $\boldsymbol{x}_i^j$ denote $(\boldsymbol{x}_i, \ldots, \boldsymbol{x}_j)$, and similarly, $\boldsymbol{X}_i^j = (\boldsymbol{X}_i, \ldots, \boldsymbol{X}_j)$.

We make frequent use of types [2, Ch. 2]. The type (i.e., empirical distribution) of a sequence $\boldsymbol{x}$ is denoted by $\hat{P}_{\boldsymbol{x}}$, and similarly for joint types $\hat{P}_{\boldsymbol{xy}}$ and conditional types $\hat{P}_{\boldsymbol{y}|\boldsymbol{x}}$. The set of all types for a given sequence length n is denoted by $\mathcal{P}_n(\mathcal{X})$. The type class $\mathcal{T}(P)$ is the set of all sequences with type P, and the conditional type class $\mathcal{T}(P_{\widetilde{X}|X})$ is the set of all $\widetilde{X}$-sequences inducing a given conditional type $P_{\widetilde{X}|X}$ for an arbitrary fixed X-sequence (whose type will be clear from the context).

For two positive sequences f_n and g_n, we write $f_n \doteq g_n$ if $\lim_{n\to\infty} \frac{1}{n} \log \frac{f_n}{g_n} = 0$, $f_n \dot{\leq} g_n$ if $\limsup_{n\to\infty} \frac{1}{n} \log \frac{f_n}{g_n} \leq 0$, and similarly for $\dot{\geq}$.

C. Structure of the Paper

In Section II, we formally introduce the channel coding setup and introduce additional notation. In Section III, we describe the recursive random codebook construction and establish its main properties. Section IV gives the main result and its proof, and Section V gives the equivalent dual expression and its direct derivation. Section VI studies the optimality of some specific distance functions.

II. Problem Setup

We consider the problem of reliable transmission over a DMC described by a conditional probability mass function $W(y|x)$, with input $x \in \mathcal{X}$ and output $y \in \mathcal{Y}$ for finite alphabets $\mathcal{X}$ and $\mathcal{Y}$. We define

$$W^n(\boldsymbol{y}|\boldsymbol{x}) = \prod_{k=1}^{n} W(y_k|x_k) \tag{1}$$

for input/output sequences $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n, \boldsymbol{y} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$. We use the notation $\boldsymbol{X}, \boldsymbol{Y}$ to denote the corresponding random variables. Infinite and continuous alphabets are addressed in Section V.

An encoder maps a message $m \in \{1, \ldots, M_n\}$ 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 $\{1, \ldots, M_n\}$ equiprobably. This mapping induces an (n, M_n)-codebook $\mathcal{C}_n = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{M_n}\}$ with rate $R_n = \frac{1}{n} \log M_n$.

The decoder has access to the codebook and, upon observing the channel output $\boldsymbol{y}$, produces an estimate of the transmitted message $\hat{m} \in \{1, \ldots, M_n\}$. We consider the family of maximum metric decoders for which the transmitted message is estimated as

$$\hat{m} = \operatorname*{arg\,max}_{\boldsymbol{x}_i \in \mathcal{C}_n} q(\boldsymbol{x}_i, \boldsymbol{y}) \tag{2}$$

where $q(\boldsymbol{x}, \boldsymbol{y}) : \mathcal{X}^n \times \mathcal{Y}^n \to \mathbb{R}$ is a generic decoding metric. Whenever two or more candidate codewords have the same decoding metric, an error will be assumed. Whenever $q(\boldsymbol{x}, \boldsymbol{y})$ is an increasing function of the channel transition law $W^n(\boldsymbol{y}|\boldsymbol{x})$ we recover the maximum-likelihood (ML) decoder. Otherwise, the decoder is said to be mismatched [8], [21]. Throughout the paper, we assume that the decoding metric $q(\boldsymbol{x}, \boldsymbol{y})$ only depends on the joint empirical distribution (or type) of $\boldsymbol{x}, \boldsymbol{y}$, i.e., $\hat{P}_{\boldsymbol{x},\boldsymbol{y}}$. In this case, we write the decoder as

$$\hat{m} = \operatorname*{arg\,max}_{\boldsymbol{x} \in \mathcal{C}_n} q(\hat{P}_{\boldsymbol{x},\boldsymbol{y}}), \tag{3}$$

where we assume that the type-dependent metric $q : \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$ is continuous (and therefore bounded) on the probability simplex.[1] An important class of such metrics is the class of *additive metrics*, taking the form

$$q(\hat{P}_{\boldsymbol{x},\boldsymbol{y}}) = \frac{1}{n}\sum_{i=1}^{n} q(x_i, y_i) = \mathbb{E}_{\hat{P}_{\boldsymbol{x},\boldsymbol{y}}}[q(X,Y)], \tag{4}$$

where $q(x,y)$ is a *single-letter* metric (abusing notation slightly), and the average is with respect to the joint empirical distribution. A notable example of a non-additive type-dependent metric is the empirical mutual information, $q(\hat{P}_{\boldsymbol{x},\boldsymbol{y}}) = I_{\hat{P}_{\boldsymbol{x},\boldsymbol{y}}}(X;Y)$.

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

$$\liminf_{n\to\infty} -\frac{1}{n}\log \Pr\big(\hat{S} \neq S\big) \geq E - \epsilon. \tag{5}$$

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

III. Random Codebook and Properties

In this section, we introduce our recursive random coding scheme, and state its main properties used for deriving the associated error exponent.

Codes that attain the Gilbert-Varshamov bound on the Hamming space [11], [12] ensure that all codewords are at least at a certain target Hamming distance Δ from each other. The generalized RGV construction is a randomized constant-composition counterpart of such codes for arbitrary DMCs and more general distance functions.

Definition 1. *Let Ω be the set of bounded, continuous, symmetric, and type-dependent functions $d(\cdot,\cdot) : \mathcal{X}^n \times \mathcal{X}^n \to \mathbb{R}$, i.e., bounded functions that satisfy $d(\boldsymbol{x},\boldsymbol{x}') = d(\boldsymbol{x}',\boldsymbol{x})$ for all $\boldsymbol{x},\boldsymbol{x}' \in \mathcal{X}^n$, that depend on $(\boldsymbol{x},\boldsymbol{x}')$ only through the joint empirical distribution $\hat{P}_{\boldsymbol{x}\boldsymbol{x}'}$, and that are continuous on the probability simplex.*

We use the notation $d(\boldsymbol{x},\boldsymbol{x}')$ and $d(\hat{P}_{\boldsymbol{x}\boldsymbol{x}'})$ interchangeably for convenience, similarly to $q(\boldsymbol{x},\boldsymbol{y})$ and $q(\hat{P}_{\boldsymbol{x}\boldsymbol{y}})$. We refer to $d \in \Omega$ as a *distance function*, though it need not be a distance in the topological sense (e.g., it may be negative).

Some examples of distance functions in Ω are as follows:

- We say that the distance function is *additive* if it can be written as

$$d(\boldsymbol{x},\boldsymbol{x}') = \frac{1}{n}\sum_{k=1}^{n} d(x_k, x'_k) \tag{6}$$

  for some single-letter function $d(x,x')$ (abusing notation slightly). Any such distance function is in Ω, as long as $d(x,x')$ is symmetric. Notable examples include the Hamming distance

$$d_{\mathrm{H}}(x,x') = \mathbb{1}\{x' \neq x\}, \tag{7}$$

  and the Bhattacharyya distance

$$d_{\mathrm{B}}(x,x') = -\log \sum_{y\in\mathcal{Y}} \sqrt{W(y|x)W(y|x')}. \tag{8}$$

  Note that the latter choice depends on the channel, and to satisfy the boundedness assumption we require that any two inputs have a common output that is produced with positive probability.
- We will later consider a distance equal to the negative mutual information, $d(P_{X\widetilde{X}}) = -I_P(X;\widetilde{X})$, which will turn out to be universally optimal subject to the constraints of our construction. For constant-composition codes, it is equivalent to the *equivocation distance* $d(P_{X\widetilde{X}}) = H_P(\widetilde{X}|X)$, which was considered in a different but related context by Blahut [18].

In the following, we describe how to construct a code $\mathcal{C}_n$ with M_n codewords of length n, such that any two distinct codewords $\boldsymbol{x},\boldsymbol{x}' \in \mathcal{C}_n$ satisfy $d(\boldsymbol{x},\boldsymbol{x}') > \Delta$ for a given function $d(\cdot,\cdot) \in \Omega$ and threshold $\Delta \in \mathbb{R}$. This guarantees that the minimum distance of the codebook exceeds Δ. The construction depends on an input distribution $P \in \mathcal{P}(\mathcal{X})$, and throughout the paper, we let $P_n \in \mathcal{P}_n(\mathcal{X})$ denote an arbitrary type with the same as support as P satisfying $\max_{x\in\mathcal{X}} |P_n(x) - P(x)| \leq \frac{1}{n}$.

Along with $P \in \mathcal{P}(\mathcal{X})$, fixing n, M_n, a distance function $d(\cdot,\cdot) \in \Omega$, and constants $\delta > 0, \Delta \in \mathbb{R}$, the construction is described by the following steps:

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

$$\mathcal{T}(P_n, \boldsymbol{x}_1) \triangleq \{\bar{\boldsymbol{x}} \in \mathcal{T}(P_n) \,:\, d(\bar{\boldsymbol{x}}, \boldsymbol{x}_1) > \Delta\} \tag{9}$$

$$= \mathcal{T}(P_n) \backslash \{\bar{\boldsymbol{x}} \in \mathcal{T}(P_n) \,:\, d(\bar{\boldsymbol{x}}, \boldsymbol{x}_1) \leq \Delta\}, \tag{10}$$

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

$$\mathcal{T}(P_n, \boldsymbol{x}_1^{i-1}) \triangleq \{\bar{\boldsymbol{x}} \in \mathcal{T}(P_n) \,:\, d(\bar{\boldsymbol{x}}, \boldsymbol{x}_j) > \Delta,\, j = 1\ldots, i-1\} \tag{11}$$

$$= \mathcal{T}(P_n, \boldsymbol{x}_1^{i-2}) \backslash \big\{\bar{\boldsymbol{x}} \in \mathcal{T}(P_n, \boldsymbol{x}_1^{i-2}) \,:\, d(\bar{\boldsymbol{x}}, \boldsymbol{x}_{i-1}) \leq \Delta\big\}. \tag{12}$$

Throughout the paper, it will be useful to generalize the notation $\mathcal{T}(P_n, \boldsymbol{x}_1^{i-1})$ as follows. For any subset $\mathcal{D} \subseteq \mathcal{T}(P_n)$, we define

$$\mathcal{T}(P_n, \mathcal{D}) \triangleq \{\boldsymbol{x} \in \mathcal{T}(P_n) :\, d(\boldsymbol{x},\boldsymbol{x}') > \Delta,\, \forall \boldsymbol{x}' \in \mathcal{D}\}. \tag{13}$$

In Lemma 1 below, we will show that 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

$$\mathrm{e}^{n(R_n+\delta)} \mathrm{vol}_{\boldsymbol{x}}(\Delta) \leq |\mathcal{T}(P_n)| \tag{14}$$

where $\mathrm{vol}_{\boldsymbol{x}}(\Delta) = |\{\bar{\boldsymbol{x}} \in \mathcal{T}(P_n) \,:\, d(\bar{\boldsymbol{x}},\boldsymbol{x}) \leq \Delta\}|$ is the "volume" of a "ball" of radius Δ according to the "distance"

[1] Similarly to [3], our analysis easily extends to the ML decoding metric, for which $q(\boldsymbol{x},\boldsymbol{y})$ may equal $-\infty$ when $W^n(\boldsymbol{y}|\boldsymbol{x}) = 0$.

$d(\cdot,\cdot)$, 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 $\boldsymbol{x} \in \mathcal{T}(P_n)$. It will be convenient to rewrite (14) as

$$\sum_{\bar{\boldsymbol{x}} \in \mathcal{T}(P_n)\,:\, d(\bar{\boldsymbol{x}},\boldsymbol{x}) \le \Delta} \frac{1}{|\mathcal{T}(P_n)|} \le \mathrm{e}^{-n(R_n+\delta)}. \tag{15}$$

A. Codebook Properties

Here we provide several lemmas characterizing the key properties of the generalized RGV construction. We begin with the fact that the construction is well-defined, in the sense that the procedure described above always produces the desired number of codewords M_n, i.e., the set $\mathcal{T}(P_n, \boldsymbol{x}_1^{i-1})$ given the previous codewords is always non-empty.

Lemma 1. *The generalized RGV codebook construction with condition* (15) *is such that for all $i \in \{1,\dots,M_n\}$, all $\boldsymbol{x}_1^{i-1}$ occurring with non-zero probability, and any $\delta > 0$, we have*

$$(1-\mathrm{e}^{-n\delta})|\mathcal{T}(P_n)| \le |\mathcal{T}(P_n, \boldsymbol{x}_1^{i-1})| \le |\mathcal{T}(P_n)|. \tag{16}$$

Proof. The upper bound is trivial, since

$$\mathcal{T}(P_n, \boldsymbol{x}_1^{M_n-1}) \subseteq \cdots \subseteq \mathcal{T}(P_n, \boldsymbol{x}_1^{i-1}) \subseteq \mathcal{T}(P_n, \boldsymbol{x}_1^{i-2}) \subseteq \cdots \subseteq \mathcal{T}(P_n). \tag{17}$$

For the lower bound, we make use of (14)–(15). After $M_n = \mathrm{e}^{nR_n}$ iterations of the above procedure, we have removed no more than $\mathrm{e}^{nR_n}\mathrm{vol}_{\boldsymbol{x}}(\Delta) \le |\mathcal{T}(P_n)|\mathrm{e}^{-n\delta}$ sequences from $\mathcal{T}(P_n)$. This implies that after iteration $M_n = \mathrm{e}^{nR_n}$,

$$|\mathcal{T}(P_n, \boldsymbol{x}_1^{M_n-1})| \ge |\mathcal{T}(P_n)| - \mathrm{e}^{nR_n}\mathrm{vol}_{\boldsymbol{x}}(\Delta) \tag{18}$$

$$\ge |\mathcal{T}(P_n)|(1-\mathrm{e}^{-n\delta}). \tag{19}$$

The lower bound in (16) for $i \in \{1,\dots,M_n\}$ follows from (19) and (17).

□

Henceforth, whenever we refer to the generalized RGV construction, this implicitly includes the condition (15) (or equivalently (14)).

The following lemmas provide upper and lower bounds on the marginal distributions of small numbers of codewords (up to three) in the RGV construction. We make use of the following exponentially vanishing quantity:

$$\delta_n \triangleq \frac{\mathrm{e}^{-n\delta}}{1-\mathrm{e}^{-n\delta}}. \tag{20}$$

We begin with the joint distribution between two codewords, as this plays the most important role in our analysis. Here and subsequently, $\Pr(\boldsymbol{x}_k, \boldsymbol{x}_m)$ is a shorthand for $\Pr(\boldsymbol{X}_k = \boldsymbol{x}_k, \boldsymbol{X}_m = \boldsymbol{x}_m)$, and similarly for other expressions such as $\Pr(\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{x}_k)$.

Lemma 2. *Under the generalized RGV construction, for any $k \in \{1,\dots,M_n-1\}$, $m > k$ and $\boldsymbol{x}_k, \boldsymbol{x}_m \in \mathcal{T}(P_n)$, if $d(\boldsymbol{x}_k, \boldsymbol{x}_m) > \Delta$ then we have*

$$\frac{(1-4\delta_n^2)}{|\mathcal{T}(P_n)|^2}\mathrm{e}^{-2\delta_n} \le \Pr(\boldsymbol{x}_k, \boldsymbol{x}_m) \le \frac{1}{(1-\mathrm{e}^{-n\delta})^2|\mathcal{T}(P_n)|^2}, \tag{21}$$

while $\Pr(\boldsymbol{x}_k, \boldsymbol{x}_m) = 0$ whenever $d(\boldsymbol{x}_k, \boldsymbol{x}_m) \le \Delta$.

Proof. See Appendix A. □

Note that here k and m are arbitrary indices, and m need not correspond to the transmitted message. In some cases, we will apply the lemma with m being the transmitted message.

For the joint distribution between three codewords, we will only require an upper bound, and it will only be used for the ensemble tightness part.

Lemma 3. *Under the generalized RGV construction, for any $i,j,k \in \{1,\dots,M_n\}$, such that $i<j<k$ and $\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{x}_k \in \mathcal{T}(P_n)$, if $\min\{d(\boldsymbol{x}_i,\boldsymbol{x}_j), d(\boldsymbol{x}_i,\boldsymbol{x}_k), d(\boldsymbol{x}_j,\boldsymbol{x}_k)\} > \Delta$ then*

$$\Pr(\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{x}_k) \le \frac{1}{(1-e^{-n\delta})^3|\mathcal{T}(P_n)|^3}, \tag{22}$$

while $\min\{d(\boldsymbol{x}_i,\boldsymbol{x}_j), d(\boldsymbol{x}_i,\boldsymbol{x}_k), d(\boldsymbol{x}_j,\boldsymbol{x}_k)\} \le \Delta$ whenever $\Pr(\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{x}_k) = 0$.

Proof. See Appendix B. □

Finally, by a basic symmetry argument, the marginal distribution of any given codeword $\boldsymbol{X}_m$ (without any conditioning) is uniform over $\mathcal{T}(P_n)$, as stated in the following.

Lemma 4. *For any message index m, the marginal distribution of codeword $\boldsymbol{X}_m$ is $\Pr(\boldsymbol{x}_m) = \frac{1}{|\mathcal{T}(P_n)|}$ for $\boldsymbol{x}_m \in \mathcal{T}(P_n)$, and zero elsewhere.*

Proof. See Appendix C. □

IV. Main Result

Using graph decomposition techniques, Csiszár and Körner [3] studied the error exponents of constant-composition codes under a decoder that uses a type-dependent decoding metric $q(\hat{P}_{\boldsymbol{x},\boldsymbol{y}})$, and derived the following achievable exponent for an arbitrary input distribution P:

$$E_q(R,P,W) = \min_{V \in \mathcal{T}_I} D(V_{Y|X}\|W|P) + \big|I(\widetilde{X};Y,X) - R\big|_+, \tag{23}$$

where

$$\mathcal{T}_I \triangleq \Big\{V_{X\widetilde{X}Y} \in \mathcal{P}(\mathcal{X}\times\mathcal{X}\times\mathcal{Y}) : V_X = V_{\widetilde{X}} = P, q(V_{\widetilde{X}Y}) \ge q(V_{XY}), I(X;\widetilde{X}) \le R\Big\}. \tag{24}$$

This exponent was shown to be at least as high as the maximum of the expurgated exponent and the random coding exponent.

The following theorem presents an exact single-letter expression for the error exponent of the recursive RGV codebook construction described in the previous section. We show in Section VI that it reduces to the exponent of [3], $E_q(R,P,W)$, when the distance function $d(\cdot,\cdot)$ is optimized.

Letting

$$E_{\mathrm{RGV}}(R,P,W,q,d,\Delta) = \min_{V_{X\widetilde{X}Y} \in \mathcal{T}_{d,q,P}(\Delta)} D(V_{Y|X}\|W|P) + \big|I(\widetilde{X};Y,X) - R\big|_+, \tag{25}$$

where

$$\mathcal{T}_{d,q,P}(\Delta) \triangleq \Big\{V_{X\widetilde{X}Y} \in \mathcal{P}(\mathcal{X}\times\mathcal{X}\times\mathcal{Y}) : V_X = V_{\widetilde{X}} = P,\, q(V_{\widetilde{X}Y}) \geq q(V_{XY}),\, d(V_{X\widetilde{X}}) \geq \Delta\Big\}, \quad (26)$$

we have the following.

Theorem 1. *For all $P \in \mathcal{P}(\mathcal{X})$, $\delta > 0$, $\Delta \in \mathbb{R}$, $d \in \Omega$, and $R > 0$ satisfying*

$$R \leq \min_{P_{X\widetilde{X}} : d(P_{X\widetilde{X}}) \leq \Delta,\, P_X = P_{\widetilde{X}} = P} I(X;\widetilde{X}) - 2\delta, \quad (27)$$

the ensemble average error probability $\bar{P}_{\mathrm{e}}^{(n)}$ of the generalized RGV construction with parameters $(n, R, P, d, \Delta, \delta)$ and the bounded continuous type-dependent decoding metric $q(\cdot)$ over the DMC W satisfies

$$\bar{P}_{\mathrm{e}}^{(n)} \dot{\leq} \mathrm{e}^{-nE_{\mathrm{RGV}}(R,P,W,q,d,\Delta)}. \quad (28)$$

In addition, if q is an additive decoding metric, then

$$\bar{P}_{\mathrm{e}}^{(n)} \dot{\geq} \mathrm{e}^{-nE_{\mathrm{RGV}}(R,P,W,q,d,\Delta+\epsilon)} \quad (29)$$

for arbitrarily small $\epsilon > 0$.

The achievability proof (i.e., upper bound on the error probability) is given in Section IV-A, and the ensemble tightness proof (i.e., lower bound on the error probability) for additive metrics is given in Section IV-B.

As will be shown in Section VI, under the rate constraint (27), if the distance function is chosen appropriately, the generalized RGV construction achieves the exponent $E_q(R, P, W)$ in (23), which in turn shows the achievability of capacity for ML decoding or the LM rate in the mismatched case [3], [22]. Moreover, for a distance function d that uniquely attains its minimum value when $X = X'$, varying Δ from its minimum to maximum value yields all possible values of rates in $(0, H(P))$, which covers the entire range of possible rates with constant composition codes.

Theorem 1 implies that the exact exponent of the coding scheme equals $E_{\mathrm{RGV}}(R, P, W, q, d, \Delta)$ whenever Δ is a continuity point. We note that while the additivity of $q(\cdot)$ is required for the derivation of the lower bound on $\bar{P}_{\mathrm{e}}^{(n)}$, the upper bound holds also for any continuous $q(\cdot)$ that need not be additive. For non-additive q, in the assertion of the lower bound (29) for non additive q, we would have to replace $E_{\mathrm{RGV}}(R, P, W, q, d, \Delta + \epsilon)$ by

$$\min_{V_{X\widetilde{X}Y} \in \mathcal{T}_{d,q,P,\epsilon}(\Delta+\epsilon)} D(V_{Y|X}\|W|P) + \big|I(\widetilde{X};Y,X) - R\big|_+, \quad (30)$$

where

$$\mathcal{T}_{d,q,P,\epsilon}(\Delta+\epsilon) \triangleq \Big\{V_{X\widetilde{X}Y} \in \mathcal{P}(\mathcal{X}\times\mathcal{X}\times\mathcal{Y}) : V_X = V_{\widetilde{X}} = P, q(V_{\widetilde{X}Y}) \geq q(V_{XY}) + \epsilon,\, d(V_{X\widetilde{X}}) \geq \Delta + \epsilon\Big\}, \quad (31)$$

that is, we would have the extra ϵ in $q(V_{\widetilde{X}Y}) \geq q(V_{XY}) + \epsilon$. While this yields the desired tightness result whenever the optimization is "continuous" with respect to the metric constraint, it is unclear in what generality such continuity holds. As a simple example, $\mathcal{T}_{d,q,P,\epsilon}(\Delta + \epsilon)$ is always empty under the erasures-only metric $q(x, y) = \mathbb{1}\{W(y|x) > 0\}$, meaning that (30) does not provide a tightness result in this case.

By a simple symmetrization argument, we can show that $E_{\mathrm{RGV}}(R, P, W, q, d, \Delta)$ is an achievable error exponent even when d is not symmetric. This is stated in the following.

Corollary 1. *Under the setup of Theorem 1 with a non-symmetric continuous type-dependent bounded distance function d, if the pair (R, Δ) satisfies* (27)*, then the error exponent $E_{\mathrm{RGV}}(R, P, W, q, d, \Delta)$ is achievable at rate R.*

Proof. We apply Theorem 1 with the symmetric distance

$$d'(\boldsymbol{x}, \boldsymbol{x}') = \min\big\{d(\boldsymbol{x}, \boldsymbol{x}'), d(\boldsymbol{x}', \boldsymbol{x})\big\}. \quad (32)$$

Notice that this choice enforces $d(\boldsymbol{x}, \boldsymbol{x}') > \Delta$ for all pairs $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ in the codebook, regardless of whether $i < j$ or $i > j$.

The exponent in (25) with symmetric distance d' simplifies as follows:

$$\begin{aligned}&\min_{\substack{V: V_X = V_{\widetilde{X}} = P,\\ q(V_{\widetilde{X}Y}) \geq q(V_{XY}),\\ \min\{d(P_{X\widetilde{X}}), d(P_{\widetilde{X}X})\} \geq \Delta}} D(V_{Y|X}\|W|P) + [I(\widetilde{X};X,Y) - R]_+ \\ &\geq \min_{\substack{V: V_X = V_{\widetilde{X}} = P,\\ q(V_{\widetilde{X}Y}) \geq q(V_{XY}),\\ d(P_{X\widetilde{X}}) \geq \Delta}} D(V_{Y|X}\|W|P) + [I(\widetilde{X};X,Y) - R]_+,\end{aligned} \quad (33)$$

since on the right-hand side we are minimizing over a larger set. Moreover, the minimization in the rate condition (27) with distance d' simplifies as follows:

$$\begin{aligned}&\min_{P_{X\widetilde{X}} : P_X = P_{\widetilde{X}} = P, \min\{d(P_{X\widetilde{X}}), d(P_{\widetilde{X}X})\} \leq \Delta} I(X;\widetilde{X}) \\ &= \min\Bigg\{\min_{P_{X\widetilde{X}} : P_X = P_{\widetilde{X}} = P, d(P_{X\widetilde{X}}) \leq \Delta} I(X;\widetilde{X}), \\ &\qquad \min_{P_{X\widetilde{X}} : P_X = P_{\widetilde{X}} = P, d(P_{\widetilde{X}X}) \leq \Delta} I(X;\widetilde{X})\Bigg\} \quad (34) \\ &= \min_{P_{X\widetilde{X}} : P_X = P_{\widetilde{X}} = P, d(P_{X\widetilde{X}}) \leq \Delta} I(X;\widetilde{X}), \quad (35)\end{aligned}$$

where the second line follows since $\min_{z \in A \cup B} f(z) = \min\big\{\min_{z\in A} f(z), \min_{z \in B} f(z)\big\}$, and the last line follows from the symmetry of mutual information. □

We briefly discuss the proof of Theorem 1. While the theorem states the error exponent, the central part of the analysis is in arriving at the following asymptotic expression for the ensemble average probability of error:

$$\begin{aligned}\bar{P}_{\mathrm{e}}^{(n)} \doteq &\sum_{\boldsymbol{x} \in \mathcal{T}(P_n), \boldsymbol{y}} \frac{1}{|\mathcal{T}(P_n)|} W^n(\boldsymbol{y}|\boldsymbol{x}) \\ &\times \min\Bigg\{1, (M_n - 1) \sum_{\substack{\boldsymbol{x}' \in \mathcal{T}(P_n) : q(\boldsymbol{x}', \boldsymbol{y}) \geq q(\boldsymbol{x}, \boldsymbol{y}) \\ d(\boldsymbol{x}', \boldsymbol{x}) \geq \Delta}} \frac{1}{|\mathcal{T}(P_n)|}\Bigg\},\end{aligned} \quad (36)$$

which holds for every type-dependent decoding metric q (not necessarily additive or continuous). This can be interpreted as a stronger (albeit asymptotic) analog of the *random coding*

union bound [23] that achieves not only the random coding exponent, but also the low-rate improvements of the expurgated exponent.

It is also worth discussing the connection of Theorem 1 with the analysis of [3] based on graph decomposition techniques. A key result shown therein is the existence of a rate-R constant composition codebook $\mathcal{C}_n$ such that each $\boldsymbol{x} \in \mathcal{C}_n$ satisfies

$$|\mathcal{T}_{\bar{V}}(\boldsymbol{x}) \cap \mathcal{C}_n| \leq \exp\{n(R - I(P, \bar{V}))\} \tag{37}$$

for all conditional types $\bar{V}$ representing a "channel" from $\mathcal{X} \to \mathcal{X}$, where $I(P, \bar{V}) = I_{P \times \bar{V}}(X; X')$. In the derivation of E_q (*cf.*, (23)), (37) is used to establish the empirical mutual information bound $I_{\hat{P}_{\boldsymbol{x},\boldsymbol{x}'}}(X; X') \leq R$ for any two codewords $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{C}_n$. It is also used to upper bound the number of output sequences $\boldsymbol{y}$ that can give rise to a given joint type $\tilde{P}_{XX'Y}$, with (37) characterizing the $\tilde{P}_{XX}$ marginal and standard techniques characterizing $\tilde{P}_{Y|XX'}$.

Although it was not shown in [3], (37) can be used to establish the achievability part of Theorem 1 for general distance functions. To see this, let $I_{\min}$ be the smallest empirical mutual information among codeword pairs $(\boldsymbol{x}; \boldsymbol{x}')$ with $d(\boldsymbol{x}; \boldsymbol{x}') \leq \Delta$, as stated in Theorem 1. If $R < I_{\min}$, then the left-hand side of (37) is zero, meaning all codeword pairs satisfy $d(\boldsymbol{x}; \boldsymbol{x}') > \Delta$. Upon noticing this fact, the rest of the proof of [3, Theorem 1] remains essentially unchanged and yields the RGV exponent.

Compared to [3] and other related works, the main advantages of our approach are as follows: (i) We provide an explicit recursive random coding construction rather than only proving existence; (ii) We establish, to our knowledge, the first ensemble tightness result for any construction achieving the expurgated exponent; (ii) We provide a direct extension to channels with continuous alphabets, whereas [3] relies heavily on combinatorial arguments and types.

A. Proof of Achievability (Upper Bound on $\bar{P}_{\mathrm{e}}^{(n)}$)

The proof is given in three steps.

Step 1: Characterizing the permitted rates

For convenience, we define

$$R' \triangleq \min_{P_{X\widetilde{X}} \in \mathcal{P}(\mathcal{X}^2) \,:\, d(P_{X\widetilde{X}}) \leq \Delta,\, P_X = P_{\widetilde{X}} = P} I(X; \widetilde{X}) - 2\delta. \tag{38}$$

Recalling that $\mathcal{T}(P_{\widetilde{X}|X})$ stands for a conditional type class [2, Ch. 2] corresponding to $\boldsymbol{x} \in \mathcal{T}(P_n)$, and letting $\mathcal{P}_n(\mathcal{X}|\boldsymbol{x})$ be the set of all conditional types, we have for n sufficiently large that

$$\begin{aligned}
&\sum_{\bar{\boldsymbol{x}} \in \mathcal{T}(P_n) \,:\, d(\bar{\boldsymbol{x}}, \boldsymbol{x}) \leq \Delta} \frac{1}{|\mathcal{T}(P_n)|} \\
&\quad \leq (n+1)^{|\mathcal{X}|^2} \max_{\substack{P_{\widetilde{X}|X} \in \mathcal{P}_n(\mathcal{X}|\boldsymbol{x}) \,:\, P_{\widetilde{X}} = P_X = P_n \\ d(P_{X\widetilde{X}}) \leq \Delta}} \frac{|\mathcal{T}(P_{\widetilde{X}|X})|}{|\mathcal{T}(P_n)|} \qquad (39) \\
&\quad \leq \exp\left(- n \left(\min_{\substack{P_{X\widetilde{X}} \in \mathcal{P}_n(\mathcal{X}^2) \,:\, d(P_{X\widetilde{X}}) \leq \Delta \\ P_X = P_{\widetilde{X}} = P}} I(X; \widetilde{X}) - \delta \right) \right) \qquad (40) \\
&\quad \leq \mathrm{e}^{-n(R' + \delta)}, \qquad (41)
\end{aligned}$$

where (39) follows since the number of conditional types is upper bounded by $(n+1)^{|\mathcal{X}|^2}$, (40) holds for n sufficiently large because $|\mathcal{T}(P_{\widetilde{X}|X})| \doteq \mathrm{e}^{nH_P(\widetilde{X}|X)}$ and $|\mathcal{T}(P_n)| \doteq \mathrm{e}^{nH(P)}$ [2, Ch. 2], and (41) follows from (38) and the fact that $\mathcal{P}_n(\mathcal{X}^2) \subseteq \mathcal{P}(\mathcal{X}^2)$. Hence, if the rate of the generalized RGV construction satisfies $R_n \leq R'$, we have

$$\sum_{\bar{\boldsymbol{x}} \in \mathcal{T}(P_n) \,:\, d(\bar{\boldsymbol{x}}, \boldsymbol{x}) \leq \Delta} \frac{1}{|\mathcal{T}(P_n)|} \leq \mathrm{e}^{-n(R_n + \delta)}, \tag{42}$$

which is precisely the condition assumed in (15).

We henceforth assume that the number of codewords of the generalized RGV construction is such that $R_n \leq R'$, and calculate the resulting average probability of error.

Step 2: Conditional error probability

We define the i-th pairwise error event given $(\boldsymbol{X}_m, \boldsymbol{Y}) = (\boldsymbol{x}_m, \boldsymbol{y})$, where $i \neq m$ as

$$\mathcal{E}_i = \{q(\boldsymbol{X}_i, \boldsymbol{y}) \geq q(\boldsymbol{x}_m, \boldsymbol{y})\}, \tag{43}$$

meaning that the random codeword $\boldsymbol{X}_i$ is favored over $\boldsymbol{x}_m$ (or the two are favored equally). The ensemble average error probability is

$$\bar{P}_e^{(n)} = \frac{1}{M_n} \sum_{m=1}^{M_n} \bar{P}_{e,m}^{(n)}, \tag{44}$$

where the probability of error assuming that the m-th codeword has been transmitted is

$$\bar{P}_{e,m}^{(n)} = \mathbb{E}[\Pr(\text{error} \,|\, \boldsymbol{X}_m, \boldsymbol{Y})], \tag{45}$$

and where

$$\Pr(\text{error} \,|\, \boldsymbol{x}_m, \boldsymbol{y}) = \Pr\left(\bigcup_{\substack{i=1 \\ i \neq m}}^{M_n} \mathcal{E}_i \,\middle|\, \boldsymbol{X}_m = \boldsymbol{x}_m, \boldsymbol{Y} = \boldsymbol{y} \right) \tag{46}$$

is the probability of decoding error for the m-th codeword assuming that the realizations of the codeword and received sequences are $\boldsymbol{x}_m$ and $\boldsymbol{y}$ (recall that ties are counted as errors). We initially perform the analysis conditioned on the transmitted and received sequences being $\boldsymbol{x}_m$ and $\boldsymbol{y}$, respectively (and implicitly on m being transmitted), and later we duly average over these choices.

Now, since only sequences $\boldsymbol{x}_i$ such that $d(\boldsymbol{x}_i, \boldsymbol{x}_m) > \Delta$ have positive probability conditioned on $\boldsymbol{X}_m = \boldsymbol{x}_m$, we have

$$\begin{aligned}
&\Pr(\mathcal{E}_i | \boldsymbol{x}_m, \boldsymbol{y}) \\
&= \sum_{\substack{\boldsymbol{x}_i \,:\, q(\boldsymbol{x}_i, \boldsymbol{y}) \geq q(\boldsymbol{x}_m, \boldsymbol{y}) \\ d(\boldsymbol{x}_i, \boldsymbol{x}_m) > \Delta}} \Pr(\boldsymbol{x}_i | \boldsymbol{x}_m, \boldsymbol{y}) \qquad (47) \\
&= \sum_{\substack{\boldsymbol{x}_i \,:\, q(\boldsymbol{x}_i, \boldsymbol{y}) \geq q(\boldsymbol{x}_m, \boldsymbol{y}) \\ d(\boldsymbol{x}_i, \boldsymbol{x}_m) > \Delta}} \Pr(\boldsymbol{x}_i | \boldsymbol{x}_m) \qquad (48) \\
&= \sum_{\substack{\boldsymbol{x}_i \,:\, q(\boldsymbol{x}_i, \boldsymbol{y}) \geq q(\boldsymbol{x}_m, \boldsymbol{y}) \\ d(\boldsymbol{x}_i, \boldsymbol{x}_m) > \Delta}} \frac{\Pr(\boldsymbol{x}_i, \boldsymbol{x}_m)}{\Pr(\boldsymbol{x}_m)} \qquad (49) \\
&\leq \frac{1}{(1 - \mathrm{e}^{-n\delta})^2} \sum_{\substack{\boldsymbol{x}_i \in \mathcal{T}(P_n) \,:\, q(\boldsymbol{x}_i, \boldsymbol{y}) \geq q(\boldsymbol{x}_m, \boldsymbol{y}) \\ d(\boldsymbol{x}_i, \boldsymbol{x}_m) > \Delta}} \frac{1}{|\mathcal{T}(P_n)|}, \qquad (50)
\end{aligned}$$

where (48) follows since $\boldsymbol{X}_i - \boldsymbol{X}_m - \boldsymbol{Y}$ forms a Markov chain, and (50) follows from Lemmas 2 and 4.

Applying the union bound to (46) and substituting (50), we obtain

$$\begin{aligned}
&\Pr(\text{error}\,|\,\boldsymbol{x}_m,\boldsymbol{y}) \\
&\le \sum_{\substack{i\in\{1,\ldots,M_n\},\\ i\neq m}} \Pr\left(\mathcal{E}_i \,\middle|\, \boldsymbol{X}_m=\boldsymbol{x}_m,\boldsymbol{Y}=\boldsymbol{y}\right) && (51)\\
&\le \frac{1}{(1-\mathrm{e}^{-n\delta})^2} \sum_{\substack{i\in\{1,\ldots,M_n\},\\ i\neq m}} \sum_{\substack{\boldsymbol{x}_i\in\mathcal{T}(P_n):\\ q(\boldsymbol{x}_i,\boldsymbol{y})\ge q(\boldsymbol{x}_m,\boldsymbol{y})\\ d(\boldsymbol{x}_i,\boldsymbol{x}_m)\ge\Delta}} \frac{1}{|\mathcal{T}(P_n)|} && (52)\\
&= (M_n-1)\frac{1}{(1-\mathrm{e}^{-n\delta})^2} \sum_{\substack{\boldsymbol{x}'\in\mathcal{T}(P_n):\\ q(\boldsymbol{x}',\boldsymbol{y})\ge q(\boldsymbol{x}_m,\boldsymbol{y})\\ d(\boldsymbol{x}',\boldsymbol{x}_m)>\Delta}} \frac{1}{|\mathcal{T}(P_n)|}, && (53)
\end{aligned}$$

where (53) follows since summands in the summation over $\boldsymbol{x}_i$ are equal for all i.

Applying the obvious inequality $\Pr(\text{error}\,|\,\boldsymbol{x}_m,\boldsymbol{y})\le 1$, and slightly enlarging the set of summands by replacing $d(\boldsymbol{x}',\boldsymbol{x})>\Delta$ by $d(\boldsymbol{x}',\boldsymbol{x})\ge\Delta$, it follows that

$$\begin{aligned}
\bar{P}_{\mathrm{e}}^{(n)} \doteq& \sum_{\boldsymbol{x}\in\mathcal{T}(P_n),\boldsymbol{y}} \frac{1}{|\mathcal{T}(P_n)|} W^n(\boldsymbol{y}|\boldsymbol{x}) \\
&\times \min\left\{1,(M_n-1)\sum_{\substack{\boldsymbol{x}'\in\mathcal{T}(P_n):\\ q(\boldsymbol{x}',\boldsymbol{y})\ge q(\boldsymbol{x},\boldsymbol{y})\\ d(\boldsymbol{x}',\boldsymbol{x})\ge\Delta}} \frac{1}{|\mathcal{T}(P_n)|}\right\}, && (54)
\end{aligned}$$

where we have averaged over $(\boldsymbol{x}_m,\boldsymbol{y})$ and used Lemma 4.

Step 3: Deducing the error exponent

Deducing the error exponent from (54) amounts to a standard analysis based on the method of types, so we provide a rather brief treatment.

Similarly to (39), the inner sum in (54) satisfies

$$\sum_{\substack{\boldsymbol{x}'\in\mathcal{T}(P_n):\\ q(\boldsymbol{x}',\boldsymbol{y})\ge q(\boldsymbol{x},\boldsymbol{y}),\\ d(\boldsymbol{x}',\boldsymbol{x})\ge\Delta}} \frac{1}{|\mathcal{T}(P_n)|} \doteq \max_{\substack{\hat{P}_{\boldsymbol{x}'|\boldsymbol{x}\boldsymbol{y}}\in\mathcal{P}_n(\mathcal{X}|\boldsymbol{x}\boldsymbol{y}):\\ q(\boldsymbol{x}',\boldsymbol{y})\ge q(\boldsymbol{x},\boldsymbol{y})\\ d(\boldsymbol{x}',\boldsymbol{x})\ge\Delta}} \frac{|\mathcal{T}(\hat{P}_{\boldsymbol{x}'|\boldsymbol{x}\boldsymbol{y}})|}{|\mathcal{T}(P_n)|}. \tag{55}$$

Applying the standard properties of types $|\mathcal{T}(\hat{P}_{\boldsymbol{x}'|\boldsymbol{x}\boldsymbol{y}})| \doteq \mathrm{e}^{nH_{\hat{P}}(\widetilde{X}|Y,X)}$ and $|\mathcal{T}(P_n)| \doteq \mathrm{e}^{nH(P_n)}$ [2, Ch. 2], we can simplify the objective on the right-hand side of (55) to $e^{-nI(\widetilde{X};X,Y)}$. Moreover, we have $W^n(\boldsymbol{y}|\boldsymbol{x}) = \mathrm{e}^{n(D(\hat{P}_{\boldsymbol{y}|\boldsymbol{x}}\|W|P_n)+H(\hat{P}_{\boldsymbol{y}|\boldsymbol{x}}))}$, which implies that $(\boldsymbol{X}_m,\boldsymbol{Y})$ has a given conditional type $V_{Y|X}$ with probability $\mathrm{e}^{-nD(V_{Y|X}\|W|P_n)}$ times a subexponential factor. Using the following continuity lemma to replace P_n by its limiting value P, we deduce the final single-letter exponent:

$$\bar{P}_{\mathrm{e}}^{(n)} \doteq \mathrm{e}^{-n\min_{V\in\mathcal{T}_{d,q,P}(\Delta)} D(V_{Y|X}\|W|P)+|I(\widetilde{X};Y,X)-R|_+}, \tag{56}$$

where $\mathcal{T}_{d,q,P}(\Delta)$ is defined in (26).

Lemma 5. *Consider a DMC W and an input distribution $P\in\mathcal{P}(\mathcal{X})$, along with continuous and bounded d, q and a threshold Δ. For any sequence $P_n\in\mathcal{P}(\mathcal{X})$ with the same support as P such that $P_n(x)\to P(x)$ for all x, we have*

$$\liminf_{n\to\infty} E_{\mathrm{RGV}}(R,P_n,W,q,d,\Delta) \ge E_{\mathrm{RGV}}(R,P,W,q,d,\Delta). \tag{57}$$

Proof. See Appendix D. □

B. Proof of Ensemble Tightness (Lower Bound on $\bar{P}_{\mathrm{e}}^{(n)}$)

We proceed in two steps.

Step 1: Lower bounding the conditional error probability

We shall use the de Caen's lower bound on the probability of a union [24] of events $\{\mathcal{E}_i\}_{i=1}^M$:

$$\Pr\left(\bigcup_{i=1}^M \mathcal{E}_i\right) \ge \sum_{i=1}^M \frac{[\Pr(\mathcal{E}_i)]^2}{\sum_{j=1}^M \Pr(\mathcal{E}_i\cap\mathcal{E}_j)}. \tag{58}$$

Explicitly taking into account the case in which $\Pr(\mathcal{E}_i)$ can be zero for some i values, in which case $\Pr\left(\bigcup_{i=1}^M \mathcal{E}_i\right) = \Pr\left(\bigcup_{i:\Pr(\mathcal{E}_i)>0}\mathcal{E}_i\right)$, we rewrite the de Caen bound as follows:

$$\Pr\left(\bigcup_{i=1}^M \mathcal{E}_i\right) \ge \sum_{\substack{i=1,\\ \Pr(\mathcal{E}_i)>0}}^M \frac{[\Pr(\mathcal{E}_i)]^2}{\Pr(\mathcal{E}_i)+\sum_{j\neq i}\Pr(\mathcal{E}_i\cap\mathcal{E}_j)}. \tag{59}$$

Recalling (44)-(46), and applying (59) to the events $\{\mathcal{E}_i\}_{i=1}^{M_n}$ defined in (43), we obtain

$$\begin{aligned}
&\Pr\left(\bigcup_{\substack{i=1\\ i\neq m}}^{M_n} \mathcal{E}_i \,\middle|\, \boldsymbol{X}_m=\boldsymbol{x}_m,\boldsymbol{Y}=\boldsymbol{y}\right) \\
&\ge \sum_{\substack{i=1,i\neq m,\\ \Pr(\mathcal{E}_i|\boldsymbol{x}_m,\boldsymbol{y})>0}}^{M_n} \frac{[\Pr(\mathcal{E}_i|\boldsymbol{x}_m,\boldsymbol{y})]^2}{\Pr(\mathcal{E}_i|\boldsymbol{x}_m,\boldsymbol{y})+\sum_{j\notin\{i,m\}}\Pr(\mathcal{E}_i\cap\mathcal{E}_j|\boldsymbol{x}_m,\boldsymbol{y})}.
\end{aligned} \tag{60}$$

We first lower bound $\Pr(\mathcal{E}_i|\boldsymbol{x}_m,\boldsymbol{y})$ using (49) along with Lemmas 2 and 4 to obtain

$$\begin{aligned}
&\Pr(\mathcal{E}_i|\boldsymbol{x}_m,\boldsymbol{y}) \\
&\ge (1-4\delta_n^2)\mathrm{e}^{-2\delta_n} \sum_{\substack{\boldsymbol{x}'\in\mathcal{T}(P_n):\, q(\boldsymbol{x}',\boldsymbol{y})\ge q(\boldsymbol{x}_m,\boldsymbol{y})\\ d(\boldsymbol{x}',\boldsymbol{x}_m)>\Delta}} \frac{1}{|\mathcal{T}(P_n)|}. && (61)
\end{aligned}$$

Next, we evaluate $\Pr(\mathcal{E}_i\cap\mathcal{E}_j|\boldsymbol{x}_m,\boldsymbol{y})$. To this end we let $\mathcal{I}_{d,\Delta}(\boldsymbol{x}_i,\boldsymbol{x}_j,\boldsymbol{x}_m)$ denote the indicator of the event of $(\boldsymbol{x}_i,\boldsymbol{x}_j,\boldsymbol{x}_m)$ mutually satisfying the pairwise d-distance constraints; i.e.,

$$\begin{aligned}
&\mathcal{I}_{d,\Delta}(\boldsymbol{x}_i,\boldsymbol{x}_j,\boldsymbol{x}_m) \\
&\quad= \mathbb{1}\{\min\{d(\boldsymbol{x}_i,\boldsymbol{x}_j),d(\boldsymbol{x}_i,\boldsymbol{x}_m),d(\boldsymbol{x}_m,\boldsymbol{x}_j)\}>\Delta\}. && (62)
\end{aligned}$$

From Lemmas 3 and 4, we obtain

$$\begin{aligned}
\Pr(\boldsymbol{x}_i,\boldsymbol{x}_j|\boldsymbol{x}_m) &= \frac{\Pr(\boldsymbol{x}_i,\boldsymbol{x}_j,\boldsymbol{x}_m)}{\Pr(\boldsymbol{x}_m)} && (63)\\
&= \frac{\Pr(\boldsymbol{x}_i,\boldsymbol{x}_m,\boldsymbol{x}_j)}{1/|\mathcal{T}(P_n)|} && (64)\\
&\le \frac{\mathcal{I}_{d,\Delta}(\boldsymbol{x}_i,\boldsymbol{x}_j,\boldsymbol{x}_m)}{(1-e^{-n\delta})^3|\mathcal{T}(P_n)|^2}. && (65)
\end{aligned}$$

Now, since $(\boldsymbol{X}_i, \boldsymbol{X}_j) - \boldsymbol{X}_m - \boldsymbol{Y}$ forms a Markov chain,

$$\begin{aligned}
&\Pr(\mathcal{E}_i \cap \mathcal{E}_j | \boldsymbol{x}_m, \boldsymbol{y}) \\
&= \sum_{\substack{\boldsymbol{x}_i, \boldsymbol{x}_j \colon q(\boldsymbol{x}_i, \boldsymbol{y}) \geq q(\boldsymbol{x}_m, \boldsymbol{y}), \\ q(\boldsymbol{x}_j, \boldsymbol{y}) \geq q(\boldsymbol{x}_m, \boldsymbol{y})}} \Pr(\boldsymbol{x}_i, \boldsymbol{x}_j | \boldsymbol{x}_m, \boldsymbol{y}) && (66) \\
&= \sum_{\substack{\boldsymbol{x}_i, \boldsymbol{x}_j \colon q(\boldsymbol{x}_i, \boldsymbol{y}) \geq q(\boldsymbol{x}_m, \boldsymbol{y}), \\ q(\boldsymbol{x}_j, \boldsymbol{y}) \geq q(\boldsymbol{x}_m, \boldsymbol{y})}} \Pr(\boldsymbol{x}_i, \boldsymbol{x}_j | \boldsymbol{x}_m) && (67) \\
&\dot{\leq} \sum_{\substack{\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{T}(P_n) \colon q(\boldsymbol{x}_i, \boldsymbol{y}) \geq q(\boldsymbol{x}_m, \boldsymbol{y}), \\ q(\boldsymbol{x}_j, \boldsymbol{y}) \geq q(\boldsymbol{x}_m, \boldsymbol{y})}} \frac{\mathcal{I}_{d,\Delta}(\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{x}_m)}{|\mathcal{T}(P_n)|^2} && (68) \\
&\leq \sum_{\substack{\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{T}(P_n) \colon \\ q(\boldsymbol{x}_i, \boldsymbol{y}) \geq q(\boldsymbol{x}_m, \boldsymbol{y}), \\ q(\boldsymbol{x}_j, \boldsymbol{y}) \geq q(\boldsymbol{x}_m, \boldsymbol{y})}} \frac{\mathbb{1}\{\min\{d(\boldsymbol{x}_i, \boldsymbol{x}_m), d(\boldsymbol{x}_m, \boldsymbol{x}_j)\} > \Delta\}}{|\mathcal{T}(P_n)|^2} && (69) \\
&= \sum_{\boldsymbol{x}_i \in \mathcal{T}(P_n) \colon q(\boldsymbol{x}_i, \boldsymbol{y}) \geq q(\boldsymbol{x}_m, \boldsymbol{y})} \frac{\mathbb{1}\{d(\boldsymbol{x}_i, \boldsymbol{x}_m) > \Delta\}}{|\mathcal{T}(P_n)|} \\
&\times \sum_{\boldsymbol{x}_j \in \mathcal{T}(P_n) \colon q(\boldsymbol{x}_j, \boldsymbol{y}) \geq q(\boldsymbol{x}_m, \boldsymbol{y})} \frac{\mathbb{1}\{d(\boldsymbol{x}_j, \boldsymbol{x}_m) > \Delta\}}{|\mathcal{T}(P_n)|} && (70) \\
&\dot{\leq} \Pr(\mathcal{E}_i | \boldsymbol{x}_m, \boldsymbol{y}) \Pr(\mathcal{E}_j | \boldsymbol{x}_m, \boldsymbol{y}). && (71)
\end{aligned}$$

where (68) follows from (65), (69) follows since by definition of $\mathcal{I}_{d,\Delta}(\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{x}_m)$ (see (62)), and (71) follows from (61).

Combining (60) and (71) yields

$$\begin{aligned}
&\Pr\left(\bigcup_{\substack{i=1, \\ i \neq m}}^{M_n} \mathcal{E}_i | \boldsymbol{x}_m, \boldsymbol{y} \right) \\
&\dot{\geq} \sum_{\substack{i=1, i \neq m, \\ \Pr(\mathcal{E}_i | \boldsymbol{x}_m, \boldsymbol{y}) > 0}}^{M_n} [\Pr(\mathcal{E}_i | \boldsymbol{x}_m, \boldsymbol{y})]^2 \Bigg(\Pr(\mathcal{E}_i | \boldsymbol{x}_m, \boldsymbol{y}) \\
&\qquad + \Pr(\mathcal{E}_i | \boldsymbol{x}_m, \boldsymbol{y}) \cdot \sum_{j \notin \{i,m\}} \Pr(\mathcal{E}_j | \boldsymbol{x}_m, \boldsymbol{y}) \Bigg)^{-1} && (72) \\
&= \sum_{\substack{i=1, \\ i \neq m}}^{M_n} \frac{\Pr(\mathcal{E}_i | \boldsymbol{x}_m, \boldsymbol{y})}{1 + \sum_{j \notin \{i,m\}} \Pr(\mathcal{E}_j | \boldsymbol{x}_m, \boldsymbol{y})}, && (73)
\end{aligned}$$

where (73) follows since fixing i we have that if $\Pr(\mathcal{E}_i | \boldsymbol{x}_m, \boldsymbol{y}) = 0$, then obviously the i-th summand on the r.h.s. of (73) is equal to zero and therefore does not affect the summation, and if $\Pr(\mathcal{E}_i | \boldsymbol{x}_m, \boldsymbol{y}) > 0$, this term can be cancelled out from both the numerator and denominator, in which case the i summand on the l.h.s. of (73) is equal to that of the r.h.s. of (73).

Since (50) and (61) imply that

$$\Pr(\mathcal{E}_i | \boldsymbol{x}_m, \boldsymbol{y}) \doteq \sum_{\substack{\boldsymbol{x}' \colon q(\boldsymbol{x}', \boldsymbol{y}) \geq q(\boldsymbol{x}_m, \boldsymbol{y}) \\ d(\boldsymbol{x}', \boldsymbol{x}_m) > \Delta}} \frac{1}{|\mathcal{T}(P_n)|}, \tag{74}$$

letting $\tilde{p}(\boldsymbol{x}_m, \boldsymbol{y})$ denote the right-hand side of (74), we obtain

$$\begin{aligned}
\Pr\left(\bigcup_{\substack{i=1, \\ i \neq m}}^{M_n} \mathcal{E}_i | \boldsymbol{x}_m, \boldsymbol{y} \right) &\dot{\geq} (M_n - 1) \cdot \frac{\tilde{p}(\boldsymbol{x}_m, \boldsymbol{y})}{1 + (M_n - 2)\tilde{p}(\boldsymbol{x}_m, \boldsymbol{y})} && (75) \\
&\geq \frac{(M_n - 1)\tilde{p}(\boldsymbol{x}_m, \boldsymbol{y})}{1 + (M_n - 1)\tilde{p}(\boldsymbol{x}_m, \boldsymbol{y})} && (76) \\
&\geq \frac{1}{2} \min\{1, (M_n - 1)\tilde{p}(\boldsymbol{x}_m, \boldsymbol{y})\}, && (77)
\end{aligned}$$

where the last step follows from the inequality $\frac{x}{1+x} \geq \frac{1}{2}\min\{1, x\}$, which holds for all $x \geq 0$.

Averaging over $(\boldsymbol{x}_m, \boldsymbol{y})$ via Lemma 4, and substituting the definition of $\tilde{p}(\boldsymbol{x}_m, \boldsymbol{y})$, we obtain the lower bound

$$\begin{aligned}
\bar{P}_{e,m}^{(n)} \dot{\geq} &\sum_{\boldsymbol{x} \in \mathcal{T}(P_n), \boldsymbol{y}} \frac{1}{|\mathcal{T}(P_n)|} W^n(\boldsymbol{y}|\boldsymbol{x}) \\
&\times \min\left\{ 1, (M_n - 1) \sum_{\substack{\boldsymbol{x}' \in \mathcal{T}(P_n) \colon q(\boldsymbol{x}', \boldsymbol{y}) \geq q(\boldsymbol{x}, \boldsymbol{y}) \\ d(\boldsymbol{x}', \boldsymbol{x}) > \Delta}} \frac{1}{|\mathcal{T}(P_n)|} \right\}. && (78)
\end{aligned}$$

Step 2: Deducing the error exponent

Applying a similar argument to that used in deriving (56), we obtain from (78) that

$$\begin{aligned}
\bar{P}_{\mathrm{e}}^{(n)} \dot{\geq} \exp\Big(- n \min_{V \in \mathcal{T}_{d,q}^{(n)}(\Delta)} &D(V_{Y|X} \| W | P_n) \\
&+ \big| I(\widetilde{X}; Y, X) - R \big|_+ \Big) && (79)
\end{aligned}$$

where

$$\begin{aligned}
&\mathcal{T}_{d,q,P}^{(n)}(\Delta) \triangleq \Big\{ V_{X\widetilde{X}Y} \in \mathcal{P}_n(\mathcal{X} \times \mathcal{X} \times \mathcal{Y}) : \\
&V_X = V_{\widetilde{X}} = P_n, q(V_{\widetilde{X}Y}) \geq q(V_{XY}), d(P_{X\widetilde{X}}) \geq \Delta \Big\}. && (80)
\end{aligned}$$

Note that this exponent differs from $E_{\mathrm{RGV}}(R, P, W, \phi, d, \Delta)$ only in that the minimization is performed over empirical distributions rather than the probability simplex. The following lemma concludes the proof of ensemble tightness; this is the only part of the analysis where the assumption of additive q is used.

Lemma 6. *Given $P \in \mathcal{P}(\mathcal{X})$ and its corresponding type $P_n \in \mathcal{P}_n(\mathcal{X})$, under any $d \in \Omega$ and additive and bounded metric q, we have for any $\epsilon > 0$ and sufficiently large n that*

$$\begin{aligned}
&\min_{V \in \mathcal{T}_{d,q}^{(n)}(\Delta)} D(V_{Y|X} \| W | P_n) + \big| I(\widetilde{X}; Y, X) - R \big|_+ \\
&\qquad \leq E_{\mathrm{RGV}}(R, P, W, q, d, \Delta + \epsilon) + \epsilon. && (81)
\end{aligned}$$

The proof of Lemma 6 is given in Appendix E.

V. Dual Expression and Continuous Alphabets

In this section, we show that in the case that the distance function d and decoding metric q are additive, the RGV exponent of Theorem 1 permits an equivalent dual expression obtained using Lagrange duality. Moreover, we explain how it

can be derived directly using cost-constrained coding [1], [4], [10], without resorting to constant-composition coding. This approach extends directly to memoryless channels with infinite or continuous alphabets under mild technical assumptions, namely, that all auxiliary cost functions involved have a finite mean with respect to P.

A. Dual expression

We begin by stating the dual form of the RGV exponent and rate condition in Theorem 1. As mentioned above, we focus on additive distances of the form (6), and additive decoding metrics of the form (4)

Theorem 2. *Under the setup of Theorem 1 with an additive distance function d and additive decoding metric q, the error exponent* (25) *can be written as*

$$\begin{aligned} &E_{\mathrm{RGV}}(R,P,W,q,d,\Delta) \\ &= \sup_{\rho\in[0,1],r\ge 0,s\ge 0,a(\cdot)} -\sum_x P(x)\log\sum_y W(y|x) \\ &\quad\times\left(\frac{\sum_{x'}P(x')e^{sq(x',y)}e^{a(x')}e^{r(d(x,x')-\Delta)}}{e^{sq(x,y)}e^{a(x)}}\right)^{\rho} - \rho R, \end{aligned} \tag{82}$$

and rate condition (27) *can be written as*

$$\begin{aligned} R \le \sup_{r\ge 0,a(\cdot)} &-\sum_x P(x) \\ &\times\log\sum_{x'}P(x')e^{a(x')-\phi_a}e^{-r(d(x,x')-\Delta)} - 2\delta, \end{aligned} \tag{83}$$

where $\phi_a = \mathbb{E}_P[a(X)]$.

Proof. The proof uses Lagrange duality analogously to the corresponding statements for the random coding and expurgated exponents [4], [10]; see Appendix F. □

The expression in (82) bears a strong resemblance to the mismatched random coding exponent for constant-composition coding [9]; in fact, the only difference is the presence of additional term $e^{r(d(x,x')-\Delta)}$.

The proof of Theorem 2 does not use the symmetry of d, and hence the equivalence holds even for non-symmetric d as per Corollary 1. The direct derivation below, however, does require a symmetric distance function, but one can still infer the achievability of the exponent for non-symmetric choices via the symmetrization argument used in Corollary 1.

B. Direct derivation via cost-constrained coding

One way of understanding (82) is by noting that it is the exponent that one obtains upon applying Gallager-type bounding techniques, e.g., Markov's inequality and $\min\{1,\alpha\} \le \min_{\rho\in[0,1]}\alpha^{\rho}$, to the asymptotic multi-letter random coding union bound expression in (36) for constant-composition coding. To our knowledge, the "dual analysis" of constant-composition random coding was initiated by Poltyrev [25].

The preceding approach permits continuous channel outputs, but requires discrete inputs. It turns out, however, that we can attain an analog of (36) 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 this approach is that it is difficult to claim ensemble tightness; we provide only achievability results.

1) Code construction: Fix an input distribution P and four auxiliary costs $a_1(x),\dots,a_4(x)$. 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

$$P_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{\mu}P^n(\boldsymbol{x})\mathbb{1}\Big\{\big|a_j(\boldsymbol{x})-\phi_j\big|\le\epsilon,\quad j=1,2,3,4\Big\}, \tag{84}$$

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 functions a_j represent *auxiliary costs* that 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. We observe that μ is the probability (under P^n) of the event in the indicator function of (84) occurring, and we immediately obtain

$$\lim_{n\to\infty}\mu = 1 \tag{85}$$

by the law of large numbers.

With the definition of $P_{\boldsymbol{X}}$ in place, we recursively generate the codewords in a similar manner to Section III:

$$\Pr(\boldsymbol{x}_1) = P_{\boldsymbol{X}}(\boldsymbol{x}_1) \tag{86}$$

$$\Pr(\boldsymbol{x}_2|\boldsymbol{x}_1) = \frac{1}{\mu_2(\boldsymbol{x}_1)}P_{\boldsymbol{X}}(\boldsymbol{x}_2)\mathbb{1}\big\{d(\boldsymbol{x}_1,\boldsymbol{x}_2)>\Delta\big\} \tag{87}$$

$$\vdots$$

$$\begin{aligned} \Pr(\boldsymbol{x}_m|\boldsymbol{x}_1^{m-1}) = \frac{1}{\mu_m(\boldsymbol{x}_1^{m-1})}&P_{\boldsymbol{X}}(\boldsymbol{x}_m) \\ &\times\mathbb{1}\big\{d(\boldsymbol{x}_i,\boldsymbol{x}_m)>\Delta\ \ \forall i<m\big\}, \end{aligned} \tag{88}$$

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 $f(\boldsymbol{x}_i\,|\,\cdot)$.

We proceed by describing the analysis in two steps. To avoid repetition, we omit certain parts of the analysis that are the same as the constant-composition case.

2) Key properties: Similarly to the constant-composition case, we seek to arrive at an upper bound of the form

$$\begin{aligned} \bar{P}_{\mathrm{e}}^{(n)} \dot{\le} &\sum_{\boldsymbol{x},\boldsymbol{y}} P_{\boldsymbol{X}}(\boldsymbol{x})W^n(\boldsymbol{y}|\boldsymbol{x}) \\ &\times\min\Bigg\{1,(M_n-1)\sum_{\substack{\boldsymbol{x}':\,q(\boldsymbol{x}',\boldsymbol{y})\ge q(\boldsymbol{x},\boldsymbol{y})\\ d(\boldsymbol{x}',\boldsymbol{x})\ge\Delta}} P_{\boldsymbol{X}}(\boldsymbol{x}')\Bigg\} \end{aligned} \tag{89}$$

that holds under the rate condition (83). Towards establishing this bound, we prove the following four important properties:

(a) For any $\boldsymbol{x}$ such that $P_{\boldsymbol{X}}(\boldsymbol{x}) > 0$, we have under $\boldsymbol{X}' \sim P_{\boldsymbol{X}}$ that

$$-\frac{1}{n}\log \Pr\left(d(\boldsymbol{x},\boldsymbol{X}') \le \Delta\right) \ge \sup_{r\ge 0, a(\cdot)} -\sum_x P(x) \times \log \sum_{x'} P(x') e^{a(x')-\phi_a} e^{-r(d(x,x')-\Delta)} - \delta, \quad (90)$$

thus matching the rate condition in (83).

(b) The normalizing constants in (86)–(88) satisfy $\mu_m(\boldsymbol{X}_1^{m-1}) \ge 1 - \mathrm{e}^{-n\delta}$ almost surely under the rate condition (83), for any choice of $\delta > 0$.

(c) The marginal distribution of any given codeword (indexed by m) satisfies $\Pr(\boldsymbol{x}_m) \doteq P_{\boldsymbol{X}}(\boldsymbol{x}_m)$.

(d) The marginal joint distribution of any two codewords (indexed by k and m) satisfies $\Pr(\boldsymbol{x}_k, \boldsymbol{x}_m) \dot{\le} P_{\boldsymbol{X}}(\boldsymbol{x}_k) P_{\boldsymbol{X}}(\boldsymbol{x}_m) \mathbb{1}\{d(\boldsymbol{x}_k, \boldsymbol{x}_m) > \Delta\}$.

As we outline below, the first two properties are used as stepping stones to obtaining the final two. Once the final two properties are established, then a near-identical analysis to that of (43)–(54) yields (89).

To establish the first property (90), we bound the probability therein for fixed $\boldsymbol{x}$:

$$\begin{aligned}
&\Pr\left(d(\boldsymbol{x},\boldsymbol{X}') \le \Delta\right) \\
&= \sum_{\boldsymbol{x}'} P_{\boldsymbol{X}}(\boldsymbol{x}') \mathbb{1}\{d(\boldsymbol{x},\boldsymbol{x}') \le \Delta\} && (91)\\
&\le \sum_{\boldsymbol{x}'} P_{\boldsymbol{X}}(\boldsymbol{x}') \mathrm{e}^{-nr(d(\boldsymbol{x},\boldsymbol{x}')-\Delta)} && (92)\\
&\le \sum_{\boldsymbol{x}'} P_{\boldsymbol{X}}(\boldsymbol{x}') \mathrm{e}^{-nr(d(\boldsymbol{x},\boldsymbol{x}')-\Delta)} \mathrm{e}^{n(a_1(\boldsymbol{x}')-\phi_1+\epsilon)} && (93)\\
&\le \sum_{\boldsymbol{x}'} P^n(\boldsymbol{x}') \mathrm{e}^{-nr(d(\boldsymbol{x},\boldsymbol{x}')-\Delta)} \mathrm{e}^{n(a_1(\boldsymbol{x}')-\phi_1+2\epsilon)}, && (94)
\end{aligned}$$

where (92) uses Markov inequality with an arbitrary parameter $r > 0$, (93) uses the fact that $a_1(\boldsymbol{x}') \ge \phi_1 - \epsilon$ by construction, and (94) holds for sufficiently large n because $\mu \to 1$ in (84). Taking the logarithm and applying Gallager's single-letterization argument [1], we get

$$\begin{aligned}
&-\log \Pr\left(d(\boldsymbol{x},\boldsymbol{X}') \le \Delta\right) \\
&\ge -\sum_{k=1}^{n} \log \sum_{x'} P(x') \mathrm{e}^{-r(d(x_k,x')-\Delta)} \mathrm{e}^{a_1(x')-\phi_1} - 2n\epsilon. && (95)
\end{aligned}$$

We now choose $a_2(x) = -\log \sum_{x'} P(x') \mathrm{e}^{r(d(x,x')-\Delta)} \mathrm{e}^{a_1(x')-\phi_1}$, which ensures that the leading term on the right-hand side of (95) is equal to $n a_2^n(\boldsymbol{x})$. 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

$$\begin{aligned}
&-\frac{1}{n}\log \Pr\left(d(\boldsymbol{x},\boldsymbol{X}') \le \Delta\right) \\
&\ge -\sum_x P(x) \log \sum_{x'} P(x') \mathrm{e}^{-r(d(x,x')-\Delta)} \mathrm{e}^{a_1(x')-\phi_1} - 3\epsilon. && (96)
\end{aligned}$$

Choosing $\epsilon = \frac{\delta}{3}$ and optimizing r and $a_1(\cdot)$, we obtain (90), thus completing the proof of the first property above.

The second property above follows easily from the first: 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

$$\begin{aligned}
1 - \mu_m(\boldsymbol{x}_1^{m-1}) &\le \sum_{i<m} \Pr\left(d(\boldsymbol{x}_i,\boldsymbol{X}') \le \Delta\right) && (97)\\
&\le e^{nR_n} \Pr\left(d(\boldsymbol{x}_i,\boldsymbol{X}') \le \Delta\right) && (98)\\
&\le e^{-n\delta}, && (99)
\end{aligned}$$

where (99) follows from (90) and the rate condition (83).

Upper bounding the indicator functions in (86)–(88) 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 third property above. The other direction requires more effort, and is deferred to Appendix G.

For the forth property above, we use (88) and the fact that $\mu_m(\boldsymbol{x}_1^{m-1}) \ge 1 - e^{-n\delta}$ to obtain

$$\begin{aligned}
&\Pr(\boldsymbol{x}_k, \boldsymbol{x}_m) \\
&= \sum_{\boldsymbol{x}_1^{k-1}, \boldsymbol{x}_{k+1}^{m-1}} \Pr(\boldsymbol{x}_1^{k-1}) \Pr(\boldsymbol{x}_k|\boldsymbol{x}_1^{k-1}) \\
&\quad \times \Pr(\boldsymbol{x}_{k+1}^{m-1}|\boldsymbol{x}_1^{k}) \Pr(\boldsymbol{x}_m|\boldsymbol{x}_1^{m-1}) && (100)\\
&\le \sum_{\boldsymbol{x}_1^{k-1}, \boldsymbol{x}_{k+1}^{m-1}} \Pr(\boldsymbol{x}_1^{k-1}) \cdot \frac{P_{\boldsymbol{X}}(\boldsymbol{x}_k)}{1-e^{-n\delta}} \cdot \Pr(\boldsymbol{x}_{k+1}^{m-1}|\boldsymbol{x}_1^{k}) \\
&\quad \times \frac{P_{\boldsymbol{X}}(\boldsymbol{x}_m)\mathbb{1}\{d(\boldsymbol{x}_k,\boldsymbol{x}_m) > \Delta\}}{1-e^{-n\delta}} && (101)\\
&= \frac{1}{(1-e^{-n\delta})^2} P_{\boldsymbol{X}}(\boldsymbol{x}_k) P_{\boldsymbol{X}}(\boldsymbol{x}_m) \mathbb{1}\{d(\boldsymbol{x}_k,\boldsymbol{x}_m) > \Delta\}. && (102)
\end{aligned}$$

3) Upper-bounding the multi-letter upper bound: Once (89) is established, the steps in deriving (82) are standard. Such an analysis requires two additional auxiliary costs, and these are given by a_3 and a_4 in (84). In particular, we set $a_3(x) = a(x)$ in (82) and

$$a_4(x) = -\log \sum_y W(y|x) \times \left(\frac{\sum_{x'} P(x') e^{sq(x',y)} e^{a(x')} e^{r(d(x,x')-\Delta)}}{e^{sq(x,y)} e^{a(x)}}\right)^{\rho}. \quad (103)$$

In fact, removing the constraint $d(\boldsymbol{x},\boldsymbol{x}') > \Delta$ from the pairwise error probability term in (89) recovers the standard random-coding union bound, which was already used in [10] to establish the exponent in (82) 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 the inequality $\mathbb{1}\{d(\boldsymbol{x},\boldsymbol{x}') \ge \Delta\} \le \mathrm{e}^{nr(d(\boldsymbol{x},\boldsymbol{x}')-\Delta)}$, similarly to (91). Due to this similarity, the details are omitted.

VI. Optimal Distance Functions

In this section, we study optimal choices for the distance function $d(\cdot,\cdot)$ in Theorem 1, thus characterizing how the codewords should be separated in order to get the best possible exponent for our construction. While some of the analysis in this section includes the constant $\delta > 0$, the best exponent will always be obtained in the limit as $\delta \to 0$.

A. Reduction to the Csiszár-Körner Exponent - Negative Mutual Information Distance

We show that when the distance function $d(\cdot,\cdot)$ is optimized, and Δ is chosen appropriately, the exponent in Theorem 1 recovers the exponent $E_q(R,P,W)$ in (23) [3].

Corollary 2. *Let $\epsilon > 0$ be given, let $q(\cdot)$ be an arbitrary type-dependent continuous decoding rule, and let R, P, and $d \in \Omega$ be given. The exponent of the ensemble average error probability of the generalized RGV construction with sufficiently small δ, $d(P_{X\widetilde{X}}) = -I(X;\widetilde{X})$, $\Delta = -(R+2\delta)$, sufficiently large n, and decoding metric $q(\cdot)$ over the DMC W is at least as high as $E_q(R,P,W) - \epsilon$.*

Proof. We claim that the choices $d(P_{X\widetilde{X}}) = -I(X;\widetilde{X})$ and $\Delta = -(R+2\delta)$ are valid for all R in the sense of satisfying the rate condition (27). To see this, note that

$$\begin{aligned}
&\min_{\substack{P_{X\widetilde{X}}:\, d(P_{X\widetilde{X}})\le\Delta \\ P_X = P_{\widetilde{X}} = P}} I(X;\widetilde{X})\Big|_{\substack{d(P_{X\widetilde{X}})=-I(X;\widetilde{X}) \\ \Delta=-(R+2\delta)}} \\
&\quad = \min_{\substack{P_{X\widetilde{X}}:\, I(X;\widetilde{X})\ge R+2\delta \\ P_X = P_{\widetilde{X}} = P}} I(X;\widetilde{X}) \qquad (104)\\
&\quad \ge R + 2\delta, \qquad (105)
\end{aligned}$$

as required. Now, under the same choices, we have

$$\begin{aligned}
&E_{\mathrm{RGV}}(R,P,W,q,d,\Delta)\Big|_{d(P_{X\widetilde{X}})=-I(X;\widetilde{X}),\,\Delta=-(R+2\delta)} \\
&\quad = \min_{V\in\mathcal{T}_{I,\delta}} D(V_{Y|X}\|W|P) + \big|I(\widetilde{X};Y,X) - R\big|_+, \qquad (106)
\end{aligned}$$

where

$$\mathcal{T}_{I,\delta} \triangleq \left\{ \begin{array}{l} V_{X\widetilde{X}Y} \in \mathcal{P}(\mathcal{X}\times\mathcal{X}\times\mathcal{Y}) : V_X = V_{\widetilde{X}} = P, \\ q(V_{\widetilde{X}Y}) \ge q(V_{XY}), I(\widetilde{X};X) \le R + 3\delta \end{array} \right\}. \qquad (107)$$

The result follows by taking $\delta \to 0$ and using the continuity of $E_q(R,P,W)$ in R [3]. □

The following proposition reveals that the above choice of (d,Δ) is the one that maximizes the general exponent given in Theorem 1.

Proposition 1. *Under the setup of Theorem 1 with*

$$R \le \min_{\substack{P_{X\widetilde{X}}:\, P_X = P_{\widetilde{X}} = P \\ d(P_{X\widetilde{X}})\le\Delta}} I(X;\widetilde{X}) - 2\delta, \qquad (108)$$

we have

$$\begin{aligned}
&E_{\mathrm{RGV}}(R,P,W,q,d,\Delta) \\
&\le E_{\mathrm{RGV}}(R,P,W,q,d,\Delta)\Big|_{d=-I(X;\widetilde{X}),\,\Delta=-(R+2\delta)}. \qquad (109)
\end{aligned}$$

Proof. From (108), we see that among all $P'_{X\widetilde{X}}$ such that $P'_X = P'_{\widetilde{X}} = P$, the condition $d(P'_{X\widetilde{X}}) \le \Delta$ implies $R + 2\delta \le I_{P'}(X;\widetilde{X})$. The contrapositive statement is that among all $P'_{X\widetilde{X}}$ such that $P'_X = P'_{\widetilde{X}} = P$, the condition $R+2\delta > I_{P'}(X;\widetilde{X})$ implies $d(P'_{X\widetilde{X}}) > \Delta$. As a result, when (108) holds, $\mathcal{T}_{d,q,P}(\Delta)$ defined in (26) satisfies

$$\mathcal{T}_{d,q,P}(\Delta) \supseteq \mathcal{T}_{I,\delta}, \qquad (110)$$

where $\mathcal{T}_{I,\delta}$ is defined in (107). Therefore,

$$\begin{aligned}
&E_{\mathrm{RGV}}(R,P,W,q,d,\Delta) \\
&\quad = \min_{V\in\mathcal{T}_{d,q,P}(\Delta)} D(V_{Y|X}\|W|P) + \big|I(\widetilde{X};Y,X) - R\big|_+ \qquad (111)\\
&\quad \le \min_{V\in\mathcal{T}_{I,\delta}} D(V_{Y|X}\|W|P) + \big|I(\widetilde{X};Y,X) - R\big|_+, \qquad (112)
\end{aligned}$$

so the exponent is upper bounded by that corresponding to $d(P_{X\widetilde{X}}) = -I(X;\widetilde{X})$ and $\Delta = -(R+2\delta)$. □

We note that the choice $d(P_{X\widetilde{X}}) = -I(X;\widetilde{X})$ is *universally optimal* in maximizing the achievable exponent in Theorem 1 (subject to (27)), in the sense that it has no dependence on the channel, decoding rule, or input distribution. This provides an interesting analogy with the decoding rule $q(P_{XY}) = I(X;Y)$, which is known to be universally optimal for achieving the regular random-coding exponent; however, it remains an open problem as to whether such a choice also attains the expurgated exponent [3].

B. A Non-Universal Optimal Distance Function

In this subsection, we show that the non-universal distance function $d(P_{X\tilde{X}}) = \beta_{R,W,q}(P_{X\tilde{X}})$ also achieves the exponent of Csiszár and Körner, where

$$\beta_{R,W,q}(P_{X\tilde{X}}) \triangleq \min_{V_{X\tilde{X}Y}\in\mathcal{T}'(P_{X\tilde{X}})} \Gamma(V_{X\tilde{X}Y}), \qquad (113)$$

with[2]

$$\Gamma(V_{X\tilde{X}Y}) \triangleq D(V_{Y|X}\|W|V_X) + \big|I(\widetilde{X};Y,X) - R\big|_+, \qquad (114)$$

and

$$\begin{aligned}
\mathcal{T}'(P_{X\tilde{X}}) \triangleq \Big\{ &V_{X\tilde{X}Y} \in \mathcal{P}(\mathcal{X}\times\mathcal{X}\times\mathcal{Y}) : \\
&V_{X\tilde{X}} = P_{X\tilde{X}}, q(V_{\tilde{X}Y}) \ge q(V_{XY}) \Big\}. \qquad (115)
\end{aligned}$$

We first provide a corollary characterizing the exponent of Theorem 1 with $d(\cdot) = \beta_{R,W,q}(\cdot)$, and then prove its equivalence to (23).

Corollary 3. *If the pair (R,Δ) satisfies the condition*

$$R \le \min_{\substack{P_{X\widetilde{X}}:\, P_X = P_{\widetilde{X}} = P \\ \beta_{R,W,q}(P_{X\widetilde{X}})\le\Delta}} I(X;\widetilde{X}) - 2\delta, \qquad (116)$$

then the ensemble average error probability $\bar{P}_{\mathrm{e}}^{(n)}$ of the RGV codebook construction with parameters $(n,R,P,\beta_{R,W,q},\Delta,\delta)$ using the continuous type-dependent decoding rule $q(\cdot)$ over the channel W satisfies

$$\bar{P}_{\mathrm{e}}^{(n)} \dot{\le} \mathrm{e}^{-n\Delta}. \qquad (117)$$

Proof. First observe that the minimization in $E_{\mathrm{RGV}}(R,P,W,q,d,\Delta)$ (see (25)) can be done in two stages: Minimize first over $P_{X\widetilde{X}}$, and then over $V_{X\widetilde{X}Y}$ that are consistent with $P_{X\widetilde{X}}$. By doing so, we obtain

$$\begin{aligned}
&E_{\mathrm{RGV}}(R,P,W,q,d,\Delta) \\
&= \min_{\substack{P_{X\widetilde{X}}:\, P_X = P_{\widetilde{X}} = P \\ d(P_{X\widetilde{X}})\ge\Delta}} \min_{V_{X\widetilde{X}Y}\in\mathcal{T}'(P_{X\widetilde{X}})} D(V_{Y|X}\|W|P)
\end{aligned}$$

[2]The dependence of Γ on (R,W,q) is left implicit to lighten notation.

$$+\left|I(\widetilde{X};Y,X)-R\right|_+, \quad (118)$$

where $\mathcal{T}'(P_{X\widetilde{X}})$ is defined in (115). From the definition of $\beta_{R,W,q}(P_{X\widetilde{X}})$ (113), we can rewrite (118) as

$$E_{\mathrm{RGV}}(R,P,W,q,d,\Delta)=\min_{\substack{P_{X\widetilde{X}}:\,P_X=P_{\widetilde{X}}=P\\ d(P_{X\widetilde{X}})\ge\Delta}}\beta_{R,W,q}(P_{X\widetilde{X}}). \quad (119)$$

Hence, by the choice $d(\cdot)=\beta_{R,W,q}(\cdot)$ we obtain

$$E_{\mathrm{RGV}}(R,P,W,q,d,\Delta)=\min_{\substack{P_{X\widetilde{X}}:\,P_X=P_{\widetilde{X}}=P\\ d(P_{X\widetilde{X}})\ge\Delta}} d(P_{X\widetilde{X}}) \quad (120)$$

$$\ge\Delta. \quad (121)$$

Combined with (27), this yields that for a pair (R,Δ) that satisfies (116), we have $\bar{P}_{\mathrm{e}}^{(n)}\dot{\le}\,\mathrm{e}^{-n\Delta}$. □

Note that while the preceding proof gives an exponent of Δ, one cannot make Δ arbitrarily large, because past a certain point the condition (116) will never be satisfied.

The following proposition shows that error exponents corresponding to Corollaries 2 and 3 are identical, and hence, both are optimal when (27) holds.

Proposition 2. *For any $P\in\mathcal{P}(\mathcal{X})$, the achievable rate-exponent pairs (R,E) resulting from Theorem 1 (i.e., taking the union over all $\delta>0$ and $\Delta>0$) are identical for the choices $d(P_{X\widetilde{X}})=-I(X;\widetilde{X})$ and $d(P_{X\widetilde{X}})=\beta_{R,W,q}(P_{X\widetilde{X}})$.*

Proof. Consider the exponent in Corollary 2 for $d(P_{X\widetilde{X}})=-I(X;\widetilde{X})$. For fixed R, the highest possible exponent E is obtained by choosing Δ such that (27) holds with equality, and then taking $\delta\to0$ to obtain the achievable pair

$$(R,E)=\left(R,\min_{V_{X\widetilde{X}Y}\in\mathcal{T}_I} D(V_{Y|X}\|W|P)+\left|I(\widetilde{X};Y,X)-R\right|_+\right), \quad (122)$$

where

$$\mathcal{T}_I\triangleq\Big\{V_{X\widetilde{X}Y}\in\mathcal{P}(\mathcal{X}\times\mathcal{X}\times\mathcal{Y}):\ V_X=V_{\widetilde{X}}=P, q(P_{\widetilde{X},Y})\ge q(P_{X,Y}), I(X;\widetilde{X})\le R\Big\}. \quad (123)$$

Next, Corollary 3 states that Δ is an achievable exponent at rate R for $d(P_{X\widetilde{X}})=\beta_{R,W,q}(P_{X\widetilde{X}})$ provided that

$$R<\min_{P_{X\widetilde{X}}:\,P_X=P_{\widetilde{X}}=P,\beta_{R,W,q}(P_{X\widetilde{X}})\le\Delta} I(X;\widetilde{X}). \quad (124)$$

The condition $\beta_{R,W,q}(P_{X\widetilde{X}})\le\Delta$ is equivalent to:

$$\beta_{R,W,q}(P_{X\widetilde{X}})\le\Delta \iff \min_{V_{X\widetilde{X}Y}:\,q(P_{\widetilde{X},Y})\ge q(P_{X,Y}),V_{X\widetilde{X}}=P_{X\widetilde{X}}}\Gamma(V_{X\widetilde{X}Y})\le\Delta \quad (125)$$

$$\iff \Gamma(V_{X\widetilde{X}Y})\le\Delta \text{ for some } V_{X\widetilde{X}Y} \text{ s.t. } q(P_{\widetilde{X},Y})\ge q(P_{X,Y}), V_{X\widetilde{X}}=P_{X\widetilde{X}}. \quad (126)$$

Using this, we can rewrite the right-hand side of (124) as

$$\min_{P_{X\widetilde{X}}:\,P_X=P_{\widetilde{X}}=P,\beta_{R,W,q}(P_{X\widetilde{X}})\le\Delta} I(X;\widetilde{X})$$

$$=\min_{\substack{P_{X\widetilde{X}}:\,P_X=P_{\widetilde{X}}=P,\\ \Gamma(V_{X\widetilde{X}Y})\le\Delta \text{ for some } V_{X\widetilde{X}Y}:\ \substack{q(P_{\widetilde{X},Y})\ge q(P_{X,Y}),\\ V_{X\widetilde{X}}=P_{X\widetilde{X}}}}} I(X;\widetilde{X}) \quad (127)$$

$$=\min_{P_{X\widetilde{X}}:\,P_X=P_{\widetilde{X}}=P}\ \min_{V_{X\widetilde{X}Y}:\,q(P_{\widetilde{X},Y})\ge q(P_{X,Y})} \begin{cases} I_P(X;\widetilde{X}) & \Gamma(V_{X\widetilde{X}Y})\le\Delta \text{ and } V_{X\widetilde{X}}=P_{X\widetilde{X}}\\ \infty & \text{otherwise}\end{cases} \quad (128)$$

$$=\min_{\substack{V_{X\widetilde{X}Y}:\,q(P_{\widetilde{X},Y})\ge q(P_{X,Y}),\\ \Gamma(V_{X\widetilde{X}Y})\le\Delta, V_X=V_{\widetilde{X}}=P}} I_V(X;\widetilde{X}), \quad (129)$$

where the last step uses the fact that $I_P(X;\widetilde{X})=I_V(X;\widetilde{X})$ whenever $V_{X\widetilde{X}}=P_{X\widetilde{X}}$. From (129), it follows that (124) can be written as

$$R<\min_{V_{X\widetilde{X}Y}\in\mathcal{V}:\,\Gamma(V_{X\widetilde{X}Y})\le\Delta} I_V(X;\widetilde{X}), \quad (130)$$

where $\mathcal{V}=\{V_{X\widetilde{X}Y}:\,q(P_{\widetilde{X},Y})\ge q(P_{X,Y}), V_X=V_{\widetilde{X}}=P\}$. We claim that (130) is equivalent to

$$\Delta<\min_{V_{X\widetilde{X}Y}\in\mathcal{V}:\,I_V(X;\widetilde{X})\le R}\Gamma(V_{X\widetilde{X}Y}). \quad (131)$$

To see this, we show that (130) implies (131), and that the complement of (130) implies the complement of (131):

- First suppose that (130) holds. This means that within $\mathcal{V}$ we have $\Gamma(V_{X\widetilde{X}Y})\le\Delta \implies R<I_V(X;\widetilde{X})$, and the contrapositive statement is that within $\mathcal{V}$ we have $R\ge I_V(X;\widetilde{X}) \implies \Gamma(V_{X\widetilde{X}Y})>\Delta$, which implies (131).
- Now suppose that (130) fails. This means that there exists $V\in\mathcal{V}$ such that $\Gamma(V_{X\widetilde{X}Y})\le\Delta$ and $R\ge I_V(X;\widetilde{X})$, which implies that (131) fails.

Finally, we note that the right-hand side of (131) is precisely $E_{\mathrm{RGV}}(R,P,W,q,d,\Delta)|_{d=-I(X;\widetilde{X})}$ (see (25)), and we recall that Δ equals the achievable exponent for $d(P_{X\widetilde{X}})=\beta_{R,W,q}(P_{X\widetilde{X}})$. Thus, (131) states that given R, this exponent can be made arbitrarily close to $E_q(R,P,W)$ in (23). Since the latter is optimal by Proposition 1, the proof is complete. □

C. Bhattacharyya and Chernoff Distances

Here we show that an additive distance function with per-letter distance

$$d_s(x,x')=-\log\sum_y W(y|x)\left(\frac{\mathrm{e}^{q(x',y)}}{\mathrm{e}^{q(x,y)}}\right)^s, \quad (132)$$

for suitably-chosen $s>0$ also recovers the maximum of the random coding and expurgated exponents. We call this the *Chernoff distance*, because it is closely related to the Chernoff bound for bounding a probability of the event of the form $\{q(\boldsymbol{X}',\boldsymbol{y})\ge q(\boldsymbol{x},\boldsymbol{y})\}$. In the case of ML decoding $q(x,y)=\log W(y|x)$, we choose $s=\frac{1}{2}$, and hence d_s reduces to the Bhattacharrya distance, which is symmetric. For general decoding metrics, we may require $s\ne\frac{1}{2}$, and thus d_s is not symmetric; however, the RGV exponent is still achievable according to Corollary 1.

We note that since we are considering bounded metrics, the distance $d_s(x,x')$ is also bounded, in accordance with

Definition 1. However, this may rule out certain choices such as $q(x,y) = \log W(y|x)$ for channels with zero-probability transitions, in which we wish to assign the value $q(x,y) = -\infty$ when $W(y|x) = 0$.

We will show that the additive distance d_s recovers both the random coding and expurgated exponents for mismatched decoding [3], [4]. This implies the *near-optimality* of d_s, in the sense that no examples are known for which $E_q(R,P,W)$ is strictly higher than the maximum of the random-coding and expurgated exponents.

Recovering the (ensemble-tight) random coding exponent is immediate: By setting Δ equal its maximum possible value, the rate condition in (27) becomes trivial, and we can lower bound the exponent in (25) by dropping the constraint $d(P_{X\widetilde{X}}) \geq \Delta$ and writing $I(\widetilde{X};Y,X) \geq I(\widetilde{X};Y)$. The resulting exponent matches that of [2], [10]. Alternatively, setting $r = 0$ in (82) gives the same exponent in the dual form.

Recovering the expurgated exponent is more difficult; we do this using the dual form in Theorem 2. Setting $\rho = 1$ in (82), and letting s coincide with the choice in (132), we obtain

$$\begin{aligned} &E_{\mathrm{RGV}}(R,P,W,q,d,\Delta) \\ &\geq -\sum_x P(x)\log\sum_{x'}P(x')\sum_y W(y|x) \\ &\quad\times\Big(\frac{\mathrm{e}^{q(x',y)}}{\mathrm{e}^{q(x,y)}}\Big)^s\frac{\mathrm{e}^{a(x')}}{\mathrm{e}^{a(x)}}e^{r(d(x,x')-\Delta)} - R \qquad (133)\\ &= -\sum_x P(x)\log\sum_{x'}P(x')\mathrm{e}^{-d_s(x,x')}\frac{\mathrm{e}^{a(x')}}{\mathrm{e}^{a(x)}}\mathrm{e}^{r(d(x,x')-\Delta)} - R. \qquad (134)\end{aligned}$$

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

$$\begin{aligned} &E_{\mathrm{RGV}}(R,P,W,q,d,\Delta) \\ &\geq -\sum_x P(x)\log\sum_{x'}P(x')\mathrm{e}^{-\frac{d_s(x,x')}{1+\rho'}}\frac{\mathrm{e}^{a(x')}}{\mathrm{e}^{a(x)}} + \Delta\frac{\rho'}{1+\rho'} - R. \qquad (135)\end{aligned}$$

Then, choosing

$$\Delta = -(1+\rho')\bigg(\sum_x P(x)\log\bigg[\sum_{x'}P(x')\mathrm{e}^{-\frac{d_s(x,x')}{1+\rho'}}\frac{\mathrm{e}^{a(x')}}{\mathrm{e}^{a(x)}}\bigg] + R + 2\delta\bigg), \qquad (136)$$

we obtain from (135) that

$$\begin{aligned} &E_{\mathrm{RGV}}(R,P,W,q,d,\Delta) \\ &\geq -\sum_x P(x)\log\sum_{x'}P(x')\mathrm{e}^{-\frac{d_s(x,x')}{1+\rho'}}\frac{\mathrm{e}^{a(x')}}{\mathrm{e}^{a(x)}} \\ &\quad-\rho'\bigg(\sum_x P(x)\log\sum_{x'}P(x')\mathrm{e}^{-\frac{d_s(x,x')}{1+\rho'}}\frac{\mathrm{e}^{a(x')}}{\mathrm{e}^{a(x)}} + R + 2\delta\bigg) - R \qquad (137)\\ &= -(1+\rho')\bigg(\sum_x P(x)\log\sum_{x'}P(x')\mathrm{e}^{-\frac{d_s(x,x')}{1+\rho'}}\frac{\mathrm{e}^{a(x')}}{\mathrm{e}^{a(x)}}\bigg) \\ &\quad-(1+\rho'+2\delta\rho')R. \qquad (138)\end{aligned}$$

Upon taking $\delta \to 0$ and optimizing over $\rho' \geq 0$, $s \geq 0$, and $a(\cdot)$, this exponent is identical to the dual form for the mismatched decoding expurgated exponent given in [4], which is known to be equivalent to the primal form given in [3].

We also need to check that the choice of Δ in (136) complies with the rate condition in (83). We choose the same $a(\cdot)$ as in the exponent, but a value different r (note that the two need not be the same). We simplify the condition as follows:

$$\begin{aligned} R &\leq -\sum_x P(x)\log\sum_{x'}P(x')\mathrm{e}^{a(x')-\phi_a}\mathrm{e}^{-r(d_s(x,x')-\Delta)} - 2\delta \qquad (139)\\ &= -\sum_x P(x)\log\sum_{x'}P(x')\mathrm{e}^{a(x')-\phi_a}\mathrm{e}^{-rd_s(x,x')} - r\Delta - 2\delta \qquad (140)\\ &= -\sum_x P(x)\log\sum_{x'}P(x')\mathrm{e}^{a(x')-\phi_a}\mathrm{e}^{-rd_s(x,x')} \\ &\quad + r(1+\rho')\bigg(\sum_x P(x)\log\bigg[\sum_{x'}P(x')\mathrm{e}^{-\frac{d_s(x,x')}{1+\rho'}}\frac{\mathrm{e}^{a(x')}}{\mathrm{e}^{a(x)}}\bigg] \\ &\qquad + R + 2\delta\bigg) - 2\delta, \qquad (141)\end{aligned}$$

where we have substituted (136).

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 (141) reduces to $R \leq R$, which is trivially satisfied.

VII. Discussion and Conclusion

In this paper, we introduced a sequential random scheme based on randomizing a generalized form of Gilbert-Varshamov codes with a general distance function. This ensemble ensures that the codewords are sufficiently separated in the input space, and simultaneously achieves both the random coding and expurgated exponents. We proved that the RGV exponent is ensemble-tight for any additive decoding metric, and to our knowledge, this is the first such result for any construction achieving the expurgated exponent. In addition, we provided dual-domain expressions, along with a direct derivation that extends beyond the finite-alphabet setting, and we presented choices of the distance function that attain the best possible exponent.

Appendix

A. Proof of Lemma 2

In the following, products of the form $\prod_{i\neq\{k,m\}}$ are a shorthand for $\prod_{i\in\{1,\ldots,m\}\setminus\{k,m\}}$. In addition, for the special case of $m = k+1$, any summations over $\boldsymbol{x}_{k+1}^{m-1}$ are void, and any terms of the form $\Pr(\boldsymbol{x}_{k+1}^{m-1}|\boldsymbol{x}_1^k)$ should be omitted (i.e., replaced by 1)

Since by assumption $k < m$, we have

$$\begin{aligned} &\Pr(\boldsymbol{x}_k,\boldsymbol{x}_m) \\ &= \sum_{\boldsymbol{x}_1^{k-1},\boldsymbol{x}_{k+1}^{m-1}} \Pr(\boldsymbol{x}_1^{k-1})\Pr(\boldsymbol{x}_k|\boldsymbol{x}_1^{k-1})\end{aligned}$$

$$\times \Pr(\boldsymbol{x}_{k+1}^{m-1}|\boldsymbol{x}_1^k)\Pr(\boldsymbol{x}_m|\boldsymbol{x}_1^{m-1}) \tag{142}$$

$$= \sum_{\boldsymbol{x}_1^{k-1},\boldsymbol{x}_{k+1}^{m-1}} \Pr(\boldsymbol{x}_1^{k-1})\Pr(\boldsymbol{x}_{k+1}^{m-1}|\boldsymbol{x}_1^k) \times \frac{\prod_{i=1}^{k-1}\mathbb{1}\{d(\boldsymbol{x}_k,\boldsymbol{x}_i)>\Delta\}}{|\mathcal{T}(P_n,\boldsymbol{x}_1^{k-1})|}\frac{\prod_{i=1}^{m-1}\mathbb{1}\{d(\boldsymbol{x}_m,\boldsymbol{x}_i)>\Delta\}}{|\mathcal{T}(P_n,\boldsymbol{x}_1^{m-1})|} \tag{143}$$

$$\geq \frac{\mathbb{1}\{d(\boldsymbol{x}_k,\boldsymbol{x}_m)>\Delta\}}{|\mathcal{T}(P_n)|^2}\sum_{\boldsymbol{x}_1^{k-1},\boldsymbol{x}_{k+1}^{m-1}}\Pr(\boldsymbol{x}_1^{k-1})\Pr(\boldsymbol{x}_{k+1}^{m-1}|\boldsymbol{x}_1^k) \times \prod_{i=1}^{k-1}\mathbb{1}\{d(\boldsymbol{x}_k,\boldsymbol{x}_i)>\Delta\}\prod_{i\notin\{k,m\}}\mathbb{1}\{d(\boldsymbol{x}_m,\boldsymbol{x}_i)>\Delta\} \tag{144}$$

$$= \frac{\mathbb{1}\{d(\boldsymbol{x}_k,\boldsymbol{x}_m)>\Delta\}}{|\mathcal{T}(P_n)|^2}\sum_{\boldsymbol{x}_1^{k-1},\boldsymbol{x}_{k+1}^{m-1}}\prod_{i\notin\{k,m\}}\Pr(\boldsymbol{x}_i|\boldsymbol{x}_1^{i-1}) \times \mathbb{1}\{d(\boldsymbol{x}_k,\boldsymbol{x}_i)>\Delta\}\mathbb{1}\{d(\boldsymbol{x}_m,\boldsymbol{x}_i)>\Delta\} \tag{145}$$

where (143) follows by noting that the two fractions appearing are precisely $\Pr(\boldsymbol{x}_k|\boldsymbol{x}_1^{k-1})$ and $\Pr(\boldsymbol{x}_m|\boldsymbol{x}_1^{m-1})$, (144) follows from Lemma 1, and (145) writes $\Pr(\boldsymbol{x}_1^{k-1})\Pr(\boldsymbol{x}_{k+1}^{m-1}|\boldsymbol{x}_1^k)$ recursively, as well as extending $\prod_{i=1}^{k-1}\mathbb{1}\{d(\boldsymbol{x}_k,\boldsymbol{x}_i)>\Delta\}$ to $\prod_{i\notin\{k,m\}}\mathbb{1}\{d(\boldsymbol{x}_k,\boldsymbol{x}_i)>\Delta\}$ since the term $\Pr(\boldsymbol{x}_1^{k-1})\Pr(\boldsymbol{x}_{k+1}^{m-1}|\boldsymbol{x}_1^k)$ is zero whenever $d(\boldsymbol{x}_k,\boldsymbol{x}_i)\leq\Delta$ for some $k<i<m$.

We now apply a recursive procedure to the summation in (145). Letting $\psi_i(\boldsymbol{x}_i,\boldsymbol{x}_1^{i-1},\boldsymbol{x}_k,\boldsymbol{x}_m)$ denote the argument to the product therein, we have

$$\sum_{\boldsymbol{x}_1^{k-1},\boldsymbol{x}_{k+1}^{m-1}}\prod_{i\notin\{k,m\}}\psi_i(\boldsymbol{x}_i,\boldsymbol{x}_1^{i-1},\boldsymbol{x}_k,\boldsymbol{x}_m) = \Bigg(\sum_{\boldsymbol{x}_1^{k-1},\boldsymbol{x}_{k+1}^{m-2}}\prod_{i\notin\{k,m,m-1\}}\psi_i(\boldsymbol{x}_i,\boldsymbol{x}_1^{i-1},\boldsymbol{x}_k,\boldsymbol{x}_m)\Bigg) \times \sum_{\boldsymbol{x}_{m-1}}\psi_{m-1}(\boldsymbol{x}_{m-1},\boldsymbol{x}_1^{m-2},\boldsymbol{x}_k,\boldsymbol{x}_m). \tag{146}$$

The summation over $\boldsymbol{x}_{m-1}$ can be expanded as follows:

$$\sum_{\boldsymbol{x}_{m-1}}\psi_{m-1}(\boldsymbol{x}_{m-1},\boldsymbol{x}_1^{m-2},\boldsymbol{x}_k,\boldsymbol{x}_m)$$

$$= \sum_{\boldsymbol{x}_{m-1}}\frac{\mathbb{1}\{\boldsymbol{x}_{m-1}\in\mathcal{T}(P_n,\boldsymbol{x}_1^{m-2})\}}{|\mathcal{T}(P_n,\boldsymbol{x}_1^{m-2})|}\mathbb{1}\{d(\boldsymbol{x}_k,\boldsymbol{x}_{m-1})>\Delta\} \times \mathbb{1}\{d(\boldsymbol{x}_m,\boldsymbol{x}_{m-1})>\Delta\} \tag{147}$$

$$= \frac{|\mathcal{T}(P_n,\boldsymbol{x}_1^{m-2},\boldsymbol{x}_k,\boldsymbol{x}_m)|}{|\mathcal{T}(P_n,\boldsymbol{x}_1^{m-2})|} \tag{148}$$

$$\geq \frac{|\mathcal{T}(P_n,\boldsymbol{x}_1^{m-2})|-2\,\mathrm{vol}_{\boldsymbol{x}}(\Delta)}{|\mathcal{T}(P_n,\boldsymbol{x}_1^{m-2})|} \tag{149}$$

$$= 1-\frac{2\,\mathrm{vol}_{\boldsymbol{x}}(\Delta)}{|\mathcal{T}(P_n,\boldsymbol{x}_1^{k-2})|} \tag{150}$$

$$\geq 1-\frac{2e^{-n(R_n+\delta)}}{1-e^{-n\delta}} \tag{151}$$

$$= 1-2\delta_n e^{-nR_n} \tag{152}$$

where (148) follows since the three indicator functions are simultaneously equal to one if and only if $\boldsymbol{x}_{m-1}\in\mathcal{T}(P_n,\boldsymbol{x}_1^{m-2},\boldsymbol{x}_k,\boldsymbol{x}_m)$, (149) follows since the only sequences that can be in $\mathcal{T}(P_n,\boldsymbol{x}_1^{m-2})$ but not $\mathcal{T}(P_n,\boldsymbol{x}_1^{m-2},\boldsymbol{x}_k,\boldsymbol{x}_m)$ are those in the d-balls centered as $\boldsymbol{x}_k$ and $\boldsymbol{x}_m$ (recall also that $\mathrm{vol}_{\boldsymbol{x}}$ does not depend on $\boldsymbol{x}$), and (151) follows from the volume upper bound and the set cardinality lower bound Lemma 1, and (152) applies the definition of δ_n in (20).

Applying the above procedure recursively to the indices $m-2$, $m-3$, and so on in (146) (skipping index k), and substituting into (145), we obtain

$$\Pr(\boldsymbol{x}_k,\boldsymbol{x}_m) \geq \frac{\mathbb{1}\{d(\boldsymbol{x}_k,\boldsymbol{x}_m)>\Delta\}}{|\mathcal{T}(P_n)|^2}\left(1-\frac{2\delta_n}{\mathrm{e}^{nR_n}}\right)^{\mathrm{e}^{nR_n}} \tag{153}$$

$$\geq \frac{\mathbb{1}\{d(\boldsymbol{x}_k,\boldsymbol{x}_m)>\Delta\}}{|\mathcal{T}(P_n)|^2}(1-4\delta_n^2)\mathrm{e}^{-2\delta_n} \tag{154}$$

where (153) also applies $m-2\leq e^{nR_n}$ in the exponent, and (154) follows from the standard inequality $\left(1-\frac{\alpha}{N}\right)^N\geq e^{-\alpha}\left(1-\frac{\alpha^2}{N}\right)$. This establishes the desired lower bound.

The upper bound in (21) simply follows by applying Lemma 1 to (143), and upper bounding the indicator functions by one.

B. Proof of Lemma 3

Recall the abbreviation in (62) (which we use with k in place of m). Recalling the assumption $i<j<k$, we have

$$\Pr(\boldsymbol{x}_i,\boldsymbol{x}_j,\boldsymbol{x}_k)$$

$$= \sum_{\boldsymbol{x}_1^{i-1},\boldsymbol{x}_{i+1}^{j-1},\boldsymbol{x}_{j+1}^{k-1}}\Pr(\boldsymbol{x}_1^{i-1})\Pr(\boldsymbol{x}_i|\boldsymbol{x}_1^{i-1})\Pr(\boldsymbol{x}_{i+1}^{j-1}|\boldsymbol{x}_1^i) \times \Pr(\boldsymbol{x}_j|\boldsymbol{x}_1^{j-1})\Pr(\boldsymbol{x}_{j+1}^{k-1}|\boldsymbol{x}_1^j)\Pr(\boldsymbol{x}_k|\boldsymbol{x}_1^{k-1}) \tag{155}$$

$$= \sum_{\boldsymbol{x}_1^{i-1},\boldsymbol{x}_{i+1}^{j-1},\boldsymbol{x}_{j+1}^{k-1}}\Pr(\boldsymbol{x}_1^{i-1})\Pr(\boldsymbol{x}_{i+1}^{j-1}|\boldsymbol{x}_1^i)\Pr(\boldsymbol{x}_{j+1}^{k-1}|\boldsymbol{x}_1^j) \times \frac{\prod_{r=1}^{i-1}\mathbb{1}\{d(\boldsymbol{x}_i,\boldsymbol{x}_r)>\Delta\}}{|\mathcal{T}(P_n,\boldsymbol{x}_1^{i-1})|}\frac{\prod_{s=1}^{j-1}\mathbb{1}\{d(\boldsymbol{x}_j,\boldsymbol{x}_s)>\Delta\}}{|\mathcal{T}(P_n,\boldsymbol{x}_1^{j-1})|} \times \frac{\prod_{t=1}^{m-1}\mathbb{1}\{d(\boldsymbol{x}_m,\boldsymbol{x}_t)>\Delta\}}{|\mathcal{T}(P_n,\boldsymbol{x}_1^{k-1})|} \tag{156}$$

$$\leq \frac{\mathcal{I}_{d,\Delta}(\boldsymbol{x}_i,\boldsymbol{x}_j,\boldsymbol{x}_k)}{(1-e^{-n\delta})^3|\mathcal{T}(P_n)|^3}\sum_{\boldsymbol{x}_1^{i-1},\boldsymbol{x}_{i+1}^{j-1},\boldsymbol{x}_{j+1}^{k-1}}\Pr(\boldsymbol{x}_1^{i-1}) \times \Pr(\boldsymbol{x}_{i+1}^{j-1}|\boldsymbol{x}_1^i)\Pr(\boldsymbol{x}_{j+1}^{k-1}|\boldsymbol{x}_1^j) \tag{157}$$

$$= \frac{\mathcal{I}_{d,\Delta}(\boldsymbol{x}_i,\boldsymbol{x}_j,\boldsymbol{x}_k)}{(1-e^{-n\delta})^3|\mathcal{T}(P_n)|^3}\sum_{\boldsymbol{x}_1^{i-1}}\Pr(\boldsymbol{x}_1^{i-1}) \times \sum_{\boldsymbol{x}_{i+1}^{j-1}}\Pr(\boldsymbol{x}_{i+1}^{j-1}|\boldsymbol{x}_1^i)\sum_{\boldsymbol{x}_{j+1}^{k-1}}\Pr(\boldsymbol{x}_{j+1}^{k-1}|\boldsymbol{x}_1^j) \tag{158}$$

$$= \frac{\mathcal{I}_{d,\Delta}(\boldsymbol{x}_i,\boldsymbol{x}_j,\boldsymbol{x}_k)}{(1-e^{-n\delta})^3|\mathcal{T}(P_n)|^3}, \tag{159}$$

where (155) substitutes the conditional codeword distributions given all previous codewords, and (157) uses Lemma 1.

C. Proof of Lemma 4

Let π be a permutation of the indices $[1,\dots,n]$, and let $\pi(\boldsymbol{x})$ be the outcome of applying the permutation π to the sequence $\boldsymbol{x}$. By the definition of the generalized RGV construction (in

particular, the fact that the codewords are drawn uniformly and d is type-dependent), we have

$$\begin{aligned}&\Pr\left(\boldsymbol{X}_1=\boldsymbol{x}_1,\boldsymbol{X}_2=\boldsymbol{x}_2,\ldots,\boldsymbol{X}_m=\boldsymbol{x}_m\right)\\&=\Pr\left(\boldsymbol{X}_1=\pi(\boldsymbol{x}_1),\boldsymbol{X}_2=\pi(\boldsymbol{x}_2),\ldots,\boldsymbol{X}_m=\pi(\boldsymbol{x}_m)\right).\end{aligned}\tag{160}$$

We now consider summing both sides over all sequences $(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{m-1})$ that are admissible in the sense of meeting the requirement $d(\boldsymbol{x}_i,\boldsymbol{x}_j)>\Delta$ for all $i,j\in\{1,...,m\}$. Clearly such a summation yields $\Pr\left(\boldsymbol{X}_m=\boldsymbol{x}_m\right)$ on the left-hand side. Moreover, for each such $(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{m-1})$, the type-dependent nature of d implies that $(\pi(\boldsymbol{x}_1),\ldots,\pi(\boldsymbol{x}_{m-1}))$ and $(\pi^{-1}(\boldsymbol{x}_1),\ldots,\pi^{-1}(\boldsymbol{x}_{m-1}))$ are also admissible. As a result, we are also summing the right-hand side over all admissible sequences, yielding

$$\Pr\left(\boldsymbol{X}_m=\boldsymbol{x}_m\right)=\Pr\left(\boldsymbol{X}_m=\pi(\boldsymbol{x}_m)\right),\tag{161}$$

which implies that $\boldsymbol{X}_m$ is distributed uniformly over $\mathcal{T}(P_n)$.

D. Proof of Lemma 5

The RGV exponent, defined in (25), is a minimization over joint distributions $V_{X\widetilde{X}Y}$ within the constraint set $\mathcal{T}_{d,q,P}(\Delta)$ given in (26).

Let $V^*_{X\widetilde{X}Y}$ denote the minimizer subject to $\mathcal{T}_{d,q,P}(\Delta)$, and let $V^*_{X\widetilde{X}Y,n}$ denote the minimizer subject to $\mathcal{T}_{d,q,P_n}(\Delta)$. Since the space of probability distributions is compact, any infinite subsequence of $V^*_{X\widetilde{X}Y,n}$ must have a further subsequence converging to some $V^*_{X\widetilde{X}Y,\infty}$. Moreover, since d and q are continuous and $V^*_{X\widetilde{X}Y,n}\in\mathcal{T}_{d,q,P_n}(\Delta)$ with $P_n\to P$, we must have $V^*_{X\widetilde{X}Y,\infty}\in\mathcal{T}_{d,q,P}(\Delta)$, from which (57) follows.

E. Proof of Lemma 6

In this appendix, we make use of the following notation, also used in Section VI-B:

$$\Gamma(V_{X\widetilde{X}Y})\triangleq D(V_{Y|X}\|W|V_X)+\left|I(\widetilde{X};Y,X)-R\right|_+.\tag{162}$$

We observe that the exponent on the right-hand side of (79) can be rewritten as

$$\min_{\substack{V_{X\widetilde{X}}\in\mathcal{P}_n(\mathcal{X}^2):\\ V_X=V_{\widetilde{X}}=P_n,\\ d(V_{X\widetilde{X}})\ge\Delta}}\ \min_{\substack{V_{Y|X\widetilde{X}}\in\mathcal{P}_n(\mathcal{Y}|V_{X\widetilde{X}}):\\ q(P_n\times V_{Y|\widetilde{X}})-q(P_n\times V_{Y|X})\ge 0}}\Gamma(V_{X\widetilde{X}}\times V_{Y|X\widetilde{X}}),\tag{163}$$

where the notation $V_{Y|X\widetilde{X}}\in\mathcal{P}_n(\mathcal{Y}|V_{X\widetilde{X}})$ means that $V_{X\widetilde{X}}\times V_{Y|X\widetilde{X}}$ is a joint empirical distribution for sequences of length n. Throughout the appendix, we will make use of the fact the minimizers must be such that

$$W(y|x)=0\implies V_{Y|X}(y|x)=0,\tag{164}$$

since otherwise the KL divergence in (162) would be infinite. Observe that within the space of joint distributions satisfying (164), the function $\Gamma(\cdot)$ is continuous.

We first show that the inner minimization can be approximated by a minimization over $V_{Y|X\widetilde{X}}\in\mathcal{P}(\mathcal{Y}|\mathcal{X}^2)$, and then we show that the outer minimization can be approximated by a minimization over $V_{X\widetilde{X}}\in\mathcal{P}(\mathcal{X}^2)$.

Inner minimization. Define $\Psi(V_{X\widetilde{X}}\times V_{Y|X\widetilde{X}})=q(P_n\times V_{Y|\widetilde{X}})-q(P_n\times V_{Y|X})$, so that the constraint in (163) is given by $\Psi(V_{X\widetilde{X}}\times V_{Y|X\widetilde{X}})\ge 0$. For any $V_{X\widetilde{X}}\in\mathcal{P}_n(\mathcal{X}^2)$, we need to show that the inner minimization in (163) can be expanded from $\mathcal{P}_n(\mathcal{Y}|V_{X\widetilde{X}})$ to $\mathcal{P}(\mathcal{Y}|\mathcal{X}^2)$. Specifically, we wish to show that for any $\epsilon>0$, it holds for sufficiently large n that

$$\begin{aligned}&\min_{V_{Y|X\widetilde{X}}\in\mathcal{P}_n(\mathcal{Y}|V_{X\widetilde{X}}):\,\Psi(V_{X\widetilde{X}}\times V_{Y|X\widetilde{X}})\ge 0}\Gamma(V_{X\widetilde{X}}\times V_{Y|X\widetilde{X}})\\&\le\min_{V_{Y|X\widetilde{X}}\in\mathcal{P}(\mathcal{Y}|\mathcal{X}^2):\,\Psi(V_{X\widetilde{X}}\times V_{Y|X\widetilde{X}})\ge 0}\Gamma(V_{X\widetilde{X}}\times V_{Y|X\widetilde{X}})+\epsilon.\end{aligned}\tag{165}$$

Since we are considering additive decoding metrics, i.e., $q(P_{XY})=\mathbb{E}_P[q(X,Y)]$, we have

$$\begin{aligned}&\Psi(V_{X\widetilde{X}}\times V_{Y|X\widetilde{X}})\\&=\sum_{x,\overline{x},y}V_{X\widetilde{X}}(x,\overline{x})V_{Y|X\widetilde{X}}(y|x,\overline{x})\cdot[q(\overline{x},y)-q(x,y)].\end{aligned}\tag{166}$$

To prove (165), fix any $\widetilde{V}_{Y|X\widetilde{X}}\in\mathcal{P}(\mathcal{Y}|\mathcal{X}^2)$ with $\Psi(V_{X\widetilde{X}}\times\widetilde{V}_{Y|X\widetilde{X}})\ge 0$, and let $V^{(n)}_{Y|X\widetilde{X}}$ be the quantized version of $\widetilde{V}_{Y|X\widetilde{X}}$ that rounds up for the highest values of $q(\overline{x},y)-q(x,y)$, and rounds down for the smallest values:

$$\begin{aligned}&V^{(n)}_{Y|X\widetilde{X}}(y|x,\overline{x})\\&=\begin{cases}\frac{1}{nV_{X\widetilde{X}}(x,\overline{x})}\lceil n\cdot V_{X\widetilde{X}}(x,\overline{x})\cdot\widetilde{V}_{Y|X\widetilde{X}}(y|x,\overline{x})\rceil\\\qquad\text{if } q(\overline{x},y)-q(x,y)>c_{x\overline{x}}\\\frac{1}{nV_{X\widetilde{X}}(x,\overline{x})}\lfloor n\cdot V_{X\widetilde{X}}(x,\overline{x})\cdot\widetilde{V}_{Y|X\widetilde{X}}(y|x,\overline{x})\rfloor\\\qquad\text{if } q(\overline{x},y)-q(x,y)<c_{x\overline{x}},\end{cases}\end{aligned}\tag{167}$$

where for each $(x,\overline{x})$, we choose $c_{x\overline{x}}$ (as well as rounding the entries with $q(\overline{x},y)-q(x,y)=c_{x\overline{x}}$ up or down as needed) in such a way that the entries of $V^{(n)}_{Y|X\widetilde{X}}(y|x,\overline{x})$ sum to one.

By this construction and the fact that $\Psi(V_{X\widetilde{X}}\times V_{Y|X\widetilde{X}})$ is a positive linear combination of the values $q(\overline{x},y)-q(x,y)$ (*cf.*, (166)), we have

$$\Psi(V_{X\widetilde{X}}\times V^{(n)}_{Y|X\widetilde{X}})\ge\Psi(V_{X\widetilde{X}}\times\widetilde{V}_{Y|X\widetilde{X}})\tag{168}$$

and

$$\begin{aligned}&\sum_{x,\widetilde{x},y}\left|V_{X\widetilde{X}}(x,\widetilde{x})V^{(n)}_{Y|X\widetilde{X}}(y|x,\widetilde{x})-V_{X\widetilde{X}}(x,\widetilde{x})\widetilde{V}_{Y|X\widetilde{X}}(y|x,\widetilde{x})\right|\\&\qquad\le\frac{|\mathcal{X}|^2|\mathcal{Y}|}{n}.\end{aligned}\tag{169}$$

In particular, (168) immediately implies that the required constraint $\Psi(V_{X\widetilde{X}}\times V^{(n)}_{Y|X\widetilde{X}})\ge 0$ is satisfied. Moreover, (169) implies that $V_{X\widetilde{X}}\times V^{(n)}_{Y|X\widetilde{X}}$ is $O\big(\frac{1}{n}\big)$-close to $V_{X\widetilde{X}}\times\widetilde{V}_{Y|X\widetilde{X}}$ (in the ℓ_1 sense), and hence $\Gamma(V_{X\widetilde{X}}\times\widetilde{V}_{Y|X\widetilde{X}})-\Gamma(V_{X\widetilde{X}}\times V^{(n)}_{Y|X\widetilde{X}})\to 0$ by the continuity of $\Gamma(\cdot)$. This proves the part of the approximation of the inner minimization, i.e., (165).

Outer minimization. Having proved (165), the double minimization (163) is upper bounded by the following double minimization:

$$\min_{\substack{V_{X\widetilde{X}}\in\mathcal{P}_n(\mathcal{X}^2):\\ V_X=V_{\widetilde{X}}=P_n,\\ d(V_{X\widetilde{X}})\geq\Delta}} \ \min_{\substack{V_{Y|X\widetilde{X}}\in\mathcal{P}(\mathcal{Y}|\mathcal{X}^2):\\ q(P_n\times V_{Y|\widetilde{X}})-q(P_n\times V_{Y|X})\geq 0}} \Gamma(V_{X\widetilde{X}}\times V_{Y|X\widetilde{X}}). \tag{170}$$

Consider the expression in (170) with $\mathcal{P}_n(\mathcal{X}^2)$ replaced by $\mathcal{P}(\mathcal{X}^2)$ and P_n replaced by P:

$$\min_{\substack{V_{X\widetilde{X}}\in\mathcal{P}(\mathcal{X}^2):\\ V_X=V_{\widetilde{X}}=P,\\ d(V_{X\widetilde{X}})\geq\Delta}} \ \min_{\substack{V_{Y|X\widetilde{X}}\in\mathcal{P}(\mathcal{Y}|\mathcal{X}^2):\\ q(P\times V_{Y|\widetilde{X}})-q(P\times V_{Y|X})\geq 0}} \Gamma(V_{X\widetilde{X}}\times V_{Y|X\widetilde{X}}). \tag{171}$$

Given the minimizer $V^*_{X\widetilde{X}}\in\mathcal{P}(\mathcal{X}^2)$ with $V^*_X=V^*_{\widetilde{X}}=P$, let $V^*_{X\widetilde{X},n}$ be the closest joint type (e.g., in the ℓ_∞ sense) that satisfies $V^*_X=V^*_{\widetilde{X}}=P_n$. It follows that $V^*_{X\widetilde{X},n}(x,\overline{x})-V^*_{X\widetilde{X}}(x,\overline{x})\to 0$.

Let $V^*_{Y|X\widetilde{X}}$ denote the minimizer in (171), and define

$$V^{\max}_{Y|X\widetilde{X},n}=\operatorname*{arg\,max}_{V_{Y|X\widetilde{X}}\in\mathcal{P}_n(\mathcal{Y}|\mathcal{X}^2)}\Psi(V_{X\widetilde{X},n}\times V_{Y|X\widetilde{X}}). \tag{172}$$

We claim that there exists a vanishing sequence ϵ_n such that

$$(1-\epsilon_n)\Psi(V^*_{X\widetilde{X},n}\times V^*_{Y|X\widetilde{X}})+\epsilon_n\Psi(V_{X\widetilde{X},n}\times V^{\max}_{Y|X\widetilde{X},n})\geq 0. \tag{173}$$

To see this, note that since $\Psi(V^*_{X\widetilde{X}}\times V^*_{Y|X\widetilde{X}})\geq 0$ by definition, we only need the second term in (173) to be large enough to overcome the rounding from $V^*_{X\widetilde{X}}$ to $V^*_{X\widetilde{X},n}$. If $\Psi(V_{X\widetilde{X},n}\times V^{\max}_{Y|X\widetilde{X},n})>0$, then this is possible by letting ϵ_n vanish sufficiently slowly. On the other hand, $\Psi(V_{X\widetilde{X},n}\times V^{\max}_{Y|X\widetilde{X},n})<0$ is impossible, since one could swap the roles of X and $\widetilde{X}$ in (172) to produce a positive quantity. The only remaining case is that $\Psi(V_{X\widetilde{X},n}\times V_{Y|X\widetilde{X}})=0$ for all $V_{Y|X\widetilde{X}}$, in which case (173) is trivial.

Using (173) and the continuity of Γ (subject to (164), which we have established to always hold), we deduce the following for any $\epsilon>0$ and sufficiently large n:

$$\min_{\substack{V_{X\widetilde{X}}\in\mathcal{P}(\mathcal{X}^2):\\ V_X=V_{\widetilde{X}}=P,\\ d(V_{X\widetilde{X}})\geq\Delta}} \ \min_{\substack{V_{Y|X\widetilde{X}}\in\mathcal{P}(\mathcal{Y}|\mathcal{X}^2):\\ q(P\times V_{Y|\widetilde{X}})-q(P\times V_{Y|X})\geq 0}} \Gamma(V_{X\widetilde{X}}\times V_{Y|X\widetilde{X}}) \tag{174}$$

$$=\Gamma(V^*_{X\widetilde{X}}\times V^*_{Y|X\widetilde{X}}) \tag{175}$$

$$\geq\Gamma(V^*_{X\widetilde{X},n}\times V^*_{Y|X\widetilde{X}})-\epsilon \tag{176}$$

$$\geq\Gamma\left(V^*_{X\widetilde{X},n}\times\left[(1-\epsilon_n)V^*_{Y|X\widetilde{X}}+\epsilon_n V^{\max}_{Y|X\widetilde{X},n}\right]\right)-2\epsilon \tag{177}$$

$$\geq\min_{\substack{V_{Y|X\widetilde{X}}\in\mathcal{P}_n(\mathcal{Y}|\mathcal{X}^2):\\ q(P_n\times V_{Y|\widetilde{X}})-q(P_n\times V_{Y|X})\geq 0}}\Gamma\left(V^*_{X\widetilde{X},n}\times V_{Y|X\widetilde{X}}\right)-2\epsilon \tag{178}$$

$$\geq\min_{\substack{V_{X\widetilde{X}}\in\mathcal{P}(\mathcal{X}^2):\\ V_X=V_{\widetilde{X}}=P_n,\\ d(V_{X\widetilde{X}})\geq\Delta-\epsilon}} \ \min_{\substack{V_{Y|X\widetilde{X}}\in\mathcal{P}_n(\mathcal{Y}|\mathcal{X}^2):\\ q(P_n\times V_{Y|\widetilde{X}})-q(P_n\times V_{Y|X})\geq 0}}\Gamma\left(V_{X\widetilde{X}}\times V_{Y|X\widetilde{X}}\right)-2\epsilon \tag{179}$$

where both (176) and (177) follow from the continuity of $\Gamma(\cdot)$, (178) follows since $(1-\epsilon_n)V^*_{Y|X\widetilde{X}}+\epsilon_n V^{\max}_{Y|X\widetilde{X},n}$ belongs to the constraint set in the minimization due to (173), and (179) follows since $d(V^*_{X\widetilde{X}})\geq\Delta\implies d(V^*_{X\widetilde{X},n})\geq\Delta-\epsilon$ by the continuity of d.

Since Δ is arbitrary in the preceding steps, we may replace Δ by $\Delta+\epsilon$ in both (174) and (179). Upon doing so, we obtain the RGV exponent with input distribution P and parameter $\Delta+\epsilon$ on the left-hand side, while recovering the expression (170) from the first step above on the right-hand side. This completes the proof of Lemma 6.

F. Primal-dual Equivalence

The primal-dual equivalence stated in Theorem 2 follows in a near-identical manner to the mismatched random coding exponent [4] (and to a lesser extent, the mismatched expurgated exponent [4]), so we omit most of the details. We first consider the exponent (82), and then the rate condition (83).

Exponent expression. The proof of equivalence for the exponent consists of three steps, interleaved with applications of the minimax theorem to swap the order of the primal and dual optimization variables:

1) Let P_{XY} be fixed, and consider the optimization problem

$$\min_{\substack{V_{X\widetilde{X}Y}:V_{XY}=P_{XY},P_{\widetilde{X}}=P,\\ q(V_{\widetilde{X}Y})\geq q(P_{XY}),d(P_{X\widetilde{X}})\geq\Delta}} D\big(V_{X\widetilde{X}Y}\|P\times P_{XY}\big), \tag{180}$$

where $P\times P_{XY}$ denotes the joint distribution $P(x)P(\widetilde{x})P_{Y|X}(y|x)$. This minimization arises from fixing the (X,Y) marginals in (25) and noting that all terms other than the mutual information $I(\widetilde{X};X,Y)$ are constant. The mutual information is equivalent to the objective function in (180), due to the equality constraints.

Applying Lagrange duality in the same way as the random coding setting [10] (see also [26, Appendix E]), we find that (180) is equivalent to[3]

$$\sup_{s\geq 0,r\geq 0,a(\cdot)}-\sum_{x,y}P_{XY}(x,y)\times\log\frac{\sum_{x'}Q(x')e^{sq(x',y)}e^{a(x')}e^{r(d(x,x')-\Delta)}}{e^{sq(x,y)}e^{a(x)}}, \tag{181}$$

where s, r, and $a(\cdot)$ are Lagrange multipliers corresponding to the metric constraint, distance constraint, and $\widetilde{X}$-marginal constraint.

2) Let $g_{s,r,a}(x,y)=-\log\frac{\sum_{x'}Q(x')e^{sq(x',y)}e^{a(x')}e^{r(d(x,x')-\Delta)}}{e^{sq(x,y)}e^{a(x)}}$ be the function being averaged in (181). Based on the definition in (25), the previous step, and the minimax theorem, the RGV exponent is given by

$$\sup_{s\geq 0,r\geq 0,a(\cdot)}\min_{V_{XY}\,:\,P_X=P}D(V_{XY}\|P\times W)+\big|\mathbb{E}_V[g_{s,r,a}(X,Y)]-R\big|_+. \tag{182}$$

[3]We have $e^{q(x,y)}$ in place of $q(x,y)$ in [10] because we are considering additive (rather than multiplicative) decoding rules.

By applying $[z]_+ = \max_{\rho\in[0,1]} \rho z$ along with the minimax theorem, we find that this is equivalent to

$$\sup_{\rho\in[0,1],s\geq 0,r\geq 0,a(\cdot)} \min_{V_{XY}\,:\,P_X=P} D(V_{XY}\|P\times W) + \rho\big(\mathbb{E}_V[g_{s,r,a}(X,Y)] - R\big). \tag{183}$$

3) A minimization problem of the form (183) was already considered in [10] (with a different choice of $g_{s,r,a}$), and it was shown that the minimization is equivalent to the expression

$$-\sum_x Q(x)\log\sum_y W(y|x)e^{\rho g_{s,r,a}(x,y)}. \tag{184}$$

Substituting the definition of $g_{s,r,a}$ completes the proof.

Rate condition expression. The primal-dual equivalence for the rate condition can be proved using similar steps to those above; here we briefly discuss another way that it can be understood.

The primal expression (27) is of the same form as the so-called *LM rate* for mismatched decoding [3], [21], [22], with $\widetilde{X}$ playing the role of Y, and d playing the role of the decoding metric. Accordingly, the primal-dual equivalence is essentially a special case of that of the LM rate, which is well-established in the mismatched decoding literature [8], [26], [27].

G. Lower Bound for Marginal Distribution in Cost-Constrained Coding

Here we show that in the cost-constrained coding setting of Section F, each $\Pr(\boldsymbol{x}_m)$ is lower bounded by $P_{\boldsymbol{X}}(\boldsymbol{x}_m)$ times a constant tending to one. Recall that the codeword distribution is of the form (88) with $1-e^{-n\delta} \leq \mu_m(\boldsymbol{x}_1^{m-1}) \leq 1$ (see the properties following (90)). We have

$$\begin{aligned}
&\Pr(\boldsymbol{x}_m)\\
&= \sum_{\boldsymbol{x}_1^{m-1}} \Pr(\boldsymbol{x}_1^{m-1})\Pr(\boldsymbol{x}_m|\boldsymbol{x}_1^{m-1}) && (185)\\
&= \sum_{\boldsymbol{x}_1^{m-1}} \Pr(\boldsymbol{x}_1^{m-1})\frac{P_{\boldsymbol{X}}(\boldsymbol{x}_m)}{\mu_m(\boldsymbol{x}_1^{m-1})}\mathbb{1}\{d(\boldsymbol{x}_i,\boldsymbol{x}_m)>\Delta,\,\forall i<m\} && (186)\\
&\geq P_{\boldsymbol{X}}(\boldsymbol{x}_m)\sum_{\boldsymbol{x}_1^{m-1}} \Pr(\boldsymbol{x}_1^{m-1})\mathbb{1}\{d(\boldsymbol{x}_i,\boldsymbol{x}_m)>\Delta,\,\forall i<m\} && (187)\\
&\geq P_{\boldsymbol{X}}(\boldsymbol{x}_m)\sum_{\boldsymbol{x}_1^{m-1}}\prod_{i=1}^{m-1}\big(\Pr(\boldsymbol{x}_i|\boldsymbol{x}_1^{i-1})\mathbb{1}\{d(\boldsymbol{x}_i,\boldsymbol{x}_m)>\Delta\}\big), && (188)
\end{aligned}$$

where (186) substitutes (88), (187) uses the fact that $\mu_m(\boldsymbol{x}_1^{m-1}) \leq 1$, and (188) is an expansion of $\Pr(\boldsymbol{x}_1^{m-1})$.

We now unravel the product one term at a time. We start by writing

$$\begin{aligned}
&\sum_{\boldsymbol{x}_1^{m-1}}\prod_{i=1}^{m-1}\big(\Pr(\boldsymbol{x}_i|\boldsymbol{x}_1^{i-1})\mathbb{1}\{d(\boldsymbol{x}_i,\boldsymbol{x}_m)>\Delta\}\big)\\
&= \sum_{\boldsymbol{x}_1^{m-2}}\prod_{i=1}^{m-2}\big(\Pr(\boldsymbol{x}_i|\boldsymbol{x}_1^{i-1})\mathbb{1}\{d(\boldsymbol{x}_i,\boldsymbol{x}_m)>\Delta\}\big)\\
&\quad\times\sum_{\boldsymbol{x}_{m-1}}\Pr(\boldsymbol{x}_{m-1}|\boldsymbol{x}_1^{m-2})\mathbb{1}\{d(\boldsymbol{x}_{m-1},\boldsymbol{x}_m)>\Delta\}. \qquad (189)
\end{aligned}$$

Henceforth, let $\mathcal{D}(\cdot)$ denote the set of possible codewords that are at a distance exceeding Δ from all codewords listed in the brackets. Substituting the conditional codeword distribution for codeword $m-1$ gives

$$\begin{aligned}
&\sum_{\boldsymbol{x}_{m-1}}\Pr(\boldsymbol{x}_{m-1}|\boldsymbol{x}_1^{m-2})\mathbb{1}\{d(\boldsymbol{x}_{m-1},\boldsymbol{x}_m)>\Delta\}\\
&= \frac{1}{\mu_{m-1}(\boldsymbol{x}_1^{m-2})}\sum_{\boldsymbol{x}_{m-1}} P_{\boldsymbol{X}}(\boldsymbol{x}_{m-1})\mathbb{1}\{\boldsymbol{x}_{m-1}\in\mathcal{D}(\boldsymbol{x}_1^{m-2},\boldsymbol{x}_m)\} && (190)\\
&= \frac{\Pr(\boldsymbol{X}'\in\mathcal{D}(\boldsymbol{x}_1^{m-2},\boldsymbol{x}_m))}{\Pr(\boldsymbol{X}'\in\mathcal{D}(\boldsymbol{x}_1^{m-2}))}, && (191)
\end{aligned}$$

where $\boldsymbol{X}' \sim P_{\boldsymbol{X}}$, and the denominator in (191) follows since $\mu_{m-1}(\boldsymbol{x}_1^{m-2}) = \Pr(\boldsymbol{X}'\in\mathcal{D}(\boldsymbol{x}_1^{m-2}))$ by definition.

Continuing, we write

$$\begin{aligned}
&\frac{\Pr(\boldsymbol{X}'\in\mathcal{D}(\boldsymbol{x}_1^{m-2},\boldsymbol{x}_m))}{\Pr(\boldsymbol{X}'\in\mathcal{D}(\boldsymbol{x}_1^{m-2}))}\\
&\geq \frac{\Pr(\boldsymbol{X}'\in\mathcal{D}(\boldsymbol{x}_1^{m-2})) - \Pr(d(\boldsymbol{X}',\boldsymbol{x}_m)\leq\Delta)}{\Pr(\boldsymbol{X}'\in\mathcal{D}(\boldsymbol{x}_1^{m-2}))} && (192)\\
&= 1 - \frac{\Pr(d(\boldsymbol{X}',\boldsymbol{x}_m)\leq\Delta)}{\Pr(\boldsymbol{X}'\in\mathcal{D}(\boldsymbol{x}_1^{m-2}))} && (193)\\
&\geq 1 - \frac{\Pr(d(\boldsymbol{X}',\boldsymbol{x}_m)\leq\Delta)}{1-e^{-n\delta}} && (194)\\
&\geq 1 - \frac{e^{-n(R_n+\delta)}}{1-e^{-n\delta}} && (195)\\
&= 1 - \delta_n e^{-nR_n}, && (196)
\end{aligned}$$

where (192) uses $\Pr(A\cap B)\geq\Pr(A)-\Pr(B^c)$, (193) applies $\mu_{m-1}(\boldsymbol{x}_1^{m-2})\geq 1-e^{-n\delta}$, (194) makes use of the bounds on $\Pr(d(\boldsymbol{X}',\boldsymbol{x}_m)\leq\Delta)$ and R_n in (90) and (83) respectively, and (196) uses the definition of δ_n in (20).

The recursion in (188) proceeds in the exact same way as the constant-composition case in Appendix A (with a factor of 2 removed), and we get

$$\Pr(\boldsymbol{x}_m) \geq P_{\boldsymbol{X}}(\boldsymbol{x}_m)\cdot(1-\delta_n^2)e^{-\delta_n}. \tag{197}$$

ACKNOWLEDGMENT

We would like to thank the reviewers for their very helpful comments and in particular for the suggestion on how to significantly shorten the proof of the achievability part of Theorem 1.

REFERENCES

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

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

[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, 1981.

[4] 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.

[5] N. Merhav, "List decoding - random coding exponents and expurgated exponents," *IEEE Trans. Inf. Theory*, vol. 60, no. 11, pp. 6749–6759, Nov 2014.
[6] ——, "The generalized stochastic likelihood decoder: Random coding and expurgated bounds," *IEEE Trans. Inf. Theory*, vol. 63, no. 8, pp. 5039–5051, Aug 2017.
[7] A. Somekh-Baruch, "On achievable rates and error exponents for channels with mismatched decoding," *IEEE Trans. Inf. Theory*, vol. 61, no. 2, pp. 727–740, Feb 2015.
[8] 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.
[9] J. Scarlett, A. Martinez, and A. Guillén i Fàbregas, "Ensemble-tight error exponents for mismatched decoders," in *50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2012, pp. 1951–1958.
[10] ——, "Mismatched decoding: Error exponents, second-order rates and saddlepoint approximations," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2647–2666, May 2014.
[11] E. N. Gilbert, "A comparison of signalling alphabets," *Bell Labs Tech. J.*, vol. 31, no. 3, pp. 504–522, 1952.
[12] 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.
[13] V. Siforov, "On noise stability of a system with error-correcting codes," *IRE Trans. Inf. Theory*, vol. 2, no. 4, pp. 109–115, 1956.
[14] V. I. Levenshtein, "A class of systematic codes," *Doklady Akademii Nauk SSSR*, vol. 131, no. 5, pp. 1011–1014, 1960.
[15] R. A. Brualdi and V. S. Pless, "Greedy codes," *J. Combinatorial Theory, Series A*, vol. 64, no. 1, pp. 10–30, 1993.
[16] J. Conway and N. Sloane, "Lexicographic codes: error-correcting codes from game theory," *IEEE Trans. Inf. Theory*, vol. 32, no. 3, pp. 337–348, 1986.
[17] A. Trachtenberg, "Error-correcting codes on graphs: lexicodes, trellises and factor graphs," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2000.
[18] R. Blahut, "Composition bounds for channel block codes," *IEEE Trans. Inf. Theory*, vol. 23, no. 6, pp. 656–674, 1977.
[19] A. Barg and G. D. Forney, "Random codes: minimum distances and error exponents," *IEEE Trans. Inf. Theory*, vol. 48, no. 9, pp. 2568–2573, Sep 2002.
[20] A. Somekh-Baruch and N. Merhav, "Exact random coding exponents for erasure decoding," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 6444 –6454, oct. 2011.
[21] 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.
[22] J. Hui, "Fundamental issues of multiple accessing," *PhD dissertation, MIT*, 1983.
[23] 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.
[24] D. De Caen, "A lower bound on the probability of a union," *Discrete Mathematics*, vol. 169, no. 1-3, pp. 217–220, 1997.
[25] G. Poltyrev, "Random coding bounds for discrete memoryless channels," *Prob. Inf. Transm.*, vol. 18, no. 1, pp. 9–21, 1982.
[26] J. Scarlett, "Reliable communication under mismatched decoding," Ph.D. dissertation, University of Cambridge, 2014, http://itc.upf.edu/biblio/1061.
[27] 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.

Anelia Somekh-Baruch (S'01-M'03) received the B.Sc. degree from Tel-Aviv University, Tel-Aviv, Israel, in 1996 and the M.Sc. and Ph.D. degrees from the Technion?Israel Institute of Technology, Haifa, Israel, in 1999 and 2003, respectively, all in electrical engineering. During 2003?2004, she was with the Technion Electrical Engineering Department. During 2005?2008, she was a Visiting Research Associate at the Electrical Engineering Department, Princeton University, Princeton, NJ. From 2008 to 2009 she was a researcher at the Electrical Engineering Department, Technion, and from 2009 she has been with the Bar-Ilan University School of Engineering, Ramat-Gan, Israel. Her research interests include topics in information theory and communication theory. Dr. Somekh-Baruch received the Tel-Aviv University program for outstanding B.Sc. students scholarship, the Viterbi scholarship, the Rothschild Foundation scholarship for postdoctoral studies, and the Marie Curie Outgoing International Fellowship.

Jonathan Scarlett (S'14 – M'15) received the B.Eng. degree in electrical engineering and the B.Sci. degree in computer science from the University of Melbourne, Australia. From October 2011 to August 2014, he was a Ph.D. student in the Signal Processing and Communications Group at the University of Cambridge, United Kingdom. From September 2014 to September 2017, he was post-doctoral researcher with the Laboratory for Information and Inference Systems at the École Polytechnique Fédérale de Lausanne, Switzerland. Since January 2018, he has been an assistant professor in the Department of Computer Science and Department of Mathematics, National University of Singapore. His research interests are in the areas of information theory, machine learning, signal processing, and high-dimensional statistics. He received the 'EPFL Fellows' postdoctoral fellowship co-funded by Marie Curie, and the NUS Early Career Research Award.

Albert Guillén i Fàbregas (S'01–M'05–SM'09) received the Telecommunication Engineering degree and the Electronics Engineering degree from the Universitat Politècnica de Catalunya, and the Politecnico di Torino, Torino, Italy, respectively, in 1999, and the Ph.D. degree in Communication Systems from Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, in 2004.

Since 2011 he has been an ICREA Research Professor at Universitat Pompeu Fabra. He is also an Adjunct Researcher at the University of Cambridge. He has held appointments at the New Jersey Institute of Technology, Telecom Italia, European Space Agency (ESA), Institut Eurécom, University of South Australia, University of Cambridge, as well as visiting appointments at EPFL, École Nationale des Télécommunications (Paris), Universitat Pompeu Fabra, University of South Australia, Centrum Wiskunde & Informatica and Texas A&M University in Qatar. His research interests are in the areas of information theory, coding theory and communication theory.

Dr. Guillén i Fàbregas is a member of the Young Academy of Europe, received both Starting and Consolidator Grants from the European Research Council, the Young Authors Award of the 2004 European Signal Processing Conference (EUSIPCO), the 2004 Nokia Best Doctoral Thesis Award from the Spanish Institution of Telecommunications Engineers, and a pre-doctoral Research Fellowship of the Spanish Ministry of Education to join ESA. He is an Associate Editor of the IEEE TRANSACTIONS ON INFORMATION THEORY, an Editor of the Foundations and Trends in Communications and Information Theory and was an Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS.