A Derivation of the Cost-Constrained Sphere-Packing Exponent

Gonzalo Vazquez-Vilar[1], Alfonso Martinez[2] and Albert Guillén i Fàbregas[234]
[1]Universidad Carlos III de Madrid, [2]Universitat Pompeu Fabra, [3]ICREA, [4]University of Cambridge
Email: {gvazquez,alfonso.martinez,guillen}@ieee.org

***Abstract*—We derive the channel-coding sphere-packing exponent under a per-codeword cost constraint. The proof is based on hypothesis testing and holds for continuous memoryless channels.**

***Index Terms*—Channel coding, reliability function, sphere-packing exponent, cost constraint, continuous channel.**

I. Introduction

The behavior of the channel-coding error probability may be quantified in terms of error exponents, defined as the rate of the error probability's exponential decay in the block length [1], [2]. Lower bounds on the exponent for discrete memoryless channels (DMC) are easily obtained by random-coding techniques. In contrast, the computation of upper bounds, satisfied by every code, is more challenging since code-specific bounds need to be optimized over each possible codebook. Nevertheless, certain bounds avoid this optimization, e. g. the sphere-packing bound [3], which is exponentially tight for rates above the critical rate of the channel [1], [3].

The sphere-packing exponent has been derived using different techniques. By building on an instance of binary hypothesis testing, Shannon, Gallager and Berlekamp [3] derived an error bound with the sphere-packing exponent (SP67); also based on hypothesis testing, Blahut proposed an alternative derivation of this bound in [4]; the sphere-packing exponent was also obtained by using combinatorial methods in [5]; and based on the method of types in [2]. The works [6], [7] addressed the tightness of the SP67 bound for short to moderate block lengths by improving the pre-exponential and rate penalty terms. Recently, the metaconverse bound [8] has been shown to have the exponential decay of the sphere-packing bound [9].

Cost constraints were first included in the derivation of the sphere-packing bound in [10], by using a geometric approach for the specific case of the Gaussian channel. The SP67 [3] can also be extended to introduce cost constraints in general memoryless semicontinuous channels [1, p. 329, footnote]. In this work, we generalize the derivation of the sphere-packing exponent in [4] to consider per-codeword cost constraints. In contrast to the derivation in [3], no assumption of constant-composition codewords is needed. This allows to extend the cost-constrained sphere-packing exponent to arbitrary continuous memoryless channels. Building on this result, we establish a connection between the cost-constrained sphere-packing and Csiszár sphere-packing exponent for constant composition codes [2, Ch. 2, Th. 5.3].

This work has been funded in part by the European Research Council under ERC grant agreement 259663, by the European Union's 7th Framework Programme under grant agreements 303633 and 333680, and by the Spanish Ministry of Economy and Competitiveness under grants RYC-2011-08150, TEC2012-38800-C03-03 and FPDI-2013-18602.

II. Preliminaries

We study the problem of transmitting M equiprobable messages over a DMC using length-n block codes. The channel law is given by $W^n(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^n W(y_i|x_i)$, $\boldsymbol{x} = (x_1, \ldots, x_n)$, $x \in \mathcal{X}$, $\boldsymbol{y} = (y_1, \ldots, y_n)$, $y \in \mathcal{Y}$. We define a separable cost function $f_n(\boldsymbol{x}) \triangleq \sum_{i=1}^n f(x_i)$ with $f(x)$ denoting a real-valued scalar cost. A cost-constrained codebook $\mathcal{C}$ is defined as a set of codewords $\{\boldsymbol{x}_m\}_{m=1}^M$ such that $f_n(\boldsymbol{x}_m) \le n\xi$, $m = 1, \ldots, M$, where ξ is the per-symbol cost cap. The coding rate is $R \triangleq \frac{1}{n}\log M$.

An encoder maps the source message $m \in \{1, \ldots, M\}$ to a length-n codeword $\boldsymbol{x}_m$, which is then transmitted over the channel. The channel output $\boldsymbol{y}$ is decoded at the receiver following a maximum likelihood (ML) criterium. Let us denote the output of the decoder as $\hat{m}(\boldsymbol{y})$. Then, the error probability incurred when a message m was transmitted is

$$\epsilon_m(\mathcal{C}) \triangleq \Pr\{\hat{m}(\boldsymbol{Y}) \neq m\}, \tag{1}$$

and the error probability averaged over all codewords is thus

$$\epsilon(\mathcal{C}) \triangleq \frac{1}{M}\sum_{m=1}^{M} \epsilon_m(\mathcal{C}). \tag{2}$$

Similarly, we define the maximal error probability as $\epsilon_{\max}(\mathcal{C}) \triangleq \max_m \epsilon_m(\mathcal{C})$.

We say that an error exponent $E > 0$ is *achievable* if there exists a sequence of codes $\mathcal{C}_n$, $n = 1, 2, \ldots$, such that the average error probability $\epsilon(\mathcal{C}_n)$ is upper-bounded as

$$\epsilon(\mathcal{C}_n) \le e^{-nE+o(n)}, \tag{3}$$

where $o(n)$ satisfies $\lim_{n\to\infty} o(n)/n = 0$.

Any achievable error exponent for a DMC is upper bounded [2]–[5] as $E \le E_{\text{sp}}(R-\delta)$, for any $\delta > 0$, where the sphere-packing exponent $E_{\text{sp}}(R)$ is given by

$$E_{\text{sp}}(R) \triangleq \sup_{\rho\ge 0}\bigl\{E_0(\rho) - \rho R\bigr\}. \tag{4}$$

Gallager's E_0 function is $E_0(\rho) \triangleq \max_P E_0(\rho, P)$, with

$$E_0(\rho, P) \triangleq -\log \sum_y \left(\sum_x P(x) W(y|x)^{\frac{1}{1+\rho}} \right)^{1+\rho}. \quad (5)$$

The error exponent achievable by a sequence of cost constrained codebooks is upper bounded as $E \leq E_{\text{sp}}^{\text{cost}}(R-\delta)$, for arbitrary $\delta > 0$, where

$$E_{\text{sp}}^{\text{cost}}(R) \triangleq \sup_{\rho \geq 0} \{ E_0^{\text{cost}}(\rho) - \rho R \}, \quad (6)$$

with $E_0^{\text{cost}}(\rho) \triangleq \max_{P \in \mathcal{P}_\xi, s \geq 0} E_0^{\text{cost}}(\rho, P, s)$, $\mathcal{P}_\xi$ being the set of input distributions satisfying the cost constraint, and

$$E_0^{\text{cost}}(\rho, P, s) \triangleq -\log \sum_y \left(\sum_x P(x) e^{s(f(x)-\xi)} W(y|x)^{\frac{1}{1+\rho}} \right)^{1+\rho}. \quad (7)$$

A. Hypothesis Testing

Based on an observation $v \in \mathcal{V}$ in some alphabet $\mathcal{V}$, consider a binary hypothesis test between the hypotheses

$$\mathcal{H}_0 : \quad V \sim P_0, \quad (8)$$
$$\mathcal{H}_1 : \quad V \sim P_1, \quad (9)$$

where P_0 and P_1 are distributions over $\mathcal{V}$. For a binary hypothesis test $T : \mathcal{V} \to \{\mathcal{H}_0, \mathcal{H}_1\}$ we define the type-I error as deciding $\mathcal{H}_1$ when the true hypothesis is $\mathcal{H}_0$; and the type-II error as deciding $\mathcal{H}_0$ when the true hypothesis is $\mathcal{H}_1$. These error probabilities are respectively given by

$$\epsilon_{\text{I}}(T) = \mathbb{P}_0\{T(V) = \mathcal{H}_1\}, \quad (10)$$
$$\epsilon_{\text{II}}(T) = \mathbb{P}_1\{T(V) = \mathcal{H}_0\}, \quad (11)$$

where $\mathbb{P}_0\{\mathcal{E}\}$ and $\mathbb{P}_1\{\mathcal{E}\}$ denote the probability of the event $\mathcal{E}$ computed with respect to P_0 and P_1, respectively. We define the smallest type-I error among all (possibly randomized) tests T with a type-II error at most β as

$$\alpha_\beta(P_0, P_1) \triangleq \min_{T : \epsilon_{\text{II}}(T) \leq \beta} \epsilon_{\text{I}}(T). \quad (12)$$

A bound on the exponential behavior of lowest type-I and type-II errors was found by Blahut in [4, Th. 10]. Let us define

$$e(r) \triangleq \sup_{\rho \geq 0} \left\{ -\rho r - \log \left(\sum_v P_0(v)^{\frac{1}{1+\rho}} P_1(v)^{\frac{\rho}{1+\rho}} \right)^{1+\rho} \right\}, \quad (13)$$

and, for $\hat{\rho}$ maximizing (13) and for every v, let us define

$$\hat{p}(v) \triangleq \frac{P_0(v)^{\frac{1}{1+\hat{\rho}}} P_1(v)^{\frac{\hat{\rho}}{1+\hat{\rho}}}}{\sum_v P_0(v)^{\frac{1}{1+\hat{\rho}}} P_1(v)^{\frac{\hat{\rho}}{1+\hat{\rho}}}}, \quad (14)$$

Theorem 1 ([4, Th. 10]): Let $\nu > 0$ be given, and let $\zeta \in (0,1)$ be arbitrary. For any $\beta \leq \zeta e^{-(r+\nu)}$ we have that

$$\alpha_\beta(P_0, P_1) \geq \left(1 - \frac{\sigma_0^2 + \sigma_1^2}{\nu^2} - \zeta \right) e^{-(e(r)+\nu)}, \quad (15)$$

where σ_i^2 denotes the variance of the random variable $\log \frac{\hat{p}(V)}{P_i(V)}$ with respect to the distribution $\hat{p}$, $i = 0, 1$.

III. Cost-Constrained Sphere-Packing

For each message $m = 1, \ldots, M$, and based on the channel output $\boldsymbol{y}$, we define a binary hypothesis test between $P_0 = W^n(\cdot|\boldsymbol{x}_m)$ and $P_1 = Q^n$, where Q^n is a distribution over $\mathcal{Y}^n$ independent of m. Consider a (possibly suboptimal) bank of tests $\{T_m\}$ defined as follows. Based on the channel decoder, the test T_m decides $\mathcal{H}_0$ if $\hat{m}(\boldsymbol{y}) = m$, and $\mathcal{H}_1$ otherwise. For any partition on the output space induced by $\hat{m}(\boldsymbol{y})$, it holds

$$\sum_{m=1}^{M} \epsilon_{\text{II}}(T_m) = \sum_{\boldsymbol{y}} Q^n(\boldsymbol{y}) = 1. \quad (16)$$

Then, since $\epsilon_{\text{II}}(T_m) \geq 0$, $m = 1 \ldots, M$, there must exist at least one message m such that $\epsilon_{\text{II}}(T_m) \leq \frac{1}{M}$. In the remainder of this paper, we fix m such that $\epsilon_{\text{II}}(T_m) \leq \frac{1}{M}$.

The error probability of this message m is

$$\epsilon_m(\mathcal{C}) = \Pr\{\hat{m}(\boldsymbol{Y}) \neq m\} = \epsilon_{\text{I}}(T_m). \quad (17)$$

As (12) is a lower bound for any test, and using that $\epsilon_{\text{II}}(T_m) \leq \frac{1}{M}$, the maximal error probability can be lower bounded as

$$\epsilon_{\max}(\mathcal{C}) \;\geq\; \epsilon_m(\mathcal{C}) \;\geq\; \alpha_{\frac{1}{M}}\big(W^n(\cdot|\boldsymbol{x}_m), Q^n\big). \quad (18)$$

We now bound the exponential behavior of (18). We define

$$\Lambda_n(\rho, Q^n, \boldsymbol{x}) \triangleq -\frac{1}{n} \log \left(\sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x})^{\frac{1}{1+\rho}} Q^n(\boldsymbol{y})^{\frac{\rho}{1+\rho}} \right)^{1+\rho}, \quad (19)$$

and the sequences

$$\nu_n' \triangleq \nu_n'' + n^{-1} \log \zeta, \quad (20)$$
$$\nu_n'' \triangleq n^{\alpha - 1}, \quad 1/2 < \alpha < 1. \quad (21)$$

For sufficiently large n, $\nu_n' > 0$. Then, we apply Theorem 1 with $\nu = n\nu_n'$, $P_0 = W^n(\cdot|\boldsymbol{x}_m)$, $P_1 = Q^n$, and $r = nR - \nu + \log \zeta$. Since $\beta = \zeta e^{-(r+\nu)} = \frac{1}{M}$, Theorem 1 yields

$$\begin{aligned} E_{\max} &\triangleq \lim_{n\to\infty} -\frac{1}{n} \log \epsilon_{\max}(\mathcal{C}_n) && (22) \\ &\leq \lim_{n\to\infty} \sup_{\rho \geq 0} \{ \Lambda_n(\rho, Q^n, \boldsymbol{x}_m) - \rho (R - \nu_n'') \} \\ &\quad + \lim_{n\to\infty} \left(\nu_n' - \frac{1}{n} \log \left(1 - \frac{\sigma_0^2 + \sigma_1^2}{(n\nu_n')^2} - \zeta \right) \right). && (23) \end{aligned}$$

The second limit in (23) vanishes since, for ν_n' in (20), $\lim_{n\to\infty} \nu_n' = 0$, and $\lim_{n\to\infty} \frac{\sigma_i^2}{(n\nu_n')^2} = \lim_{n\to\infty} \frac{\sigma_i^2}{n^{2\alpha}} = 0$, since the variances σ_i^2, $i = 0, 1$, are proportional to n, and $2\alpha > 1$. Then,

$$E_{\max} \leq \lim_{n\to\infty} \sup_{\rho \geq 0} \{ \Lambda_n(\rho, Q^n, \boldsymbol{x}_m) - \rho (R - \nu_n'') \}. \quad (24)$$

For each n, we choose Q^n such that we obtain the lowest upper bound in (24). We next show that, for an arbitrary $\delta > 0$,

$$\begin{aligned} E_{\max} &\leq \lim_{n\to\infty} \inf_{Q^n} \sup_{\rho \geq 0} \{ \Lambda_n(\rho, Q^n, \boldsymbol{x}_m) - \rho (R - \nu_n'') \} && (25) \\ &\leq \lim_{n\to\infty} \sup_{\rho \geq 0} \inf_{Q^n} \{ \Lambda_n(\rho, Q^n, \boldsymbol{x}_m) - \rho (R - \delta) \}, && (26) \end{aligned}$$

where Q^n is not allowed to depend on $\boldsymbol{x}_m$.

In order to show (26), assume first that the value of ρ achieving the saddlepoint in (25) is finite. Then, there exists $\bar{\rho} < \infty$ such that

$$E_{\max} \leq \lim_{n\to\infty} \inf_{Q^n} \sup_{\rho \geq 0} \left\{\Lambda_n(\rho, Q^n, \boldsymbol{x}_m) - \rho\,(R - \nu_n'')\right\} \quad (27)$$

$$= \lim_{n\to\infty} \inf_{Q^n} \max_{0\leq\rho\leq\bar{\rho}} \left\{\Lambda_n(\rho, Q^n, \boldsymbol{x}_m) - \rho\,(R - \nu_n'')\right\} \quad (28)$$

$$= \lim_{n\to\infty} \sup_{0\leq\rho\leq\bar{\rho}} \inf_{Q^n} \left\{\Lambda_n(\rho, Q^n, \boldsymbol{x}_m) - \rho\,(R - \nu_n'')\right\}, \quad (29)$$

where in (28) we used that the saddle point is achieved at $\rho < \bar{\rho}$; and (29) follows from the Kneser-Fan minimax theorem [11, Th. 4.2], since, for fixed Q^n, the bracketed term in (28) is concave in ρ, and, for fixed ρ it is convex in Q^n. Then, (26) follows from (29) by increasing the range over which the maximization over ρ is performed, and by using that, for arbitrary $\delta > 0$ and sufficiently large n, $\nu_n'' \leq \delta$.

When the saddle point in (25) is attained at $\rho \to \infty$, we cannot apply Kneser-Fan minimax theorem. Let $\Lambda_n'(\cdot)$ denote the derivative of $\Lambda_n(\cdot)$ with respect to ρ. Since the optimizer in (25) is $\rho \to \infty$, it follows that $\Lambda_n'(\rho, Q^n, \boldsymbol{x}_m) \geq R - \nu_n''$ for all Q^n, $\rho \geq 0$. Using that, for sufficiently large n, $\Lambda_n'(\rho, Q^n, \boldsymbol{x}_m) \geq R - \nu_n'' \geq R - \delta$, the bound (26) becomes $E_{\max} \leq \infty$, which is trivially true. Then, (26) holds regardless the value of the optimizing ρ.

The dependence on the sequence of codebooks is present in (26) through $\Lambda_n(\rho, Q^n, \boldsymbol{x}_m)$. This dependence is circumvented by making use of the following property [1, Thm. 5.6.5].

Theorem 2: For $\rho \geq 0$, let

$$\mu_0(\boldsymbol{y}, \rho) \triangleq \sum_{\boldsymbol{x}} \widehat{P}^n(\boldsymbol{x}) W^n(\boldsymbol{y}|\boldsymbol{x})^{\frac{1}{1+\rho}}, \quad (30)$$

where $\widehat{P}^n$ is an exponent-achieving distribution, i. e., $\widehat{P}^n(\boldsymbol{x}) = \prod_{i=1}^n \widehat{P}(x_i)$, $\widehat{P} = \arg\max_P \left\{E_0\left(\rho, P\right)\right\}$.

Then, for any $\rho \geq 0$ and any $\boldsymbol{x}$, it holds that

$$\sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x})^{\frac{1}{1+\rho}} \mu_0(\boldsymbol{y}, \rho)^{\rho} \;\geq\; \sum_{\boldsymbol{y}} \mu_0(\boldsymbol{y}, \rho)^{1+\rho}. \quad (31)$$

In (31), only the left-hand side depends on $\boldsymbol{x}$. We define

$$Q_{0,\rho}^n(\boldsymbol{y}) \triangleq \frac{\mu_0(\boldsymbol{y}, \rho)^{1+\rho}}{\sum_{\boldsymbol{y}} \mu_0(\boldsymbol{y}, \rho)^{1+\rho}}. \quad (32)$$

Eq. (32) corresponds to that in [4, Cor. 4] (see also [2, p. 193, Prob. 23], [3, Eq. (4.20)]). Using Theorem 2 it can be verified that $\Lambda_n\big(\rho, Q_{0,\rho}^n, \boldsymbol{x}_m\big) \leq E_0(\rho)$ for every code in the sequence. Hence, by letting $Q^n = Q_{0,\rho}^n$ in (26) we obtain

$$E_{\max} \leq \sup_{\rho\geq 0} \big\{E_0(\rho) - \rho(R - \delta)\big\} = E_{\mathrm{sp}}(R - \delta). \quad (33)$$

In order to introduce a cost constraint into this formulation we make use of the following extension of Theorem 2.

Theorem 3: For $\rho \geq 0$, let

$$\mu_1(\boldsymbol{y}, \rho) \triangleq \sum_{\boldsymbol{x}} \widehat{P}^n(\boldsymbol{x}) e^{\hat{s}(f_n(\boldsymbol{x}) - n\xi)} W^n(\boldsymbol{y}|\boldsymbol{x})^{\frac{1}{1+\rho}}, \quad (34)$$

where $\widehat{P}^n(\boldsymbol{x}) = \prod_{i=1}^n \widehat{P}(x_i)$, and

$$\{\widehat{P}, \hat{s}\} = \arg \max_{P\in\mathcal{P}_\xi, s\geq 0} E_1(\rho, P, s). \quad (35)$$

For any $\boldsymbol{x}$ such that $f_n(\boldsymbol{x}) \leq n\xi$, it holds that

$$\sum_{\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x})^{\frac{1}{1+\rho}} \mu_1(\boldsymbol{y}, \rho)^{\rho} \;\geq\; \sum_{\boldsymbol{y}} \mu_1(\boldsymbol{y}, \rho)^{1+\rho}. \quad (36)$$

Proof: See Appendix A. ∎

We define

$$Q_{1,\rho}^n(\boldsymbol{y}) = \frac{\mu_1(\boldsymbol{y}, \rho)^{1+\rho}}{\sum_{\boldsymbol{y}'} \mu_1(\boldsymbol{y}', \rho)^{1+\rho}}. \quad (37)$$

Substituting $Q^n = Q_{1,\rho}^n$ in (19) yields

$$\Lambda_n\big(\rho, Q_{1,\rho}^n, \boldsymbol{x}_m\big) \leq -\frac{1}{n} \log\left(\left(\sum_{\boldsymbol{y}} \mu_1(\boldsymbol{y}, \rho)^{1+\rho}\right)^{\frac{1}{1+\rho}}\right)^{1+\rho} \quad (38)$$

$$= E_0^{\mathrm{cost}}(\rho). \quad (39)$$

where in (38) we applied Theorem 3 for $\boldsymbol{x}_m$ satisfying the cost constraint; and (39) follows from the definition of $\mu_1(\boldsymbol{y}, \rho)$ and the E_0^{cost} function.

Hence, from (26) and (38)-(39), we obtain

$$E_{\max} \leq \sup_{\rho\geq 0} \big\{E_0^{\mathrm{cost}}(\rho) - \rho(R - \delta)\big\} = E_{\mathrm{sp}}^{\mathrm{cost}}(R - \delta). \quad (40)$$

For any code, $\epsilon(\mathcal{C}_n) \geq \frac{1}{2}\epsilon_{\max}(\mathcal{C}_n')$ where $\mathcal{C}_n'$ is an expurgated code obtained by removing from $\mathcal{C}_n$ the $M/2$ codewords with highest error probability [4, Th. 20]. As the rate R is unaffected by the expurgation, combining (33) and (40) yields the following result.

Theorem 4: For a memoryless channel W^n, let E denote the error exponent achievable by a sequence of codebooks $\mathcal{C}_n$, $n = 1, 2, \ldots$, such that, for each value of n, the codewords satisfy an individual (separable) cost constraint $f_n(\boldsymbol{x}_m) \leq n\xi$, $m = 1, \ldots, M$. It follows that, for any $\delta > 0$,

$$E \leq \sup_{\rho\geq 0} \big\{\min\big(E_0(\rho), E_0^{\mathrm{cost}}(\rho)\big) - \rho(R - \delta)\big\}. \quad (41)$$

If the cost constraint is active for any P achieving $E_1(\rho)$, then $E_1(\rho) \leq E_0(\rho)$. However, if the cost constraint is non-active, $E_0(\rho) \leq E_1(\rho)$. Therefore, neither $E_{\mathrm{sp}}(R)$ nor $E_{\mathrm{sp}}^{\mathrm{cost}}(R)$ dominates in general.

This theorem applies to codebooks satisfying a per-codeword cost constraint. Extending Theorem 4 to codebooks satisfying an average cost constraint is still an open problem.

IV. Connection with Constant Composition Codes

For a given n, consider a constant composition code with empirical distribution P_n, i. e., every codeword $\boldsymbol{x}$ belonging to $\mathcal{C}_n$ has a composition equal to P_n. We fix Q^n to be an arbitrary product distribution, $Q^n(\boldsymbol{y}) = \prod_{i=1}^n Q(y_i)$. In this case $\Lambda_n\big(\rho, Q^n, \boldsymbol{x}\big)$ is independent of the specific code,

$$\Lambda_n\big(\rho, Q^n, \boldsymbol{x}\big) = -\sum_{x} \mathsf{P}_n(x) \log\left(\sum_{y} W(y|x)^{\frac{1}{1+\rho}} Q(y)^{\frac{\rho}{1+\rho}}\right)^{1+\rho}. \quad (42)$$

For any sequence of constant composition codes such that $\mathsf{P}_n \to \mathsf{P}$ as $n \to \infty$, from (26) and (42) it follows that

$$E_{\max} \le \sup_{\rho \ge 0} \big\{ E_1(\rho, \mathsf{P}) - \rho(R - \delta) \big\}, \tag{43}$$

where

$$E_1(\rho, P) \triangleq \min_Q \left\{ - \sum_x P(x) \log \left(\sum_y W(y|x)^{\frac{1}{1+\rho}} Q(y)^{\frac{\rho}{1+\rho}} \right)^{1+\rho} \right\}. \tag{44}$$

Eq. (43) corresponds to Csiszár sphere-packing bound for constant composition codes [2, Ch. 2, Th. 5.3]. Optimizing (43) over compositions that satisfy the cost constraint and applying the expurgation argument, we obtain

$$E \le \sup_{\rho \ge 0} \big\{ E_1(\rho) - \rho(R - \delta) \big\}, \tag{45}$$

where $E_1(\rho) \triangleq \max_{P \in \mathcal{P}_\xi} E_1(\rho, P)$.

We next show that (45) coincides with (41) in Theorem 4. To this end, let us consider the dual formulation of $E_1(\rho)$, which is given by a double optimization over distributions P satisfying the cost constraint, and over functions $a : \mathcal{X} \to \mathbb{R}$ with finite average $\bar{a} \triangleq \sum_x P(x) a(x)$ [12, Th. 3.4]:

$$E_1(\rho) = \max_{P \in \mathcal{P}_\xi, a} E_1(\rho, P, a), \tag{46}$$

$$E_1(\rho, P, a) \triangleq -\log \sum_y \left(\sum_x P(x) e^{a(x) - \bar{a}} W(y|x)^{\frac{1}{1+\rho}} \right)^{1+\rho}. \tag{47}$$

Note the similarity between $E_0^{\text{cost}}(\rho, P, s)$ in (7) and (47). While $a(x)$ in the definition of $E_1(\rho, P, a)$ in (47) is an arbitrary function to be optimized, $f(x)$ in (7) denotes the cost function, which is given.

Appendix B derives the optimality conditions for the optimization problem in (46). Let us define $P_0 \triangleq \arg\max_P E_0(\rho, P)$. When $P_0 \in \mathcal{P}_\xi$, the cost constraint is not active in (46), and the maximum is attained for $a(x) = \bar{a}$, $\forall x$. Hence, in this case $E_1(\rho, P, a)$ becomes $E_0(\rho, P)$, and $E_1(\rho) = E_0(\rho)$. In contrast, for $P_0 \notin \mathcal{P}_\xi$, the optimizing $a(\cdot)$ is $a(x) = s f(x)$, $\forall x$, for some $s \ge 0$. Using that the cost constraint holds with equality in this case, we obtain $E_1(\rho) = E_0^{\text{cost}}(\rho)$. By combining both possibilities, (46) yields $E_1(\rho) = \min\big\{E_0(\rho), E_0^{\text{cost}}(\rho)\big\}$ and (45) coincides with (41).

Appendix A
Proof of Theorem 3

Let $\rho \ge 0$ be fixed. Let us define

$$\Phi_y(P, s) \triangleq \sum_x P(x) e^{s(f(x) - \xi)} W(y|x)^{\frac{1}{1+\rho}}. \tag{48}$$

$$\Psi_y(P, s) \triangleq \sum_x P(x)(f(x) - \xi) e^{s(f(x) - \xi)} W(y|x)^{\frac{1}{1+\rho}}. \tag{49}$$

We study the optimality conditions of the following optimization problem, which is equivalent to $\max_{P \in \mathcal{P}_\xi, s \ge 0} E_1(\rho, P, s)$,

$$\begin{aligned} \min_{P,s} \quad & \sum_y \Phi_y(P, s)^{1+\rho}, \\ \text{subject to} \quad & s \ge 0,\ P(x) \ge 0, \\ & \sum_x P(x) = 1, \\ & \sum_x P(x) f(x) \le \xi. \end{aligned} \tag{50}$$

The Lagrangian of the optimization problem in (50) is

$$\begin{aligned} \mathcal{L}(P, s) = & \sum_y \Phi_y(P, s)^{1+\rho} - \sigma s - \sum_x \eta_x P(x) \\ & - \lambda \left(\sum_x P(x) - 1 \right) - \gamma \sum_x P(x) \big(\xi - f(x) \big), \end{aligned} \tag{51}$$

where $\sigma \ge 0$, $\eta_{\boldsymbol{x}} \ge 0$, $\lambda \in \mathbb{R}$ and $\gamma \ge 0$ are the Lagrange multipliers associated to the respective constraints in (50).

Let us denote by $\widehat{P}$, $\hat{s}$ the values of P, s optimizing (50). Similarly, let us define $\widehat{\Phi}_y \triangleq \Phi_y\big(\widehat{P}, \hat{s}\big)$, $\widehat{\Psi}_y \triangleq \Psi_y\big(\widehat{P}, \hat{s}\big)$. By taking the derivative of (51) with respect to $P(x)$ and equating it to zero we obtain the following optimality condition

$$\begin{aligned} (1+\rho) \sum_y e^{\hat{s}(f(x) - \xi)} W(y|x)^{\frac{1}{1+\rho}} \big(\widehat{\Phi}_y\big)^\rho & \\ = \eta_x + \lambda + \gamma \big(\xi - f(x) \big). & \end{aligned} \tag{52}$$

Likewise, by taking the derivative of (51) with respect to s and equating it to zero yields the condition

$$(1+\rho) \sum_y \widehat{\Psi}_y \big(\widehat{\Phi}_y\big)^\rho = \sigma. \tag{53}$$

Multiplying both sides of (52) by $\widehat{P}(x)$, summing over x, and using that due to complementary slackness [13, Sec. 5.5.2], $\eta_x \widehat{P}(x) = 0$, $\gamma \sum_x \widehat{P}(x)\big(\xi - f(x)\big) = 0$, we obtain

$$\lambda = (1+\rho) \sum_y \big(\widehat{\Phi}_y\big)^{1+\rho}. \tag{54}$$

Multiplying (52) by $\widehat{P}(x)(f(x) - \xi)$, summing over x, yields

$$\begin{aligned} (1+\rho) \sum_y \widehat{\Psi}_y \big(\widehat{\Phi}_y\big)^\rho = & \sum_x \eta_x \widehat{P}(x) \big(f(x) - \xi \big) \\ & + \lambda \sum_x \widehat{P}(x) \big(f(x) - \xi \big) - \gamma \sum_x \widehat{P}(x) \big(f(x) - \xi \big)^2. \end{aligned} \tag{55}$$

Substituting (53) in (55), using that $\eta_x \widehat{P}(x) = 0$, we obtain

$$\sigma = \lambda \sum_x \widehat{P}(x) \big(f(x) - \xi \big) - \gamma \sum_x \widehat{P}(x) \big(f(x) - \xi \big)^2. \tag{56}$$

If the cost constraint is non-active, i. e., $\sum_x \widehat{P}(x) f(x) < \xi$, the corresponding Lagrange multiplier is $\gamma = 0$ due to complementary slackness. If the cost constraint is active, $\sum_x \widehat{P}(x) f(x) = \xi$, and from (56) we obtain

$$\sigma = -\gamma \sum_x \widehat{P}(x) \big(f(x) - \xi \big)^2. \tag{57}$$

Since $\sigma \geq 0$, $\gamma \geq 0$, from (57) we conclude that $\sigma = \gamma = 0$, so in either case $\gamma = 0$. Substituting (54) into (52), yields

$$\sum_y e^{\hat{s}(f(x)-\xi)} W(y|x)^{\frac{1}{1+\rho}} \big(\widehat{\Phi}_y\big)^{\rho} = \sum_y \big(\widehat{\Phi}_y\big)^{1+\rho} + \frac{\eta_x}{1+\rho}, \tag{58}$$

and since $\eta_x \geq 0$,

$$\sum_y e^{\hat{s}(f(x)-\xi)} W(y|x)^{\frac{1}{1+\rho}} \big(\widehat{\Phi}_y\big)^{\rho} \geq \sum_y (\widehat{\Phi}_y)^{1+\rho}. \tag{59}$$

Finally, we use the definition of $\mu_1(\boldsymbol{y},\rho)$ in (34) to write

$$\begin{aligned}\sum_{\boldsymbol{y}} & W^n(\boldsymbol{y}|\boldsymbol{x})^{\frac{1}{1+\rho}} \mu_1(\boldsymbol{y},\rho)^{\rho} \\ &\geq \sum_{\boldsymbol{y}} e^{\hat{s}(f_n(\boldsymbol{x})-n\xi)} W^n(\boldsymbol{y}|\boldsymbol{x})^{\frac{1}{1+\rho}} \mu_1(\boldsymbol{y},\rho)^{\rho} && (60)\\ &\geq \sum_{\boldsymbol{y}} \mu_1(\boldsymbol{y},\rho)^{1+\rho}, && (61)\end{aligned}$$

where in (60) we used that, by assumption, $f_n(\boldsymbol{x}) \leq n\xi$; and (61) follows from factorizing (60) and applying (59) to each of the factors. The step (60) holds with equality for $f_n(\boldsymbol{x}) = n\xi$, and the step (61) is tight as long as $\widehat{P}^n(\boldsymbol{x}) > 0$.

APPENDIX B
OPTIMALITY CONDITIONS FOR $E_1(\rho)$ IN (46)

The Lagrangian of the optimization problem (46) is

$$\begin{aligned}\mathcal{L}(P,a) &= (1+\rho)\sum_x P(x)a(x) \\ &\quad - \log \sum_y \left(\sum_x P(x) e^{a(x)} W(y|x)^{\frac{1}{1+\rho}}\right)^{1+\rho} \\ &\quad - \lambda\left(\sum_x P(x) - 1\right) - \gamma \sum_x P(x)\big(f(x)-\xi\big), \qquad (62)\end{aligned}$$

where $\lambda \in \mathbb{R}$ and $\gamma \geq 0$ are the Lagrange multipliers associated to the constraints $\sum_x P(x) = 1$ and $\sum_x P(x)f(x) \leq \xi$, respectively.

Let $\widehat{P}$, $\hat{a}(\cdot)$, denote the values of P, $a(\cdot)$ optimizing (46). Setting the derivative of $\mathcal{L}(P,a)$ with respect to $a(x)$ to zero, we obtain the following optimality condition,

$$\frac{\sum_y e^{\hat{a}(x)} W(y|x)^{\frac{1}{1+\rho}} \left(\sum_{x'} \widehat{P}(x') e^{\hat{a}(x')} W(y|x'')^{\frac{1}{1+\rho}}\right)^{\rho}}{\sum_y \left(\sum_{x''} \widehat{P}(x'') e^{\hat{a}(x'')} W(y|x'')^{\frac{1}{1+\rho}}\right)^{1+\rho}} = 1, \tag{63}$$

for $x \in \mathcal{X}$. Equating to zero the derivative of $\mathcal{L}(P,a)$ with respect to $P(x)$, and using (63) it follows that for $\widehat{P}$, $\hat{a}(\cdot)$,

$$(1+\rho)\big(\hat{a}(x)-1\big) - \lambda - \gamma\big(f(x)-\xi\big) = 0. \tag{64}$$

Due to complementary slackness, $\gamma \sum_x \widehat{P}(x)\big(f(x)-\xi\big) = 0$. Then, multiplying (64) by $\widehat{P}(x)$, summing over $x \in \mathcal{X}$, yields

$$\lambda = (1+\rho)\sum_x \widehat{P}(x)\big(\hat{a}(x)-1\big). \tag{65}$$

Substituting (65) in (64), we obtain that, for $x \in \mathcal{X}$,

$$\hat{a}(x) - \sum_{\bar{x}} \widehat{P}(\bar{x})\hat{a}(\bar{x}) = \frac{\gamma}{1+\rho}\big(f(x)-\xi\big). \tag{66}$$

When the cost constraint is inactive its associated Lagrange multiplier is $\gamma = 0$. Hence, from (66) we obtain that $\hat{a}(x) = \sum_{\bar{x}} \widehat{P}(\bar{x})\hat{a}(\bar{x})$, $x \in \mathcal{X}$, is a constant. On the other hand, when the cost constraint is active, $\sum_x \widehat{P}(x)f(x) = \xi$ and $\gamma \geq 0$. Then, (66) yields $\hat{a}(x) = sf(x)$ with $s = \frac{\gamma}{1+\rho} \geq 0$.

ACKNOWLEDGEMENT

The authors wish to thank an anonymous reviewer for his useful remarks on the compactness condition in the minimax theorem.

REFERENCES

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

[2] I. Csiszár and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.

[3] C. E. Shannon, R. G. Gallager, and E. R. Berlekamp, "Lower bounds to error probability for coding on discrete memoryless channels. I," *Inf. Contr.*, vol. 10, no. 1, pp. 65–103, Jan. 1967.

[4] R. Blahut, "Hypothesis testing and information theory," *IEEE Trans. Inf. Theory*, vol. 20, no. 4, pp. 405–417, Jul. 1974.

[5] E. A. Haroutunian, "Estimates of the error exponents for the semicontinuous memoryless channel," *Probl. Per. Inf.*, vol. 4, pp. 37–48, 1968, in Russian.

[6] A. Valembois and M. Fossorier, "Sphere-packing bounds revisited for moderate block lengths," *IEEE Trans. Inf. Theory*, vol. 50, no. 12, pp. 2998–3014, Dec. 2004.

[7] G. Wiechman and I. Sason, "An improved sphere-packing bound for finite-length codes over symmetric memoryless channels," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 1962–1990, May 2008.

[8] Y. Polyanskiy, H. 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.

[9] Y. Polyanskiy, "Saddle point in the minimax converse for channel coding," *IEEE Trans. Inf. Theory*, vol. 59, no. 5, pp. 2576–2595, May 2013.

[10] C. E. Shannon, "Probability of error for optimal codes in a gaussian channel," *Bell Syst. Tech. J.*, vol. 38, pp. 611–656, 1959.

[11] M. Sion, "On general minimax theorems," *Pacific J. of Math.*, vol. 8, no. 1, pp. 171–176, 1958.

[12] G. S. Poltyrev, "Random coding bounds for discrete memoryless channels," *Probl. Per. Inf.*, vol. 18, no. 1, pp. 12–26, 1982, in Russian.

[13] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, US: Cambridge University Press, 2004.