On exponential error bounds for random codes on the DMC

Andrea Montanari, G. David Forney, Jr.

1 Introduction

The purpose of this note is to show that Gallager's celebrated exponential error bounds for the DMC can be obtained by a more or less straightforward application of large-deviation theory and convex analysis. Probably these results have been derived previously by others, such as Csiszár and Körner.

2 DMC/RCE/MLD Model

A discrete memoryless channel (DMC) is characterized by discrete input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$ and a transition probability matrix $\{p(x|y), x \in \mathcal{X}, y \in \mathcal{Y}\}$, such that the probability that the sequence $\mathbf{y} = (y_1, y_2, \ldots, y_N)$ is received if $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ is transmitted is

$$p(\mathbf{y}|\mathbf{x}) = \prod_{k=1}^{N} p(y_k|x_k).$$

A block code of length N and rate R for such a channel is a set $\mathcal{C}$ of $M = e^{NR}$ code sequences $\mathbf{x}_i, 1 \leq i \leq M$, of length N.

A random code ensemble (RCE) of length N and rate R for such a channel is characterized by an input probability distribution $\{p(x), x \in \mathcal{X}\}$, and consists of a set of $M = e^{NR}$ code sequences of length N in which each symbol in each sequence is chosen independently at random according to the distribution $p(x)$.

Thus if a codeword $\mathbf{x}$ is selected at random and sent over the channel, the joint probability $p(\mathbf{x}, \mathbf{y})$ is the product of independent per-symbol joint probabilities $p(x_k, y_k) = p(x_k)p(y_k|x_k)$. The per-symbol marginal output probability distribution is $p(y) = \sum_{\mathcal{Y}} p(x, y)$.

Let $\{q(x, y), x \in \mathcal{X}, y \in \mathcal{Y}\}$ denote the relative frequency distribution of the symbol pairs (x, y) in a particular sample $(\mathbf{x}, \mathbf{y})$ drawn according to the RCE/DMC model. Similarly let $\{q(x), x \in \mathcal{X}\}$ and $\{q(y), y \in \mathcal{Y}\}$ denote the marginal relative frequency distributions, and $q(y|x)$ and $q(x|y)$ the conditional relative frequency distributions. Such sample distributions are called "types."

We will be interested in the average probability of decoding error $\Pr(E)$ under a maximum-likelihood decoding (MLD) rule. One way of characterizing such a rule is as follows. Define the negative log likelihood function

$$h(x, y) = -\log p(y|x),$$

where in these notes all logarithms will be natural. A physicist might call $h(x, y)$ the Hamiltonian of the coupling between x and y. Given a received sequence $\mathbf{y}$, the MLD rule is then to choose the codeword $\mathbf{x} \in \mathcal{C}$ that minimizes the sample average "energy"

$$H(\mathbf{x}, \mathbf{y}) = \langle h(x, y) \rangle = \sum_{\mathcal{X}, \mathcal{Y}} q(x, y) h(x, y) = -\log p(\mathbf{y}|\mathbf{x}).$$

3 Large-deviation analysis

We use an output-centered analysis in which we first choose the "correct" codeword $\mathbf{x}$ and the output sequence $\mathbf{y}$ according to the joint probability distribution $p(\mathbf{x}, \mathbf{y})$. We can do this by first choosing $\mathbf{y}$ according to $p(\mathbf{y})$ and then $\mathbf{x}$ according to $p(\mathbf{x}|\mathbf{y})$.

We then choose the remaining $M - 1$ "incorrect" codewords $\mathbf{x}'$ independently according to the probability distribution $p(\mathbf{x}')$. The joint probability of any incorrect codeword $\mathbf{x}'$ and the output sequence $\mathbf{y}$ is therefore $p(\mathbf{x}', \mathbf{y}) = p(\mathbf{x}')p(\mathbf{y})$.

An error event E occurs only if the sample average coupling $H(\mathbf{x}, \mathbf{y})$ between the correct codeword $\mathbf{x}$ and the output sequence $\mathbf{y}$ is greater than or equal to the minimum coupling $\min_{\mathbf{x}' \neq \mathbf{x}} H(\mathbf{x}', \mathbf{y})$ between $\mathbf{y}$ and any incorrect codeword $\mathbf{x}' \neq \mathbf{x}$.

3.1 Output sequence

By large-deviation theory, the probability $P(q||p)$ under $p(y)$ that the output sequence $\mathbf{y}$ has relative frequency distribution (type) $q(y)$ is given by

$$P(q||p) \approx e^{-ND(q||p)},$$

where "$\approx$" means that the exponent is correct, and $D(q||p)$ is the Kullback-Leibler divergence

$$D(q||p) = \sum_{\mathcal{Y}} q(y) \log \frac{q(y)}{p(y)}.$$

The divergence is a strictly convex function of the sample distribution $q(y)$, with a unique minimum of 0 when $q(y) = p(y)$.

3.2 Correct codeword

For the correct codeword, given the output sequence $\mathbf{y}$ and its type $q(y)$, the exponent is given by the minimum type exponent. For N large, given $q(y)$, the total probability under $p(x|y)$ of all sequences of type $q(x|y)$ is approximately equal to

$$P_c(q(x|y)|q(y)) \approx \exp\left\{-N \sum_{\mathcal{Y}} q(y) D_y(q||p_c)\right\},$$

where

$$D_y(q||p_c) = \sum_{\mathcal{X}} q(x|y) \log \frac{q(x|y)}{p(x|y)}.$$

Now, given $q(y)$, let $q(x|y)$ be any set of conditional probability distributions such that $\langle h(x, y) \rangle \geq \tau$. Then the probability of the event $\{\langle h(x, y) \rangle \geq \tau\}$ is lowerbounded by the probability of any type $q(x|y)$ satisfying this constraint:

$$\Pr(\langle h(x, y) \rangle \geq \tau | q(y)) \geq \max_{q(x|y): \langle h(x,y) \rangle \geq \tau} P_c(q(x|y)|q(y)).$$

Thus we obtain a lower bound involving the minimal type exponent

$$E_t(\tau) = \min_{q(x|y): \langle h(x,y) \rangle \geq \tau} \sum_{\mathcal{Y}} q(y) D_y(q||p_c).$$

To find this exponent, we may introduce a Lagrange multiplier s and a Lagrangian

$$\sum_{\mathcal{Y}} q(y) D_y(q||p_c) - s(\langle h(x,y)\rangle - \tau) = \sum_{\mathcal{X},\mathcal{Y}} q(y)q(x|y) \log \frac{q(x|y)}{p(x|y)} - s(\sum_{\mathcal{X},\mathcal{Y}} q(y)q(x|y)h(x,y) - \tau).$$

Then

$$E_t(\tau) = \max_s \min_{q(x|y)} \sum_{\mathcal{Y}} q(y) D_y(q||p_c) - s(\langle h(x,y)\rangle - \tau).$$

Fixing s, this is solved for each $y \in \mathcal{Y}$ by a tilted distribution $q(x|y) = Z_c(y,s)^{-1} p(x|y) e^{sh(x,y)}$, where $Z_c(y,s) = \sum_{\mathcal{X}} p(x|y) e^{sh(x,y)}$. Substituting back, we obtain

$$E_t(\tau) = \max_s s\tau - \sum_{\mathcal{Y}} q(y) \log Z_c(y,s).$$

We recognize this as the exponent of the Chernoff upper bound

$$E_c(\tau) = \max_s s\tau - \mu_c(s),$$

where $\mu_c(s)$ is the average semi-invariant moment-generating function

$$\mu_c(s) = \mathsf{E}_y \log \mathsf{E}_{x|y}[e^{sh(x,y)}] = \sum_{\mathcal{Y}} q(y) \log \sum_{\mathcal{X}} p(x|y) e^{sh(x,y)},$$

which physicists will recognize as an average negative free energy.

In conclusion, given the output type $q(y)$, the exponent of the probability that $\langle h(x,y)\rangle \geq \tau$ for the correct codeword is the minimal type exponent

$$E_t(\tau) = \min_{q(x|y):\langle h(x,y)\rangle \geq \tau} \sum_{\mathcal{Y}} q(y) D_y(q||p_c),$$

or equivalently the Chernoff exponent

$$E_c(\tau) = \max_s s\tau - \mu_c(s).$$

3.3 Incorrect subset

Similarly, for each incorrect codeword, given $q(y)$, the total probability under $p(x)$ of all sequences with conditional type $q(x|y)$ is exponentially equal to

$$P_i(q(x|y)|q(y)) \approx \exp\{-N \sum_{\mathcal{Y}} p(y) D_y(q||p_i)\},$$

where

$$D_y(q||p_i) = \sum_{\mathcal{X}} q(x|y) \log \frac{q(x|y)}{p(x)}.$$

Now, given $q(y)$, let $q(x|y)$ be any set of conditional distributions on $\mathcal{X}$ such that $\langle h(x,y)\rangle \leq \tau$. Then the probability of the event $\{\langle h(x,y)\rangle \leq \tau\}$ is lowerbounded by:

$$\Pr(\langle h(x,y)\rangle \leq \tau | q(y)) \geq \max_{q(x|y):\langle h(x,y)\rangle \leq \tau} P_c(q(x|y)|q(y)).$$

Thus we obtain a lower bound involving the minimal type exponent

$$E_{t,i}(\tau) = \min_{q(x|y):\langle h(x,y)\rangle \le \tau} \sum_{\mathcal{Y}} p(y) D_y(q||p_i).$$

In this case, solving with a Lagrange multiplier s, we find that this is solved by tilted distributions $q(x|y) = Z_i(y,s)^{-1} p(x) e^{sh(x,y)}$, where $Z_i(y,s) = \sum_{\mathcal{X}} p(x) e^{sh(x,y)}$. This yields the exponent of the Chernoff upper bound

$$E_i(\tau) = \max_s s\tau - \mu_i(s),$$

where

$$\mu_i(s) = \sum_{\mathcal{Y}} q(y) \log Z_i(y,s) = \sum_{\mathcal{Y}} q(y) \log \sum_{\mathcal{X}} p(x) e^{sh(x,y)}.$$

4 Channel capacity

We now use these results to show that $\Pr(E)$ decreases exponentially with N for all rates R less than the mutual information $I(X;Y)$ under the joint probability $p(x,y) = p(x)p(y|x)$.

The expectation of the coupling $h(x,y) = -\log p(y|x)$ under $p(x,y)$ is the conditional entropy

$$H(Y|X) = -\sum_{\mathcal{X},\mathcal{Y}} p(x,y) \log p(y|x).$$

The derivation above, or the Chernoff-bound law of large numbers, shows that for the correct codeword the probability that $\langle h(x,y)\rangle \ge \tau$ decreases exponentially with N if $\tau > H(Y|X)$.

If $q(y) = p(y)$, which is the most likely distribution, then the exponent of the probability that any incorrect codeword has coupling $\langle h(x,y)\rangle \le \tau$ is

$$E_i(\tau) = \max_s s\tau - \mu_i(s) = \max_s s\tau - \sum_{\mathcal{Y}} p(y) \log \sum_{\mathcal{X}} p(x) e^{sh(x,y)}.$$

The expected number of incorrect codewords with coupling τ is thus

$$\mathcal{N}(\tau) \approx e^{N(R - E_i(\tau))}.$$

For $R > E_i(\tau)$, by the laws of large numbers, there are almost surely an exponentially large number of incorrect codewords with coupling τ. For $R < E_i(\tau)$, on the other hand, the probability that there are any incorrect codewords with coupling τ is exponentially small.

We note that for $s = -1$ and $\tau = H(Y|X)$ the exponent $s\tau - \mu_i(s)$ is equal to

$$-H(Y|X) - \sum_{\mathcal{Y}} p(y) \log \sum_{\mathcal{X}} p(x)p(y|x) = -H(Y|X) + H(Y) = I(X;Y),$$

and furthermore that $\mu_i'(-1) = H(Y|X)$, so that this is indeed the minimum exponent. In short,

$$E_i(H(Y|X)) = I(X;Y).$$

By continuity, for any $\varepsilon > 0$ there is a $\tau > H(Y|X)$ such that $E_i(\tau) > I(X;Y) - \varepsilon$. Then for any $R < I(X;Y) - \varepsilon$, the probability that any incorrect codeword has coupling τ is exponentially small. Thus $\Pr(E)$ is exponentially small with MLD for every $R < I(X;Y)$.

The same argument shows that $\Pr(E)$ is exponentially small for every $R < I(X;Y)$ with the following typical sequence decoding method: decode to the unique codeword $\mathbf{x} \in \mathcal{C}$ with $\langle h(x,y)\rangle \le H(Y|X) + (I(X;Y) - R)/2$, provided that there is one and only one such codeword.

5 Error exponent

We now extend the above argument to find the correct error exponent $E(R)$ for all $R < I(X;Y)$.

We will first find the error exponent $E(R|q(y))$ for a given output type $q(y)$, and then minimize the joint exponent $E(R|q(y)) + D(q||p)$ to find $E(R)$; *i.e.*,

$$E(R) = \min_{q(y)} E(R|q(y)) + D(q||p).$$

The error exponent $E(R|q(y))$ is the exponent of the minimum over τ of the joint probability that $\langle h(x,y)\rangle \geq \tau$ for the correct codeword and $\langle h(x',y)\rangle \leq \tau$ for some incorrect codeword. For the correct codeword, we have

$$E_c(\tau) = \max_{s\geq 0} s\tau - \mu_c(s) = \max_{s\geq 0} s\tau - \sum_{\mathcal{Y}} q(y)\log\sum_{\mathcal{X}} p(x|y)e^{sh(x,y)}.$$

For each incorrect codeword, we have

$$E_i(\tau) = \max_{s} s\tau - \mu_i(s) = \max_{s} s\tau - \sum_{\mathcal{Y}} q(y)\log\sum_{\mathcal{X}} p(x)e^{sh(x,y)}.$$

By the union bound, the exponent of the probability that any incorrect codeword has $\langle h(x',y)\rangle \leq \tau$ is

$$E_{i,\min}(\tau) = \begin{cases} E_i(\tau) - R, & \text{if } R \leq E_i(\tau); \\ 0, & \text{if } R \geq E_i(\tau). \end{cases}$$

This expression may be written as

$$E_{i,\min}(\tau) = \max_{0\leq\rho\leq 1} \rho(E_i(\tau) - R).$$

(The parameter ρ may be characterized as a range-limited Lagrange multiplier, from a certain derivation of the sphere-packing bound.) Then we have

$$\begin{aligned} E(R|q(y)) &= \min_{\tau} E_c(\tau) + E_{i,\min}(\tau); \\ &= \min_{\tau}\max_{0\leq\rho\leq 1} E_c(\tau) + \rho(E_i(\tau) - R). \end{aligned}$$

Since $E_c(\tau)$ increases with τ, it is clear that we are not interested in any τ greater than the solution of $R = E_i(\tau)$, which we call the GV coupling $\tau_{\mathrm{GV}}(R)$. Thus the optimum τ satisfies $R \leq E_i(\tau)$. If $R < E_i(\tau)$, then the maximum occurs for $\rho = 1$; if $R = E_i(\tau)$, then $\tau = \tau_{\mathrm{GV}}(R)$, and we are free to maximize over $0 \leq \rho \leq 1$.

5.1 Calculation

As we have just seen, $E(R|q(y)) = E_0(\rho, q(y)) - \rho R$, where

$$E_0(\rho, q(y)) = \min_{\tau} E_c(\tau) + \rho E_i(\tau) = \min_{\tau} E_0(\rho,\tau),$$

where

$$\begin{aligned} E_0(\rho,\tau) &= \max_{s_1,s_2} (s_1 + s_2\rho)\tau - \sum_{\mathcal{Y}} q(y)\log\sum_{\mathcal{X}} p(x|y)e^{s_1 h(x,y)} - \sum_{\mathcal{Y}} q(y)\log\left(\sum_{\mathcal{X}} p(x)e^{s_2 h(x,y)}\right)^{\rho} \\ &= \max_{s} s\tau - \tilde{\mu}(\rho, s), \end{aligned}$$

where

$$\tilde{\mu}(\rho, s) = \min_{s_1} \sum_{\mathcal{Y}} q(y) \log \sum_{\mathcal{X}} p(x|y) e^{s_1 h(x,y)} + \sum_{\mathcal{Y}} q(y) \log \left(\sum_{\mathcal{X}} p(x) e^{(s-s_1)h(x,y)/\rho} \right)^{\rho}.$$

Since

$$E_0(\rho, \tau) = \max_s s\tau - \tilde{\mu}(\rho, s),$$

has the form of a Fenchel (Lagrange) transform, we have the inverse transform

$$\tilde{\mu}(\rho, s) = \max_\tau s\tau - E_0(\rho, \tau).$$

In particular,

$$-\tilde{\mu}(\rho, 0) = \min_\tau E_0(\rho, \tau).$$

In other words, in the expression

$$E_0(\rho, q(y)) = \min_\tau \max_s s\tau - \tilde{\mu}(\rho, s),$$

the maximum over s occurs at $s = 0$. Therefore

$$E_0(\rho, \tau) = \max_s s\tau - \tilde{\mu}(\rho, s) = -\tilde{\mu}(\rho, 0),$$

independent of τ. Now

$$\begin{aligned} -\tilde{\mu}(\rho, 0) &= -\min_{s_1} \sum_{\mathcal{Y}} q(y) \log \sum_{\mathcal{X}} p(x|y) e^{s_1 h(x,y)} + \sum_{\mathcal{Y}} q(y) \log \left(\sum_{\mathcal{X}} p(x) e^{-s_1 h(x,y)/\rho} \right)^{\rho} \\ &= \sum_{\mathcal{Y}} q(y) \log p(y) - \min_{s_1} \sum_{\mathcal{Y}} q(y) \log \left(\sum_{\mathcal{X}} p(x) p(y|x)^{1-s_1} \right) \left(\sum_{\mathcal{X}} p(x) p(y|x)^{s_1/\rho} \right)^{\rho}. \end{aligned}$$

The derivative with respect to s_1 is zero when $s_1 = \rho/(1+\rho)$, so

$$E_0(\rho, q(y)) = -\tilde{\mu}(\rho, 0) = \sum_{\mathcal{Y}} q(y) \log p(y) - \sum_{\mathcal{Y}} q(y) \log \left(\sum_{\mathcal{X}} p(x) p(y|x)^{1/(1+\rho)} \right)^{1+\rho}.$$

Finally, we need to find the $q(y)$ that minimizes

$$E_0(\rho, q(y)) + D(q||p) = \sum_{\mathcal{Y}} q(y) \log q(y) - \sum_{\mathcal{Y}} q(y) \log \left(\sum_{\mathcal{X}} p(x) p(y|x)^{1/(1+\rho)} \right)^{1+\rho},$$

subject to $\sum_{\mathcal{Y}} q(y) = 1$. We write $f(\rho, y) = (\sum_{\mathcal{X}} p(x) p(y|x)^{1/(1+\rho)})^{1+\rho}$; then we find $q(y) = f(\rho, y)/Z(\rho)$, where

$$Z(\rho) = \sum_{\mathcal{Y}} f(y) = \sum_{\mathcal{Y}} \left(\sum_{\mathcal{X}} p(x) p(y|x)^{1/(1+\rho)} \right)^{1+\rho},$$

and

$$E_0(\rho) = \min_{\{q(y)\}} E_0(\rho, q(y)) + D(q||p) = -\log Z(\rho) = -\log \sum_{\mathcal{Y}} \left(\sum_{\mathcal{X}} p(x) p(y|x)^{1/(1+\rho)} \right)^{1+\rho},$$

which is Gallager's celebrated $E_0(\rho)$ function. Then

$$E(R) = \max_{0 \le \rho \le 1} E_0(\rho) - \rho R.$$

5.2 Recapitulation

In summary, we have shown that the correct error exponent of the DMC/RCE/MLD model is

$$E(R) = \max_{0 \le \rho \le 1} \min_{\{q(y)\}} \min_{\tau} \max_{s_1, s_2} E_0(\rho, \tau, s_1, s_2, q(y)),$$

where

$$E_0(\rho, \tau, s_1, s_2, q(y)) = (s_1 + s_2\rho)\tau - \sum_{\mathcal{Y}} q(y) \log \left(\sum_{\mathcal{X}} p(x|y) e^{s_1 h(x,y)} \right) \left(\sum_{\mathcal{X}} p(x) e^{s_2 h(x,y)} \right)^{\rho} \frac{p(y)}{q(y)}.$$

Moreover, we have shown that the maximum occurs for the following parameters:

$$\begin{aligned} s_1 &= \frac{\rho}{1+\rho}; \\ s_2 &= -\frac{1}{1+\rho}; \\ q(y) &= \frac{1}{Z(\rho)} \left(\sum_{\mathcal{X}} p(x) p(y|x)^{1/(1+\rho)} \right)^{1+\rho}. \end{aligned}$$

The optimizing values of τ and ρ are specified only implicitly. The optimum τ is given by

$$\tau(\rho) = \mu_c'(s_1) = \mu_i'(s_2) = \sum_{\mathcal{Y}} q(y) \frac{\sum_{\mathcal{X}} p(x) p(y|x)^{1/(1+\rho)} h(x,y)}{\sum_{\mathcal{X}} p(x) p(y|x)^{1/(1+\rho)}}.$$

The optimum ρ is the ρ that satisfies

$$E_0'(\rho) = R,$$

provided that $R \ge E'(1) = R_{\text{crit}}$. For $R \le R_{\text{crit}}$, the optimizing ρ is 1, and the optimizing τ is constant at $\tau_{\text{crit}} = \tau(1)$.

As $\rho \to 0$, we have $s_1 \to 0, s_2 \to -1$, and $q(y) \to p(y)$. The optimum τ decreases to $\mu_c'(0) = \mathsf{E}[h(x,y)] = H(Y|X)$, and $E_0'(\rho) \to E_0'(0) = I(X;Y)$. In other words, $\tau_{\text{GV}}(I(X;Y)) = H(Y|X)$, because $E_i(H(Y|X)) = I(X;Y)$.

For $R > R_{\text{crit}}$, we have $\rho < 1$, and therefore $E_i(\tau) = R$, or equivalently $\tau = \tau_{\text{GV}}(R)$. In other words, the error exponent is equal to the exponent of the probability that the correct codeword coupling equals or exceeds $\tau_{\text{GV}}(R)$, in which case there will with high probability be an exponentially large number of incorrect codewords with equal or smaller coupling. On the other hand, for $R < R_{\text{crit}}$, in the typical error scenario there is only one incorrect codeword with less coupling than the correct codeword.

In the high-rate region we therefore have the following parametric form for $E(R)$:

$$\begin{aligned} E(\tau) &= E_c(\tau); \\ R(\tau) &= E_i(\tau), \end{aligned}$$

where the parameters of $E_c(\tau)$ and $E_i(\tau)$ are implicitly optimized as functions of τ. As $\tau \to H(Y|X)$, we have $E(\tau) \to 0$ and $R(\tau) \to I(X;Y)$, whereas as $\tau \to \tau_{\text{crit}}$, we have $E(\tau) \to R_0 - R_{\text{crit}}$ and $R(\tau) \to R_{\text{crit}}$.

Finally, since $R(\tau) = E_i(\tau)$ implies $\tau = \tau_{\text{GV}}(R)$, we can combine these into the "closed-form" expression

$$E(R) = E_c(\tau_{\text{GV}}(R)),$$

where again all parameters are implicitly optimized as functions of R.