ENEE 739C: Advanced Topics in Signal Processing: Coding Theory
Instructor: Alexander Barg

Lecture 6 (**draft;** 9/6/03). Error exponents for Discrete Memoryless Channels

http://www.enee.umd.edu/~abarg/ENEE739C/course.html

Let us take a step back and suppose that information is transmitted over a discrete memoryless channel $\mathscr{W}$ with input and output alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$. The channel is defined by a stochastic matrix W. This means that a letter $x \in \mathcal{X}$ is received as $y \in \mathcal{Y}$ with probability $w_{x,y} = W(y|x)$ given by the corresponding entry of W. The term "stochastic matrix" means that its entries are between 0 and 1 and for every $x \in \mathcal{X}$

$$\sum_{y\in\mathcal{Y}} W(y|x) = 1$$

(each row defines a probability distribution). Given n-strings $\mathbf{x} \in \mathcal{X}^n, \mathbf{y} \in \mathcal{Y}^n$ we can use the absence of memory in the channel to compute

$$W^n(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{n} W(y_i|x_i).$$

We would like to derive an exponential error bound for transmission over $\mathscr{W}$. This will also give a lower bound on the capacity of the channel.

In fact, the basic ideas are already in place from the previous lecture. The bulk of the effort is spent on developing the relevant language. My goal here is to show that the concepts involved in the derivation for the BSC have a rather general nature.

We begin with an observation that in the general case the Hamming distance cannot be a useful measure of likelihood because the channel does not have to be symmetric. Therefore, one has to work with the distribution of all possible sequences of symbols (compositions) from $\mathcal{X}, \mathcal{Y}$. Next linear codes also cannot be of much use in general because the capacity achieving distribution for $\mathscr{W}$ does not have to be uniform (recall your information theory course).

Our nearest goal is to develop a machinery which enables us to work conveniently with compositions, thinking of them as of probability distributions. We will get to meaningful results on p.3.

Types

We still prefer to call n-strings of letters vectors. The *type* of a vector $\mathbf{x} \in \mathcal{X}^n$ is a probability distribution P given by

$$P_{\mathbf{x}}(a) = \frac{1}{n}|\{i : x_i = a\}| \quad (a \in \mathcal{X}).$$

In this case we write $T(\mathbf{x}) = P$. The set of all vectors in $\mathcal{X}^n$ of type P is denoted by T_P. A code $\mathcal{C} \subset \mathsf{T}_P$ is called a *constant composition code*.

The set of all types on $\mathcal{X}^n$ is denoted by $\mathcal{P}(\mathcal{X}^n)$. There are not that many types:

Lemma 1.

$$|\mathcal{P}(\mathcal{X}^n)| = \binom{n+q-1}{q-1} \le n^q \quad (n, q \ge 2).$$

Proof : The first equality follows since $|\mathcal{P}(\mathcal{X}^n)|$ is the number of partitions of $n+q$ into a sum of q positive terms, i.e., the needed binomial. To prove the inequality, proceed by induction. Take arbitrary n and $q = 2$, then the claim is true: $\binom{n+1}{1} = n+1 < n^2$. Now let us fix n and perform induction step on q:

$$\binom{n+(q+1)-1}{q} = \frac{n+q}{q}\binom{n+q-1}{q-1} \le \frac{n+q}{q} n^q \le n^{q+1},$$

where the first of the two inequalities follows by the induction hypothesis and the second is implied by the inequality $\frac{n+q}{q} < n$ which holds true for any $n, q \ge 2$. ∎

Let

$$H(P) = -\sum_{i=1}^{q} p_i \log_2 p_i$$

be the entropy of a probability distribution $P = (p_1, p_2, \ldots, p_q)$. The entropy of a vector $\mathbf{x} \in \mathsf{T}_P$ is defined as $H(\mathbf{x}) = H(P)$. Clearly if $T(\mathbf{x}) = P$ then $P^n(\mathbf{x}) = 2^{-nH(P)}$.

Similarly to the volume of the sphere $\mathcal{S}_w$ we would like to estimate the size $|\mathsf{T}_P|$. The answer is really the same.

Lemma 2.

$$|\mathcal{P}(\mathcal{X}^n)|^{-1} 2^{nH(P)} \le |\mathsf{T}_P| \le 2^{nH(P)}.$$

Proof : The Stirling formula will work in this case as well since

$$|\mathsf{T}_P| = \binom{n}{i_1, i_2, \ldots, i_q}.$$

To spare the reader cumbersome computations, let us take an easier path. First, $|\mathsf{T}_P| P^n(\mathbf{x}) = P^n(\mathsf{T}_P) \le 1$, so $|\mathsf{T}_P| \le \exp[nH(P)]$. It remains to prove that

$$P^n(\mathsf{T}_P) \ge |\mathcal{P}(\mathcal{X}^n)|^{-1}.$$

For that observe that

$$1 = P^n(\mathcal{X}^n) = \sum_{P' \in \mathcal{P}(\mathcal{X}^n)} P^n(\mathsf{T}_{P'}).$$

The needed result will follow if we show that $P^n(\mathsf{T}_P) \ge P^n(\mathsf{T}_{P'})$. We have

$$|\mathsf{T}_{P'}| = \frac{n!}{\prod_{a \in \mathcal{X}} (nP'(a))!}$$

and so

$$\begin{aligned}
\frac{P^n(\mathsf{T}_{P'})}{P^n(\mathsf{T}_P)} &= \frac{|\mathsf{T}_{P'}| \prod_{a \in \mathcal{X}} P(a)^{nP'(a)}}{|\mathsf{T}_P| \prod_{a \in \mathcal{X}} P(a)^{nP(a)}} \\
&= \prod_{a \in \mathcal{X}} \frac{(nP(a))!}{(nP'(a))!} P(a)^{n(P'(a)-P(a))} \\
&\le \prod_{a \in \mathcal{X}} (nP(a))^{n(P(a)-P'(a))} P(a)^{n(P'(a)-P(a))} \quad (\text{since } n!/m! \le n^{n-m}) \\
&= \prod_{a \in \mathcal{X}} n^{n(P(a)-P'(a))} = 1.
\end{aligned}$$

∎

Let us check ourselves: suppose $|\mathcal{X}| = 2, \mathbf{x} \in \mathcal{X}^n$, and $\mathrm{wt}(\mathbf{x}) = w = \omega n$. Let $P = (p, 1-p)$ be another binomial distribution. We have

$$P^n(\mathbf{x}) = p^{\omega n}(1-p)^{(1-\omega)n} = 2^{-n(h_2(\omega) + D(\omega \| p))}.$$

Let $\mathcal{S}_{\omega n}$ be the set of all vectors of weight ωn, then

$$\begin{aligned}
P^n(\mathcal{S}_{\omega n}) &\cong 2^{-nD(\omega \| p)} \\
|\mathsf{T}_{\mathbf{x}}| &\cong 2^{-nh_2(\delta)}.
\end{aligned}$$

These expressions are familiar from the previous lectures.

Types replace the Hamming weight. What about the distance? The *joint type* $T(\mathbf{x}_1, \mathbf{x}_2)$ for a pair of distinct codewords $\mathbf{x}_1, \mathbf{x}_2$ is a probability distribution defined by

$$P_{\mathbf{x}_1, \mathbf{x}_2}(a, b) = \frac{1}{n} |\{i : x_{1,i} = a, x_{2,i} = b\}|$$

for every pair of letters $a, b \in \mathcal{X}$.

Example. Let $q = 3$. We have $T(0012) = (1/2, 1/4, 1/4)$, $T(0012, 1110) = (0, 1/2, 0, 0, 1/4, 0, 1/4, 0, 0)$, where we assumed the lexicographic order on $\mathcal{X}^2$.

Entropies

Given $\mathbf{x}$ its entropy is $H(\mathbf{x}) = H(P)$ where $T(\mathbf{x}) = P$. For two vectors $\mathbf{x}, \mathbf{y}$ their joint entropy $H(\mathbf{x}, \mathbf{y})$ is the entropy of the joint type $T(\mathbf{x}, \mathbf{y})$. Define the *conditional entropy*

$$H(\mathbf{y}|\mathbf{x}) = H(\mathbf{x}, \mathbf{y}) - H(\mathbf{y})$$

(recall that this is the usual way the conditional entropy is expressed for two random variables, e.g., [2, p.16]).

It may be disturbing that the conditional entropy was defined in the absence of a conditional probability distribution (type). To bridge the gap, define a stochastic matrix V whose rows are numbered by $\mathcal{X}$, columns by $\mathcal{Y}$, and the entry v_{xy} equals the frequency of seeing the letter $y \in \mathcal{Y}$ in the coordinates where $\mathbf{x}$ has letter $x \in \mathcal{X}$.

Example. Let $q = 3, \mathcal{X} = \mathcal{Y}$, $\mathbf{x} = (000112), \mathbf{y} = (011120)$. We have

$$V = \begin{pmatrix} 1/3 & 2/3 & 0 \\ 0 & 1/2 & 1/2 \\ 1 & 0 & 0 \end{pmatrix}$$

Further, the joint type

$$T(\mathbf{x}, \mathbf{y}) = \Big(\frac{1}{6}, \frac{1}{3}, 0, 0, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, 0, 0\Big).$$

can be also computed as $T(x, y) = P(x)V(y|x)$ (check this!). □

Generalizing, let V be a stochastic matrix on $\mathcal{X} \times \mathcal{Y}$ $T(\mathbf{x}) = P$. Define the probability distribution

$$PV(y) = \sum_x P(x)V(y|x)$$

and suppose $\mathbf{y}$ has type $T(\mathbf{y}) = PV$. In this situation we shall say that $\mathbf{y}$ has *conditional type* $T_V(\mathbf{x})$.

We can also write the *conditional entropy* $H(\mathbf{y}|\mathbf{x}) = H(V|P)$ as

$$H(V|P) = -\sum_{x \in \mathcal{X}} P(x) \sum_{y \in \mathcal{Y}} V(y|x) \log V(y|x) = \mathsf{E}\Big(-\sum_{y \in \mathcal{Y}} V(y|x) \log V(y|x)\Big)$$

How many vectors $\mathbf{y}$ of conditional type $T_V(\mathbf{x})$ are there? The answer is computed similarly to the above:

$$|T_V(\mathbf{x})| \cong \exp(nH(V|P))$$

Let $T(\mathbf{x}_1) = P, T(\mathbf{x}_2) = PV$. The *mutual information* $I(\mathbf{x}_1, \mathbf{x}_2)$ between $\mathbf{x}_1$ and $\mathbf{x}_2$ (sometimes denoted as $I(x \wedge y)$) is defined as

$$I(\mathbf{x}_1, \mathbf{x}_2) = \sum_{x, x' \in \mathcal{X}} P_{\mathbf{x}_1, \mathbf{x}_2}(x, x') \log_2 \frac{P_{\mathbf{x}_1, \mathbf{x}_2}(x, x')}{P(x)PV(x')}$$

$$= H(PV) - H(V|P)$$

(again this is consistent with the standard definition of mutual information).

GV bound and the entropy distance distribution

We have created enough concepts to do the job in the general case. Next we must struggle with these chimeras proving things that are completely obvious. We begin with the "distance distribution" of random codes.

Theorem 3. *For any $P \in \mathcal{P}(\mathcal{X}^n)$ there exists an $(n, M = 2^{nR})$ code $\mathcal{C} \subset \mathsf{T}_P$ such that for any $Q \in \mathcal{P}(\mathcal{X}^n \times \mathcal{X}^n)$*

$$|\{\mathbf{c}_i, \mathbf{c}_j : i \neq j, T(\mathbf{c}_i, \mathbf{c}_j) = Q\}| \leq 2^{n(R - I(\mathbf{c}_i, \mathbf{c}_j)) + \epsilon} \quad (\epsilon > 0).$$

In particular, for every pair of distinct vectors $\mathbf{c}, \mathbf{c}' \in \mathcal{C}$

$$R \geq I(\mathbf{c}, \mathbf{c}') - \epsilon.$$

The second part of this theorem, which is a direct generalization of the GV bound, is due to R. Blahut [1].

Remark. Before proving this let us verify for some simple example that this theorem is consistent with our idea of the Gilbert bound. Let $q = 2$, then a type P is simply the set of all vectors of some fixed weight $w = \omega n$ (also called the *Johnson space*). So let $\mathbf{x}, \mathbf{x}'$ be two vectors of weight w and suppose that $\mathrm{d}(\mathbf{x}, \mathbf{x}') = 2\delta n$. We have

$$T(\mathbf{x}) = T(\mathbf{x}') = (\omega, 1 - \omega)$$
$$T(\mathbf{x}, \mathbf{x}') = (1 - \omega - \delta, \delta, \delta, \omega - \delta)$$

and

$$H(\mathbf{x}'|\mathbf{x}) = H(\mathbf{x}', \mathbf{x}) - H(\mathbf{x}) = \omega h_2\Big(\frac{\delta}{\omega}\Big) + (1 - \omega) h_2\Big(\frac{\delta}{1 - \omega}\Big)$$

Hence the GV bound on the rate of an $(n, M = 2^{nR}, d)$ code with distance in the Johnson space is given by

$$\begin{aligned} R \geq I(\mathbf{x}, \mathbf{x}') &= H(\mathbf{x}) - H(\mathbf{x}'|\mathbf{x}) \\ &= h_2(\omega) - h_2\Big(\frac{\delta}{\omega}\Big) - (1 - \omega) h_2\Big(\frac{\delta}{1 - \omega}\Big) \end{aligned}$$

Is this indeed the GV bound in the Johnson space? The answer will be found in the homework. However there is one positive sign: let $\omega = 1/2$, then we get the familiar $R \geq 1 - h_2(\delta/2)$. This reflects the fact that the middle layer (vectors of weight $n/2$) is not so much different from all of the Hamming space $\mathscr{H}_2^n$. In particular, the best codes that we know how to construct in either of these two spaces have essentially the same size and distance distribution.

Proof : (of Theorem 3). Choose M codewords of the code C from $\mathcal{X}^n$ randomly and independently. Let $Q \in \mathcal{P}(\mathcal{X}^n \times \mathcal{X}^n)$ be a joint type with both marginal distributions $Q(\cdot, \mathbf{x}')$ and $Q(\mathbf{x}, \cdot)$ equal to P. We have for $i \neq j$

$$\begin{aligned} \Pr[T(\mathbf{c}_i, \mathbf{c}_j) = Q] = \frac{|\mathsf{T}_Q|}{|\mathsf{T}_P|^2} &\leq \frac{\exp[nH(\mathbf{x}, \mathbf{x}')]}{|\mathcal{P}(\mathcal{X}^n)|^{-2} \exp[2nH(P)]} \\ &= |\mathcal{P}(\mathcal{X}^n)|^2 2^{-nI(\mathbf{c}_i, \mathbf{c}_j)}. \end{aligned}$$

Hence for a given i

$$\mathsf{E}|\{j \neq i, T(\mathbf{c}_i, \mathbf{c}_j) = Q\}| \leq M |\mathcal{P}(\mathcal{X}^n)|^2 2^{-nI(\mathbf{c}_i, \mathbf{c}_j)}.$$

Let

$$F_i = \sum_{\substack{Q \in \mathcal{P}(\mathcal{X}^n \times \mathcal{X}^n) \\ Q(\cdot, x') = Q(x, \cdot) = P}} |\{j \neq i, T(\mathbf{c}_i, \mathbf{c}_j) = Q\}| 2^{nI(\mathbf{c}_i, \mathbf{c}_j)}.$$

We have

$$\mathsf{E} \sum_{i=1}^{M} F_i \leq M^2 n^{2q} |\mathcal{P}(\mathcal{X}^n \times \mathcal{X}^n)|.$$

Therefore, there exists a code for which $\sum_i F_i$ satisfies this inequality. In this code at least $\lfloor M/2 \rfloor$ codewords satisfy

$$F_i \leq 2M n^{2q} |\mathcal{P}(\mathcal{X}^n \times \mathcal{X}^n)| \leq 2M n^{q(q+2)}.$$

These codewords form a code that satisfies both claims to be proved. ∎

Exercise. Suppose that $\mathcal{X}$ is an additive group and $\mathcal{C}$ is an additive code (i.e., an additive subgroup of $\mathcal{X}^n$). Let $P_0 = (1, 0, \dots, 0)$ be the type of the all-zero vector. Deduce from the above theorem that there exists a code of size $M \cong 2^{nR}$ such that for any $\epsilon > 0$ and any type $P \neq P_0$

$$|\mathcal{C} \cap \mathsf{T}_P| \le 2^{n(R+H(P)-\log_2 q+\epsilon)}.$$

In particular, for any $\mathbf{x} \in \mathcal{C}\backslash\{0\}$ with $T(\mathbf{x}) = P$ we have

$$R \ge \log_2 q - H(P) - \epsilon.$$

Recall the GV bound and the weight profile of random linear binary codes, compare it with these results.

Decoding and error exponents

Given a channel $\mathscr{W} : \mathcal{X} \to \mathcal{Y}$ we will transmit with a code $\mathcal{C}$ whose existence was proved in Theorem 3. Let $\mathbf{x} \in \mathcal{C}$ be the transmitted vector and $\mathbf{y} \in \mathcal{Y}^n$ be the vector received from the channel. Let us compute the probability that the conditional type of $\mathbf{y}$ is $T_V(\mathbf{x})$:

$$\begin{aligned} W^n[T_V(\mathbf{x})|\mathbf{x}] &= \sum_{\mathbf{y}\in T_V(\mathbf{x})} \prod_{i=1}^n W(y_i|x_i) = |T_V(\mathbf{x})| \prod_{x\in\mathcal{X}} \prod_{y\in\mathcal{Y}} W(y|x)^{V(y|x)n} \\ &= \exp(-nD(V\|W|P)), \end{aligned}$$

where the function

$$D(V\|W|P) = \sum_{x,y} P(x)V(y|x)\log_2 \frac{V(y|x)}{W(y|x)}$$

(the average over x of the K.-L, distance between $V(y|x)$ and $W(y|x)$) is called the *conditional divergence*. Denoting $Q = T(\mathbf{x}, \mathbf{y})$, we can also write $D(V\|W|P) = D(Q\|W)$.

Maximum mutual information decoding. Given a vector $\mathbf{y}$ received from the channel $\mathscr{W}$, decode to the (unique) codevector $\mathbf{x}$ such that

$$I(\mathbf{x}, \mathbf{y}) = \max_{\mathbf{x}'\in\mathcal{C}} I(\mathbf{x}', \mathbf{y}).$$

Theorem 4. *Let $\mathcal{C} \subset (\mathcal{X})^n$ be a code with good entropy distance distribution used over the channel $\mathscr{W}$. Then the exponent of the average error probability of decoding is bounded below as*

$$E(\mathcal{C}, \mathscr{W}) \ge \min_V [D(V\|W|P) + (I(P, V) - R)^+], \tag{1}$$

where V ranges over all stochastic matrices $V : \mathcal{X} \to \mathcal{Y}$ and $(a)^+ = \max(a, 0)$.

We do not give the proof here; see [3].

We will write $E(R, P)$ to denote the right-hand side of (1).

Note a close analogy between the error exponent for the BSC and (1). In particular, the role of V in (1) is analogous to that of the parameter ρ in the argument for the BSC in the sense that both account for the likelihood of codewords different from the transmitted one. Moreover, with some work we can recover parts (b)-(c) of the Theorem of the previous lecture from this result (in a way the general result is easier because explicit optimization is impossible).

It is not so difficult to see that $E(R, P) > 0$ for $R < I(P, W)$ and becomes zero when $R = I(P, W)$. Hence the capacity of the channel $\mathscr{C} \ge \max_P I(P, W)$. We all know of course that this in fact is an exact equality [2, p. 184,198]. The distribution P which furnishes the maximum to $I(P, W)$ is called the *capacity achieving distribution* of the channel $\mathscr{W}$.

A plethora of properties of the function $E(R, P)$ is found in the exercises in [4, Sect. 2.5] and in [6, 7]; of them we will mention only one. The function $E(R, P)$ can be written in an equivalent form (Gallager's famous result [5, 6]) as follows:

$$E(R, P) = \max_{0\le\rho\le 1} \Big\{ -\rho R - \log \sum_y \big(\sum_x P(x) W^{\frac{1}{1+\rho}}(y|x)\big)^{1+\rho} \Big\}$$

The curve $E(R, p)$ usually (not always) looks qualitatively like the error exponent for the BSC from the previous lecture, except that it does not include the expurgation part. The value$\rho = \rho_0$ which maximizes the above expression for a given R is the negative slope of the tangent to $E(R, p)$ at this point. The straight line part corresponds to the value $\rho_0 = 1$. The maximum value of the R such that $\rho_0 = 1$ is the critical rate R_{crit} of the channel. For the sphere packing part of the curve ρ_0 changes from 1 to 0 as the rate increases and becomes 0 at $R = \mathscr{C}$ (the capacity), where $E(R, P)$ also reaches zero.

References

1. R. E. Blahut, *Composition bounds for channel block codes*, IEEE Trans. Inform. Theory **23** (1977), no. 6, 656–674.
2. T. M. Cover and J. A. Thomas, *Elements of information theory*, John Wiley & Sons, New York e.a., 1991.
3. I. Csiszár, *The method of types*, IEEE Trans. Inform. Theory **44** (1998), no. 6, 2505–2523, Information theory: 1948–1998.
4. I. Csiszár and J. Körner, *Information theory. Coding theorems for discrete memoryless channels*, Akadémiai Kiadó, Budapest, 1981.
5. R. G. Gallager, *A simple derivation of the coding theorem and some applications*, IEEE Trans. Inform. Theory **11** (1965), 3–18.
6. ———, *Information theory and reliable communication*, John Wiley & Sons, New York e.a., 1968.
7. A. J. Viterbi and J. K. Omura, *Principles of digital communication and coding*, McGraw-Hill, 1979.