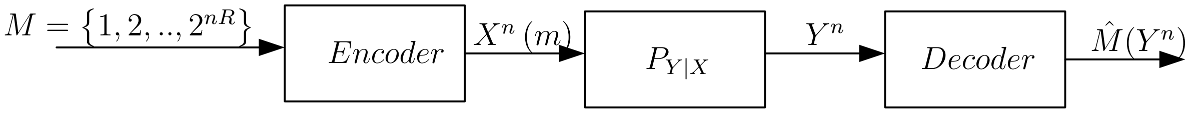
Mathematical methods in communication	June 16th, 2009
Lecture 8	
Lecturer: Haim Permuter	*Scribe: Roni Winik and Itzhak Tamo*

I. Gaussian Channel Coding

We consider the following channel coding problem:



$$P_e = P_r(\hat{M} \neq M)$$

$$P_{max} = \max_i P_r(\hat{M} \neq M | m = i)$$

Fig. 1. Communication system

For finite alphabet we saw that $C = \max_{P_X} I(X;Y)$. We now consider the Gaussian channel as an example of a continuous alphabet, where Z_i is a Gaussian white noise and

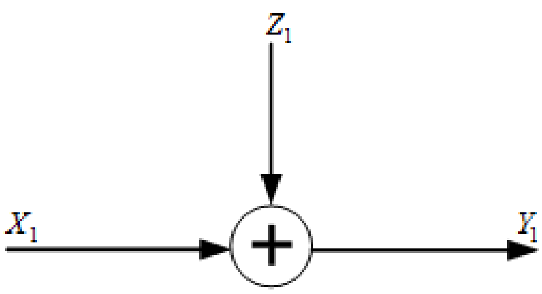


Fig. 2. Gaussian channel where $Y_i = X_i + Z_i$, and X_i is the input of the channel at time i, Y_i is the output of the channel at time i, and Z_i is a Gaussian noise, i.i.d. with variance σ_z^2.

is independent of the input X_i. If there is no constraint on the input, one can transmit an unlimited amount of information in one usage of the channel by using a large input. However, in reality we have a power constraint on the input, i.e., $\frac{1}{n}\mathbb{E}[\sum_{i=1}^{n} X_i^2] \leq P$. Note that Let us define a code for the Gaussian channel:

Definition 1 (A code for the Gaussian channel with a power limit constraint) An $(2^{nR}, n)$ code for the Gaussian channel with power constraint P consists of the following:

1) An index set $\{1, 2, ..., 2^{nR}\}$
2) An encoder function
$$f : \{1, 2, ..., 2^{nR}\} \mapsto \mathcal{X}^n, \tag{1}$$
that satisfies the power constraint for each message, i.e.,
$$\frac{1}{n}\sum_{i=1}^{n} x_i^2(m) \leq P, \quad \forall m \in \{1, 2, ..., 2^{nR}\}. \tag{2}$$
3) A decoding function
$$g : \mathcal{Y}^n \mapsto \{1, 2, ..., 2^{nR}\}, \tag{3}$$

Definition 2 (Achievable rate) The rate R is said to be achievable for a Gaussian channel with a power constraint P, if there exists a sequence of $(2^{nR}, n)$ codes with codewords satisfying the power constraint such that the maximal probability of error, P_{max}, tends to zero.

Definition 3 (Capacity of a channel) The capacity of the channel, denoted as C, is the supremum of all achievable rates.

We will show that for continuous alphabet with power constraint P,
$$C = C^I \tag{4}$$
where
$$C^I = \max_{f_X : \mathbb{E}[X^2] \leq P} I(X; Y). \tag{5}$$

Theorem 1 (Capacity of continues alphabet channel with power constraint) If $P_{Y|X}$ is a continuous alphabet channel with power constraint P , then $C = C^I$.

The proof follows the similar steps as the case of finite alphabet. The converse uses Fano's inequality, and the achievability uses a random codes argument, joint typicality decoder, and the following lemma

Lemma 1 (Upper bound on the probability of two independent sequences to be jointly typical) Consider a joint pdf $f_{X,Y}$. Let X^n be in the typical set, i.e., $X^n \in A_\epsilon(X)$, where

$$A_\epsilon^n = \left\{ X^n : \left| -\frac{1}{n} \log_2 \prod_{i=1}^{n} f_X(x_i) - h(x) \right| \leq \epsilon \right\}. \tag{6}$$

Let Y^n be drawn i.i.d. according to the pdf f_Y. Then

$$\Pr\{(X^n, Y^n) \in A_\epsilon^{(n)}(X,Y)\} \leq 2^{-n(I(X;Y)+\epsilon)}. \tag{7}$$

II. Proof for theorem 1

Proof: **Achivability**: In the achievability proof, we show that if $R < C^I$, then R is an achievable rate. Let $R < C^I$, we will see that there exists a sequence of $(2^{nR}, n)$ codes, such that maximal probability of error P_{max} tends to zero . Meaning that R is an achievable rate.

Encoder: Let f_X be a pdf that satisfies the power constraint $E(X^2) \leq P$. We generate the codebook with each element chosen i.i.d $\sim f_X$. Thus, $X_i(m), i = 1, 2, ..., n, m = 1, 2, ..., 2^{nR}$. Forming codewords $X^n(1), X^n(2), ..., X^n(2^{nR})$.

Decoder: The decoder looks at the list of the codewords and searches for one that is jointly typical with the received vector Y^n. If there is only one such codeword the decoder declares it to be the transmitted codeword. Otherwise the decoder declares an error. The decoder also declares an error if chosen codeword does not satisfy the power constraint.

Probability of error: without loss of generality, assume that codeword $m = 1$ was sent. Thus, $Y^n = X^n(1) + Z_n$. The decoder declares an error if one of the following events occur:

1) The power constraint is violated

$$E_0 = \left\{ \frac{1}{n} \sum_{i=1}^{n} X_i^2(1) > P \right\} \tag{8}$$

2) The received codeword ,Y^n, and the transmitted codeword ,$X^n(1)$, are not jointly typical.

$$E_1^c = \{(X^n(1), Y^n) \notin A_\epsilon^{(n)}\} \tag{9}$$

3) There exits $X^n(i), i \neq 1$, such that $(X^n(i), Y^n)$ are jointly typical

$$E_i = \{(X^n(i), Y^n) \in A_\epsilon^{(n)}\}, i = 2, 3, ..., 2^{nR} \tag{10}$$

We can see that:

1) By the law of large numbers we know that $\frac{1}{n}\sum_{i=1}^n X_i^2(1) \to P$, therefore $P(E_0) \to 0$ as $n \to \infty$.
2) By the joint AEP Theorem of the last lecture we know that $P((X^n(1), Y^n) \in A_\epsilon^{(n)}) \to 1$ as $n \to \infty$ therefore $P(E_1^c) \to 0$.
3) For $i \neq 1$, $X^n(i)$ and Y^n are independent, therefore again by Lemma 1 $P((X^n(i), Y^n) \in A_\epsilon^{(n)}) \leq 2^{-n(I(X;Y)-3\epsilon)}$.

Then

$$\begin{aligned}
P_e^{(n)} &= P(\hat{M} \neq M | M = 1) && (11)\\
&= P(E_0 \cup E_1^c \cup_{i=2}^{2^{nR}} E_i) && (12)\\
&\overset{(a)}{\leq} P(E_0) + P(E_1^c) + \sum_{i=2}^{2^{nR}} P(E_i) && (13)\\
&\leq \epsilon + \epsilon + \sum_{i=2}^{2^{nR}} 2^{-n(I(X;Y)-3\epsilon)} && (14)\\
&\overset{(b)}{\leq} 2\epsilon + 2^{3n\epsilon} 2^{-n(I(X;Y)-R)} \leq 3\epsilon && (15)
\end{aligned}$$

Where:
(a) Only one error can occur at the same time.
(b) For n sufficiently large $R < I(X;Y) - 3\epsilon$.
Thus, $P_e^{(n)}$ tends to zero. As in the discrete case, deleting the worst half of the codewords (those with most probable error) results in arbitrarily low maximal error probability, P_{max}. In particular the power constraint is satisfied by each of the remaining codewords since

these codewords have probability of error 1, and must belong to the worst half of the codewords. Hence R is an achievable rate therefore $C^I \leq C$.

Converse: Now we will show that if R is an achievable rate then $R < C^I$. In the discrete case which holds also for the continuous one, we obtained , $nR \leq \sum_{i=1}^{n} I(X_i; Y_i) + n\epsilon_n$. Where $\epsilon_n = \frac{1}{n} + RP_e^{(n)}$.

Let P_i be the average power of the ith coordinate of the codebook, i.e. $P_i = \frac{1}{2^{nR}} \sum_{m=1}^{2^{nR}} X_i^2(m)$. Since each of the codeword satisfies the power constraint so does their average, and hence

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} P_i &= \frac{1}{n}\sum_{i=1}^{n}\int_x f_{X_i}(x)x^2 dx && (16) \\ &= \int_x \underbrace{\frac{1}{n}\sum_{i=1}^{n} f_{x_i}(x)}_{\bar{f}_{\bar{X}}} x^2 dx \leq P && (17) \end{aligned}$$

Therefore $\bar{f}_X$ is a pdf which satisfies the power constraint. Lets define the notation $\mathcal{I}(f_X; f_{Y|X}) \triangleq I(X;Y)$. Thus,

$$\begin{aligned} R &\overset{(a)}{\leq} \frac{1}{n}\sum_{i=1}^{n} I(X_i; Y_i) + \epsilon_n && (18) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathcal{I}(f_{X_i}; f_{Y_i|X_i}) + \epsilon_n && (19) \\ &\overset{(b)}{\leq} \mathcal{I}(\frac{1}{n}\sum_{i=1}^{n} f_{X_i}; f_{Y|\bar{X}}) + \epsilon_n && (20) \\ &= \mathcal{I}(\bar{f}_{\bar{X}}; f_{Y|\bar{X}}) + \epsilon_n && (21) \end{aligned}$$

where:

(a) Using Fano's inequality, similarly, to the discrete case.

(b) Using Jensen inequality and the fact that $\mathcal{I}(f_X; f_{Y|X})$ is concave in f_X.

Since, $\epsilon_n \to 0$, therefore, $R \leq \mathcal{I}(\bar{f}_X; f_{Y|X})$. Since C is the supremum of all achievable rates, we get $C \leq C^I$. And that completes the proof. ■

III. Parallel Gaussian channels and convex optimization

A. Parallel Gaussian channels

We saw in the last lecture that for a Gaussian channel with power constraint,$\mathbb{E}[X^2] \leq P$ and $Z \sim \mathcal{N}(0, N)$ we get that

$$C = \max_{f(x), E(x^2) \leq P} I(X; Y) = \frac{1}{2} \log(1 + \frac{P}{N}). \tag{22}$$

In this section we consider k independent Gaussian channels in parallel with common power constraint. where $Z_1, Z_2, ..., Z_k$ Gaussian white noise and are independent. This is

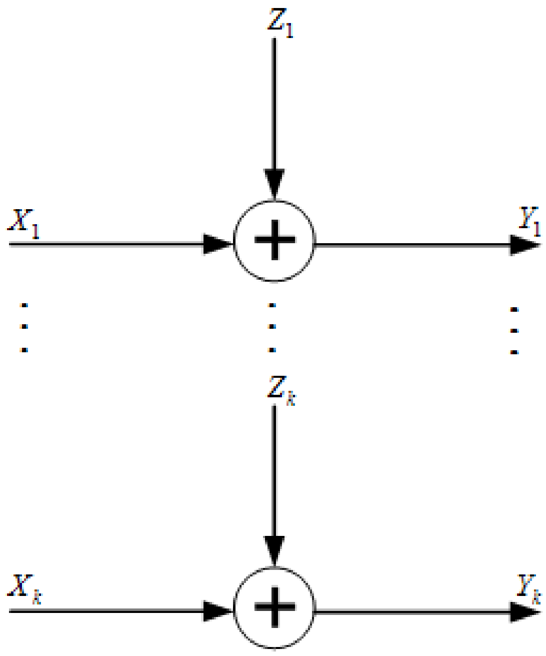


Fig. 3. Parallel Gaussian channels.

a very usefull model where each sub-channel might model a different frequency, different state on a different time slot, or different link that is available.

Our goal is to maximize the capacity of the channel subject to the power constraint. Since the channel is memoryless with the input $X_1, X_2, ...X_k$ and the output $Y_1, Y_2, ...Y_k$ the capacity of the channel is

$$C = \max_{f(x_1, x_2, ..., x_k), \mathbb{E}[\sum_{i=1}^k x^2] \leq P} I(X_1, X_2, ..., X_k; Y_1, Y_2, ..., Y_k) \tag{23}$$

Furthermore, since $Z_1, Z_2, ..., Z_k$ are independent, we get that:

$$\begin{aligned}
I(X_1, X_2, ..., X_k; Y_1, Y_2, ..., Y_k) &= h(Y_1, Y_2, ..., Y_k) - h(Y_1, Y_2, ..., Y_k | X_1, X_2, ..., X_k) \\
&\overset{(a)}{=} h(Y_1, Y_2, ..., Y_k) - h(Z_1, Z_2, ..., Z_k | X_1, X_2, ..., X_k) \\
&\overset{(b)}{=} h(Y_1, Y_2, ..., Y_k) - h(Z_1, Z_2, ..., Z_k) \\
&\overset{(c)}{=} h(Y_1, Y_2, ..., Y_k) - \sum_{i=1}^{k} h(Z_i) \\
&\leq \sum_{i=1}^{k} h(Y_i) - \sum_{i=1}^{k} h(Z_i) \\
&\overset{(d)}{\leq} \sum_{i=1}^{k} \frac{1}{2} \log(1 + \frac{P_i}{N_i}). \qquad (24)
\end{aligned}$$

Where:

(a) follows from the definition of the additive channel

(b) follows form the fact that noise Z is independent of the inputs X_i.

(c) follows from the fact that Z_i is independent of Z_j for any $i \neq j$.

(d) follows from the upper bound on differential entropy for a given variance P_i.

Thus $C \leq \sum_i \frac{1}{2} \log(1 + \frac{P_i}{N_i})$ and equality is achieved[1] if

$$(X_1, X_2, ..., X_k) \sim \mathcal{N}\left(0, \begin{bmatrix} P_1 & 0 & \cdots & 0 \\ 0 & P_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & P_k \end{bmatrix}\right). \qquad (25)$$

In other words, to maximize C the input $(X_1, X_2, ..., X_k)$ should be independent, and for each $1 \leq i \leq k$ $X_i \sim \mathcal{N}(0, P_i)$. Now our problem is reduced to finding the power distribution between the channels, i.e., $P_1, P_2, ..., P_k$. This optimization problem can be written as follows:

$$\begin{aligned}
\min_{x} \quad & -\sum_{i=1}^{k} \frac{1}{2} \log(1 + \frac{P_i}{N_i}) \\
s.t. \quad & -P_i(x) \leq 0, \quad 1 \leq i \leq m
\end{aligned}$$

[1]Please verify as an exercise that indeed all inequalities in (24) become equalities under this input definition.

$$\sum_{i=1}^{k} = P, \quad 1 \leq j \leq l \tag{26}$$

We will now use a special tool of convex optimization problems called the Karush Kuhn Tucker conditions (KKT) to solve it.

B. Convex Optimization

In this subsection we use a convex optimization tool called KKT condition to solve a specific optimization problem. We first give some background on this tool.

Consider the problem of minimizing $f_0(x)$ subject to the constraints that $f_i(x) \leq 0, 1 \leq i \leq m$ (m inequality constraints) and $h_j(x) = 0, 1 \leq j \leq l$ (l equality constraints).

In a formal mathematical notation, the problem may be written as

$$\begin{aligned} \min_{x} \quad & f_0(x) \\ s.t. \quad & f_i(x) \leq 0, \quad 1 \leq i \leq m \\ & h_j(x) = 0, \quad 1 \leq j \leq l \end{aligned} \tag{27}$$

We now introduce the formal definition of a *convex optimization problem.*

Definition 4 (Convex optimization problem) If $f_0(x),\ f_i(x), 1 \leq i \leq m$ are convex functions and $h_j(x) = 0, \quad 1 \leq j \leq l$ are affine functions, i.e., $h_j(x) = A_j x + b$, then the problem is called a convex optimization problem.

We now introduce a dual problem and give sufficient and necessary conditions for a solution to be optimal for a convex optimization problem.

Definition 5 (Dual Function) The *Dual Function*, $L(X, \lambda, \nu) : \Re^n \times \Re^m \times \Re^l \rightarrow \Re$

$$L(X, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{l} \nu_i h_i(x). \tag{28}$$

Where

λ_i- Lagrange multiplier (for the inequalities)

ν_i- Lagrange multiplier (for the equalities).

The necessary conditions for a vector $\tilde{x}$ to be a local minimum are the KKT conditions which is given below.

Definition 6 (KKT conditions) There exist constants $\lambda_i, 1 \leq i \leq m$ and $\nu_j, 1 \leq j \leq l$ such that

1) $\nabla_x L(\tilde{x}, \lambda, \nu) = 0$
2) $f_i(\tilde{x}) \leq 0,\ 1 \leq i \leq m$
3) $h_j(\tilde{x}) = 0,\ 1 \leq j \leq l$
4) $\lambda_i f_i(\tilde{x}) = 0,\ 1 \leq i \leq m$
5) $\lambda_i \geq 0,$

where $\nabla_x L(\tilde{x}, \lambda, \nu)$ is the derivative with respect to x and if x is vector that its the partial derivative with respect to each element in x.

In general KKT condition are necessary condition for having a local minimum. For a convex optimization problem they are necessary and sufficient for having a global minimum.

Theorem 2 (Sufficient and necessary condition for a convex optimization problem) In the special case (which is also our case) where $f_0(x)$ and $f_i(x)$ are convex functions and $h_j(x)$ are affine functions i.e. $h_j(x) = A_j x + b$, namely, the optimization problem is a convex optimization problem, the KKT conditions are sufficient and the vector $\tilde{x}$ is the global minimum.

C. Waterfilling

Now get back our optimization problem. We would like to find the global minimum of the function

$$f_0(P_1, ..., P_k) = -\sum_i \frac{1}{2} \log(1 + \frac{P_i}{N_i}). \tag{29}$$

Such that

$$f_i(P_1, ..., P_k) = -P_i \leq 0. \tag{30}$$

And

$$h(P_1, ..., P_k) = \sum_i P_i - P = 0 \quad (31)$$

writing the functional

$$L(P_1, ..., P_k) = -\sum \frac{1}{2}\log(1 + \frac{P_i}{N_i}) - \sum_i \lambda_i P_i + \nu(\sum_j P_j - P). \quad (32)$$

Differentiating with respect to P_i yields

$$\frac{\partial}{\partial P_i} L = -\frac{N_i}{N_i + P_i}\frac{1}{N_i} - \lambda_i + \nu = 0. \quad (33)$$

Thus

$$P_i + N_i = \frac{1}{\nu - \lambda_i}. \quad (34)$$

If $P_i > 0$ then by 5th KKT condition we get that $\lambda_i = 0$ and thus $P_i = \frac{1}{\nu} - N_i$. However by the 2nd KKT condition, $P_i \geq 0$,therefore $P_i = [\frac{1}{\nu} - N_i]^+ = \max(0, \frac{1}{\nu} - N_i)$.
Note that if $P_i = 0$ then $\lambda_i \neq 0$.
The solution is illustrated graphically in following figure:

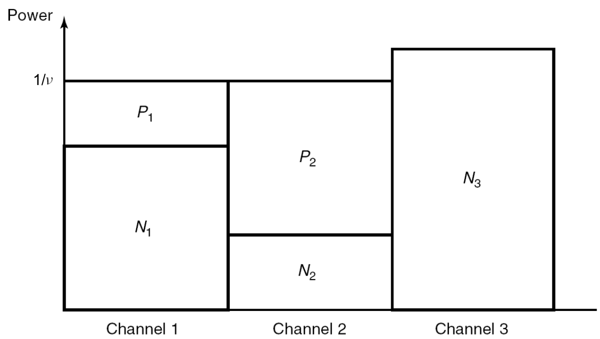


Fig. 4. Water-filling for parallel.

The vertical levels indicate the noise levels in the various channels. As signal power is increased from zero, we distribute the power to the channels with the lowest noise. When the available power is increased still further, some of the power is put into noisier channels. The process by which the power is distributed among the various bins is identical to the way in which water distributes itself in a vessel. Hence this process is referred to as 'water filling'.