6.441 Transmission of Information Apr 20, 2006

Lecture 17

Lecturer: Madhu Sudan *Scribe: Matt Willis and Matt Willsey*

Gaussian Channels (continued)

1 Overview

In this lecture, we will continue our discussion of the All White Gaussian Noise Channel (AWGN). In particular, we review the Coding Theorem for the AWGN as discussed last lecture, and prove the converse to the Coding Theorem for the AWGN. Towards the end of our discussion, we will address Parallel Gaussian Channels, which is the case when there are n communication channels, each with independent (and possibly different) noise characteristics. We briefly comment on the generalization of this analysis to the Colored Gaussian Noise Model, where the noise properties of different channels may be linked, and are no longer independent.

2 Review From Previous Lecture

A Gaussian channel, with an input alphabet, $X \in \mathbb{R}^n$, output alphabet, $Y \in \mathbb{R}^n$, and subject to the power constraint is defined as follows:

$$\begin{aligned} Y &= X + Z \\ Z &\sim N(0, \sigma^2). \end{aligned}$$

As discussed in the previous lecture,we impose the following power constraints to maintain a finite capacity:

$$\begin{aligned} var(X) &\leq P \\ E[X] &= 0. \end{aligned}$$

Also from last lecture, we calculated the channel capacity, C, for a given input distribution $p(x)$ to be:

$$\begin{aligned} C &= \max_{p(x)}\{I(X;Y)\} \\ &= \frac{1}{2}\log(2\pi e(P+\sigma^2)) \text{ } \textit{bits per transmission} \end{aligned}$$

The mutual information is maximized when $X \sim N(0, P)$.

3 Coding Theorem

In this section we will prove both the coding theorem and the converse coding theorem.

3.1 Proof of the Coding Theorem

We will begin by defining an encoding function, E, that has messages in a set of size 2^{Rn} and maps it to n real numbers (since we used the channel n times) as shown below:

$$E\{1, 2, ..., 2^{nR}\} \rightarrow \mathbb{R}^n$$

We pick E such that it is chosen at random. In other words we will ensure that every symbol that we transmit achieves the following distribution:

$$(E(m))_i \quad \sim \quad N(0, P - \epsilon)$$

where m is the message and $(E(m))_i$ is i.i.d. over (m,i). Note that the variance of $(E(m))_i$ is $P - \epsilon$ so that we do not exceed the total power of nP in n transmissions. Next we will establish the following notation:

$\underline{\text{X}}$ denotes the transmitted sequence $\underline{X} = E(m)$
$\underline{\text{Y}}$ denotes the received sequence $\underline{Y} = \underline{X} + \underline{Z}$

We will now try to prove that if R is less than C, then the probability of error is very small.

Goal: if $R < I(x; y)$ then Pr(error) is small

Now there are three sources of decoding error when transmitting m. Generally speaking, the sources of error can depend on the encoding of m, the encoding of some other message m' (where $m' \neq m$), or the error introduced by the channel which is a random variable. The three sources of error correspond to the events below.

First let E_0 be when the power of a realized encoding, E(m), is too large:

$$||E(m)||_2^2 \geq nP.$$

Note that that the 2 in the subscript above indicates an $l^2 - norm$. Remembering that $Exp[(E(m))_i^2] = (P - \epsilon)$, the law of large numbers tells us that the likelihood that we exceed nP in the above equation is very small. Thus,

$$Pr[E_0] \rightarrow 0.$$

It is important to remember that this error is simply a violation of the power constraint.

Now let E_1 be when the noise causes Z to be too large.

$$||Z||_2^2 \geq n(\sigma^2 + \epsilon)$$

Once again the law of large numbers tells us that

$$Pr[E_1] \rightarrow 0.$$

since Z will converge to its mean.

Thus, the previous two errors, E_0 and E_1, simply discusses the likelihood that random variables differ significantly from their expectation.

Now let $E_2(m')$ be defined as the event when probability of the encoding of some message, m', is too close to the encoding of the true message, m. In other words

$$||\underline{Y} - E(m')||_2^2 \leq n(\sigma^2 + \epsilon)$$

We claim that the probability of this event is

$$Pr[E_2(m')] \leq 2^{-I(X;Y)n},$$

which will be proved to follow from joint AEP.

Proof:

Let us consider two random variables, $(\underline{X}, \underline{Y})$, picked jointly according to our channel model (where $\underline{X} = E(m)$). Now let us also consider two additional random variables, $(\widetilde{\underline{X}}, \widetilde{\underline{Y}})$, picked jointly according to the channel model (where $\widetilde{\underline{X}} = E(m')$) and independent of the first two random variables. We now consider the following two subclaims:

1. $Pr[\widetilde{\underline{X}}, \underline{Y}$ *are jointly typical*$] \leq 2^{-I(X;Y)n}$
2. $E_2(m')$ *occurs if* $\widetilde{\underline{X}} = E(m')$ *and* $\underline{Y} = Channel(E[m])$ *are jointly typical.*

Generally speaking, subclaim 2 states that E_2 occurs when the encoding of E(m') is too close to E(m). Subclaim 1 then gives us $Pr[E_2]$.

Now let us draw a picture of this situation, which can be seen in Figure 1. For large n, X will be located at a radius of $\sqrt{nP}$. A particular Y associated with a particular X will be located within a ball of radius $\sqrt{n\sigma^2}$ of X as shown in the figure. However, for large n, most of the volume for realizable values of Y will be located around a radius of

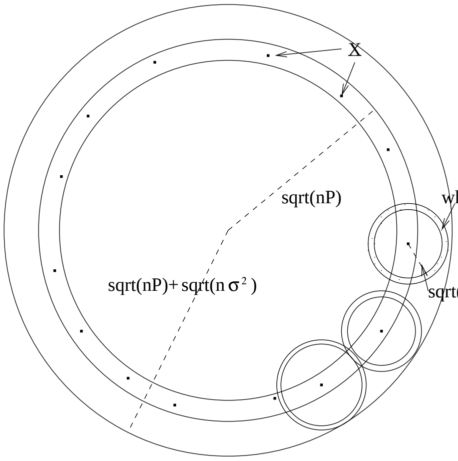

Figure 1: *Graphical Illustration Demonstrating* E_2

$\sqrt{nP + n\sigma^2}$, which is the outermost ring in the figure. E_2 occurs when $\underline{\widetilde{X}}$ falls within a small ball centered around $\underline{\mathrm{X}}$ (which is the event that $\underline{\widetilde{X}}$ and $\underline{\mathrm{Y}}$ are jointly typical).

Thus, if we pick an $\underline{\widetilde{X}}$ independent of $\underline{\mathrm{X}}$, the probability of that $\underline{\widetilde{X}}$ and $\underline{\mathrm{Y}}$ are jointly typical is roughly the volume of a small ball over the total volume of the biggest ball.

$$\begin{aligned} Pr[E_2(m')] &= \sqrt{n\sigma^2}^n / \sqrt{n(P+\sigma^2)}^n \\ &= [\sigma^2/(P+\sigma^2)]^{n/2} \\ &= 2^{-I(X;Y)n} \end{aligned}$$

The last equality holds since we have already shown that $I(X;Y) = log(1 + P/\sigma^2)^{1/2}$.

The above derivation is somewhat ad-hoc; however, now we shall formally prove subclaim 1. To determine the probability that $\underline{\widetilde{X}}$ and $\underline{\mathrm{Y}}$ are typical, we integrate the joint probability over the jointly typical set. Since $\underline{\widetilde{X}}$ and $\underline{\mathrm{Y}}$ are independent, we are able to write the joint probability as the product of the marginal probabilities of $\underline{\widetilde{X}}$ and $\underline{\mathrm{Y}}$.

$$\begin{aligned} Pr[(\underline{\widetilde{X}}, \underline{Y})\ joint\ typ.] &= \int_{joint\ typ.\ set} P_{\underline{\widetilde{X}}}(\underline{\widetilde{X}}) P_{\underline{Y}}(\underline{Y}) d_{\underline{\widetilde{X}}} d_{\underline{Y}} \\ &\leq Vol(joint\ typ.\ set) max_{\underline{\widetilde{X}}}[P_{\underline{\widetilde{X}}}(\underline{\widetilde{X}})] max_{\underline{Y}}[P_{\underline{Y}}(\underline{Y})] \\ &\leq 2^{h(x;y)n} \cdot 2^{-h(x)n} \cdot 2^{-h(y)n} \\ &= 2^{-I(X;Y)n} \end{aligned}$$

Note the second inequality hold since $\underline{\widetilde{X}}$ and $\underline{\mathrm{Y}}$ are both contained in the jointly typical set ($Vol(joint\ typ.\ set) \approx 2^{h(x;y)n},\ max_{\underline{\widetilde{X}}}[P_{\underline{\widetilde{X}}}(\underline{\widetilde{X}})] \approx 2^{-h(x)n},\ and\ max_{\underline{Y}}[P_{\underline{Y}}(\underline{Y})] \approx 2^{-h(y)n}$). The above derivation formally proves subclaim 1, which is:

$$Pr[E_2(m')] = 2^{-I(X;Y)n}.$$

Since there are 2^{Rn} messages,

$$Pr[\exists\ m'\ s.t.\ E_2(m')\ occurs] = 2^{Rn} \cdot 2^{-I(X;Y)n}.$$

Therefore if $R < I(X;Y)$, then

$$Pr[\exists\ m'\ s.t.\ E_2(m')\ occurs] \to 0,$$

which proves the coding theorem.

3.2 Converse to the Coding Theorem

The goal of this section is to demonstrate that the probability of error approaching zero implies that the channel rate R is below capacity, i.e.:

$$p_{err} \to 0 \Longrightarrow R \leq C.$$

By assumption, for a given rate R we have an input alphabet containing messages M where

$$M \in \ \{1, 2, ..., 2^{Rn}\}$$

as well as an encoding function E:

$$E : M \to X^n.$$

Our channel is described mathematically as

$$Y^n = X^n + Z^n$$

We begin the proof by noting that M, X^n, and Y^n form a Markov chain ($M \to X^n \to Y^n$), which allows us to apply Fano's Inequality:

$$H(M|Y^n) \leq 1 + nRp_{err} = O(n),$$

where $O(n) \to 0$ as $p_{err} \to 0$ (this can also be seen by using the full-fledged Fano's Inequality, $H(p_{err}) + p_{err}\log(|X^n| - 1) \geq H(M|Y^n)$). Now consider the quantity

$$I(M; Y^n) = H(M) - H(M|Y^n) = nR - O(n)$$

where $H(M) = nR$ for a uniform input distribution of messages. Due to the Markov Chain ($M \to X^n \to Y^n$), the Data Processing Inequality yields:

$$I(X^n; Y^n) \geq I(M; Y^n) = nR - o(n)$$

We make use of the fact that $I(X^n; Y^n) \leq \sum_i I(X_i; Y_i)$ which can be seen from the following steps (see Cover and Thomas for details):

$$\begin{aligned} I(X^n; Y^n) &= h(Y^n) - h(Y^n|X^n) \\ &= h(Y^n) - h(Z^n) \\ &\leq \sum_i^n h(Y_i) - h(Z^n) \\ &= \sum_i^n h(Y_i) - \sum_i^n h(Z_i) \\ &= \sum_i^n I(X_i; Y_i). \end{aligned}$$

This substitution yields:

$$\sum_i^n I(X_i; Y_i) \geq nR - O(n)$$

Let us say that the i^{th} transmission contains power P_i, where by the power constraint, $\sum_i P_i \leq nP$. As we demonstrated last lecture, we can maximize each $I(X_i; Y_i)$ to be $\frac{1}{2}\log(1 + \frac{P_i}{\sigma^2})$ by choosing the normal distribution as input distribution. Therefore,

$$\sum_i^n I(X_i; Y_i) \geq nR - O(n)$$

$$\sum_i^n \frac{1}{2}\log(1 + \frac{P_i}{\sigma^2}) \geq nR - O(n).$$

Due to symmetry, the left hand side of the equation is maximized when each P_i of equal value. Therefore,

$$\begin{aligned} \sum_i^n \frac{1}{2}\log(1+\frac{P}{\sigma^2}) &\geq nR - O(n) \\ nC &\geq nR - O(n) \\ C &\geq R - O(n)/n \end{aligned}$$

where as explained above $O(n) \rightarrow 0$ as $p_{err} \rightarrow 0$, and this proves the converse.

4 Parallel Gaussian Channels

In previous sections, we discussed the use of one AWGN channel with noise characterized by $Z = N(0, \sigma^2)$. Now we consider the case of Parallel Gaussian Channels, where the user has n such channels at his disposal. Each channel is allowed to have its own noise characteristic ($Z_i = N(0, \sigma_i^2)$), which is independent from other channels. We still impose a power constraint, but now it states that the power used over all n channels must be limited. This is a fairly realistic model that might be used to describe a radio broadcasting station, where each channel represents a different broadcast frequency, and each frequency experiences a different atmospheric dispersion. In fact, we probably already have some intuition concerning how to use a parallel channel. By way of building up an intuition on how to use such a n channel system, consider the following two examples:

- **Example 1**
  In this example, we have n identical channels, with $\sigma_1^2 = \sigma_2^2 = ... = \sigma_n^2$. It is obvious that we would want to distribute the power equally to each channel, so that $P_i = \frac{P}{n}$.
- **Example 1**
  In this example, $\sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2 = 1$, and $\sigma_{k+1}^2 = ... = \sigma_n^2 = \infty$. There is no reason to put any energy into an infinitely noisy channel, as we have no way of interpolating the input given the noisy output. Instead, we distribute the power evenly among the first k channels.

The intuition behind these two examples suggests that the effective way to use a Gaussian parallel channel is to weight the power distribution more heavily among the channels with better noise characteristics.

Now, for a more formal discussion: the i^{th} channel (where $i \in \{1, 2, ..., n\}$) of a Parallel Gaussian Channel is characterized as follows:

$$\begin{gathered} Y_i = X_i + Z_i \\ Z = N(0, \sigma_i^2) \end{gathered}$$

For each channel with power P_i, the power constraint is:

$$\sum_i^n P_i \leq P$$

Or, in terms of channel values:

$$\text{Exp}[\sum_i^n X_i^2] \leq P$$

The quantity of interest, capacity C, is defined as:

$$C = \max_{p(x_1,...,x_n:\sum_i P_i \leq P} I(X_1, ..., X_n; Y_1, ..., Y_n)$$

Following the analysis of last section, we recognize that each channel is maximized (achieves channel capacity) with an input Gaussian distribution subject to the particular channel's power constraint, and therefore:

$$C \leq \sum_i^n \frac{1}{2} \log(1 + \frac{P_i}{\sigma_i^2})$$

The analysis leading up to this equation follows the same reasoning as the one channel case, but now we no longer fix σ, but instead we allow each σ_i to vary independently of the others. The task is to maximize the right side of the equation above subject to the power constraint $\sum_i^n P_i \leq P$,, or equivalently, to maximize the following expression:

$$C \leq \sum_i^n \frac{1}{2} \log(\frac{Q_i}{\sigma_i^2})$$
$$Q_i = P_i + \sigma_i^2,$$

subject to the constraint, $\sum_i^n Q_i \leq P + \sum_i^n \sigma_i^2$. In either case, this is an optimization problem subject to a power constraint; as pointed out in the text, it can be solved using the technique of Lagrange multipliers. First, we form the appropriate Lagrange multiplier expression:

$$J(P_1, P_2, ..., P_n) = \sum \frac{1}{2} \log(1 + \frac{P_i}{\sigma_i^2}) + \lambda(\sum_i P_i)$$

Then differentiate with respect to P_i:

$$\frac{1}{2}\frac{1}{P_i + \sigma_i^2} + \lambda = 0$$
$$\Rightarrow P_i = \nu - \sigma_i^2$$

Where ν can be solved for by substituting the solved P_i's into the power constraint. (It should be noted that for physical reasons, $P_i \geq 0$, and therefore you must bound each P_i below by zero).

The preceding discussion explains the mathematics behind it, but there is a more intuitive approach to understanding the optimization process, through the process of "water-filling." In this analogy, there is a finite amount of "water" (i.e., a limitation on the power constraint) that can be poured into these n channels. It is desirable to put more water into channels that are useful and have low noise characteristics, and less water into noisier channels that have a lower capacity. So, how do we go about distributing the water? Refer to the Figure 10.4 in Cover and Thomas; the water will seek its own level, and naturally pool more deeply into the lower noise channels. A nice feature of the "water-filling" analogy is that it automatically takes into account the fact that $P_i \geq 0$; that is, if a channel is too noisy, it doesn't get negative power, rather, it simply gets no power at all.

5 General Colored Gaussian Channel

This section was only briefly covered in the last six minutes of lecture, but a brief summary is given.

A general colored gaussian channel can be characterized by three parameters. These parameters are the number of parallel channels, k, the total power constraint, P, and a $k \times k$ covariance matrix, K_z. If the K_z is diagonal, then we are dealing with the case explained in section 4 where the noise on each channel is independent from the noise on every other channel. However, in general K_z is not diagonal. Thus, we would then like to understand the capacitance of the total channel and how to relate it to the case shown in section 4. To do this we use linear algebra to diagonalize K_z as shown below.

$$K_z = Q \cdot \Lambda \cdot Q^T$$

In the above equation, $QQ^T = I$, and the diagonal matrix, Λ, is given as

$$\begin{aligned} \Lambda_{i,j} &= \lambda_{i,i}; \quad if\ i = j \\ & \quad 0; \quad\quad else. \end{aligned}$$

Thus we can convince ourselves that the capacity is as the capacity in the parallel gaussian channel with independent channel noise, total power, P, and $\sigma_i^2 = \lambda_i$. It is important to note that any covariance matrix can be diagonalized. Thus we can extend the colored gaussian channel to the case explained in detail in section 4.