6.441 Transmission of Information April 17, 2006

Lecture 16

Lecturer: Madhu Sudan *Scribe: Imad Jabbour*

1 Overview

In this lecture, we discuss the information-theoretic aspect of an Additive White Gaussian Noise (AWGN) channel. This channel is often used in communication theory to model many practical channels. We derive the capacity, and give an overview of the *Channel Coding Theorem* for AWGN channels. But first, we highlight some key facts from the previous lecture.

2 Review From Previous Lecture and Applications

In the previous lecture, we defined *differential entropy* $h(\cdot)$ and outlined some of its properties. We also defined *mutual information* for continuous random variables. In this section, we give a quick overview of the aforementioned material and illustrate some concepts with examples.

Definition 1 (Differential entropy of a continuous random variable) *The differential entropy* $h(X)$ *of a continuous random variable* X *with pdf* $f_X(x)$ *and support set* $\mathbb{R}$ *is defined as*

$$h(X) = -\int_{\mathbb{R}} f_X(x) \log f_X(x) dx \tag{1}$$

Definition 2 (Differential entropy of a continuous random vector) *The differential entropy* $h(\mathbf{X})$ *of a continuous random vector* $\mathbf{X}$ *with pdf* $f_{\mathbf{X}}(\mathbf{x})$ *and support set* $\mathbb{R}^n$ *is defined as*

$$h(\mathbf{X}) = -\int_{\mathbb{R}^n} f_{\mathbf{X}}(\mathbf{x}) \log f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \tag{2}$$

Theorem 1 (Differential entropy is invariant to translations) *The differential entropy of a continuous random variable* X *does not change if* X *is translated by a constant* c.

$$h(X + c) = h(X) \tag{3}$$

Theorem 2 (Scaling changes differential entropy) *The differential entropy of a continuous random variable* X *changes if* X *is scaled by a constant* a.

$$h(aX) = h(X) + \log |a| \tag{4}$$

Corollary:
More generally, for a continuous random vector $\mathbf{X}$ in $\mathbb{R}^n$, and for any invertible $n \times n$ matrix $\mathbf{A}$, we can write:

$$h(\mathbf{AX}) \quad = \quad h(\mathbf{X}) + \log |\mathbf{det}(\mathbf{A})|, \tag{5}$$

where $|\mathbf{det}(\mathbf{A})|$ denotes the absolute value of the determinant of $\mathbf{A}$.

Example (On the differential entropy of an additive-noise channel) Consider the channel: $Y = X + Z$, where the input X, the output Y and the additive noise Z are random variables distributed according to well-behaved pdf's. Furthermore, assume that X and Z are independent.
Now, let's consider the two-dimensional vector map

$$\begin{bmatrix} X \\ Z \end{bmatrix} \quad \rightarrow \quad \begin{bmatrix} X \\ X + Z \end{bmatrix}, \tag{6}$$

which translate into:

$$\mathbf{A}\begin{bmatrix} X \\ Z \end{bmatrix} = \begin{bmatrix} X \\ X+Z \end{bmatrix}, \tag{7}$$

and leads to $\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$. Therefore, $|\mathbf{det}(\mathbf{A})| = 1$. By using Equation 5, we get:

$$h\left(\mathbf{A}\begin{bmatrix} X \\ Z \end{bmatrix}\right) = h\left(\begin{bmatrix} X \\ X+Z \end{bmatrix}\right) \tag{8}$$

$$= h\left(\begin{bmatrix} X \\ Z \end{bmatrix}\right) + \log(1) \tag{9}$$

$$= h\left(\begin{bmatrix} X \\ Z \end{bmatrix}\right), \tag{10}$$

which implies that $h\left(\begin{bmatrix} X \\ X+Z \end{bmatrix}\right) = h\left(\begin{bmatrix} X \\ Z \end{bmatrix}\right)$.
By the chain rule for entropy, we get: $h(X)+h(X+Z|X) = h(X)+h(Z|X)$, i.e. $h(X+Z|X) = h(Z|X)$. This means that

$$h(Y|X) = h(X+Z|X) = h(Z|X) = h(Z), \tag{11}$$

where the last equality follows from the fact that X and Z are independent. [1] This says that given X, the uncertainty remaining in Y is the same as the differential entropy of Z.

Definition 3 (Mutual information between continuous random variables) *The mutual information $I(X;Y)$ between two random variables X and Y, with joint density $f_{X,Y}(x,y)$, and marginal densities $f_X(x)$ and $f_Y(y)$ respectively, is defined as*

$$I(X;Y) = \int\int f_{X,Y}(x,y)\log\frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}dxdy \tag{12}$$

From the definition, we can easily show that

$$I(X;Y) = h(X) - h(X|Y) = h(Y) - h(Y|X) \tag{13}$$

Theorem 3 (Relation of differential entropy to discrete entropy) *Consider a random variable X with a Riemann-integrable density $f_X(x)$. Suppose we divide the range of X into bins of length ϵ. Let $H(X_\epsilon)$ denote the entropy of the discretized version of X. Then*

$$h(X) = \lim_{\epsilon\to 0}[H(X_\epsilon) + \log\epsilon] \tag{14}$$

Example (On the mutual information between discrete and continuous r.v.'s) As a simple application of Eqs. 13 and 14, we would like to consider the case where X is a discrete-valued random variable, and Y is continuous-valued random variable with a Riemann integrable density $f_Y(y)$. Let Y_ϵ denote the ϵ-disretization of Y. The mutual information between X and Y can be written as

$$I(X;Y) = H(X) - H(X|Y) \tag{15}$$

But does the symmetry property of mutual information hold? That is, can we write $H(X) - H(X|Y) = h(Y) - h(Y|X)$? This turns out to be true, because of the following:

$$h(Y) - h(Y|X) = \lim_{\epsilon\to 0}[H(Y_\epsilon) + \log\epsilon] - \lim_{\epsilon\to 0}[H(Y_\epsilon|X) + \log\epsilon] \quad \text{By Eq. 14} \tag{16}$$

$$= \lim_{\epsilon\to 0}[H(Y_\epsilon) - H(Y_\epsilon|X)] \tag{17}$$

$$= I(X;Y_\epsilon) \tag{18}$$

$$= H(X) - \lim_{\epsilon\to 0} H(X|Y_\epsilon) \tag{19}$$

[1] This result can be generalized for the case of a input vector $\mathbf{X}^n$, noise vector $\mathbf{Z}^n$ and output vector $\mathbf{Y}^n$. In this case, $\mathbf{A}$ is a $2n \times 2n$ matrix whose determinant's absolute value is 1.

This says that the mutual information between X and Y is the limit of the mutual information between their quantized versions. Now, using the fact that $\lim_{\epsilon \to 0} H(X|Y_\epsilon) = H(X|Y)$, we get the desired result, i.e.

$$H(X) - H(X|Y) = h(Y) - h(Y|X) \tag{20}$$

Example (On the capacity of the *"6.441 Channel"*) Recall that the *"6.441 Channel"* is an additive-noise channel that has an input X distributed on the [-1, 1] interval, and a uniformly distributed noise on the (-ϵ, +ϵ) interval. It follows that the output Y is distributed between (-1 - ϵ) and (+1 + ϵ), and that the capacity C of this channel can be bounded as follows

$$\log\left(1 + \left\lfloor \frac{1}{\epsilon} \right\rfloor\right) \leq C \leq \log\left(1 + \frac{1}{\epsilon}\right) \text{ bits}, \tag{21}$$

with the upper and lower bound being equal if $\frac{1}{\epsilon}$ is an integer. However, the lower bound may be loose if, for instance, $\epsilon = 1.5$, which leads to $C \geq 0$. Yet, we know that we can achieve a capacity of at least $\frac{1}{2}$ bit if we represent the *"6.441 Channel"* as a binary erasure channel. Indeed, and as shown in Figure 1, the *"6.441 Channel"* can be thought of as a BER channel if we map a subset $\mathcal{S}_1$ from the support set of X to the input value 0 of the BER channel, and the complement of $\mathcal{S}_1$ to the input value 1 of the BER channel.

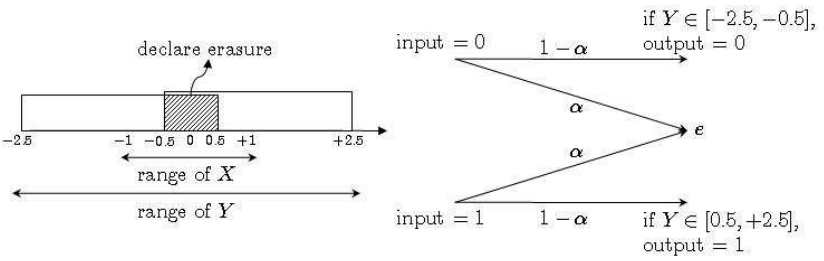

Figure 1: *"6.441 Channel"* and BER Channel

Under the above scenario, the erasure probability α is guaranteed to be at most 0.5, which means that the capacity of the *"6.441 Channel"* is at least 0.5 bit.

3 Capacity of the AWGN Channel

In this section, we derive the capacity of the AWGN channel. But before doing that, let's start by stating some facts about the AWGN channel.

3.1 What is an AWGN Channel?

An AWGN channel (see Figure 2) is a continuous-alphabet, time-discrete memoryless channel where, at each time unit, the output Y can be written as the sum of the input X and the noise Z

$$Y = X + Z \qquad Z \sim \mathcal{N}(0, \sigma^2) \tag{22}$$

The additive noise Z is assumed to be independent of the channel input X, and is represented a zero-mean Gaussian random variable with variance σ^2, and with density

$$f_Z(z) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{z^2}{2}} \tag{23}$$

A zero-mean Gaussian random variable is extensively used in the literature to model noise, since it serves as a good approximation to the cumulative effect of a large number of small random sources of noise (by the Central Limit Theorem). The term *white* is used to indicate that the noise's spectral density is flat over the frequency band of interest. In the time-domain, this says that the covariance function looks like a short duration pulse around $t = 0$. Roughly speaking, this means that the noise samples are mutually independent.

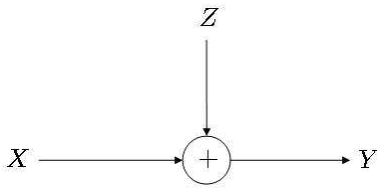

Figure 2: The AWGN Channel

3.2 Power Constraint

As we have previously mentioned, the noise Z is assumed to be independent of the signal X. But without further conditions, the capacity of this channel may be infinite, and this can happen if the noise variance is zero or if the input is unconstrained. This suggests that we need some sort of constraint on the channel input, and a good choice is the power constraint. As a result, an AWGN channel is usually specified by an upper-bound P on the signal

$$\mathbb{E}[X^2] \leq P, \tag{24}$$

which is equivalent to the constraint

$$Var(X) \leq P \quad \text{and} \quad \mathbb{E}[X] = 0 \tag{25}$$

Under these conditions, an AWGN channel is specified by a set of two parameters $\{\sigma^2, P\}$. Intuitively, if we double both the noise variance and the signal power, the capacity should remain unchanged; this suggests that the capacity of the channel should be a function of the ratio $\frac{P}{\sigma^2}$, an idea which we are going to formalize in what follows.

3.3 Information Capacity of an AWGN Channel

In this subsection, we define the information capacity of the AWGN channel as the maximum of the mutual information between the input and the output over all distributions of the input that satisfy the power constraint defined above.

Definition 4 (Information capacity of an AWGN channel) *The information capacity of the AWGN channel with power constraint P is defined as*

$$C = \max_{f(x):\mathbb{E}[X^2]\leq P} I(X;Y) \tag{26}$$

By expanding the mutual information, we get

$$\begin{aligned} I(X;Y) &= h(Y) - h(Y|X) && (27)\\ &= h(Y) - h(X+Z|X) && (28)\\ &= h(Y) - h(Z), && (29) \end{aligned}$$

where Eq. 29 follows from the result in Eq. 11. Recall that if $Z \sim \mathcal{N}(0, \sigma^2)$, then its differential entropy is: $h(Z) = \frac{1}{2}\log(2\pi e\sigma^2)$. We also remark that since X and Z are independent, and using the fact that $Var(X) \leq P$, then

$$\begin{aligned} Var(Y) &= Var(X) + Var(Z) && (30)\\ &\leq P + \sigma^2 && (31) \end{aligned}$$

Moreover, we use the fact that the Gaussian distribution maximizes the entropy for a given variance. Applying this fact to the received signal Y, whose variance is upper-bounded by $P + \sigma^2$, we get that

$h(Y) \leq \frac{1}{2}\log[2\pi e(P+\sigma^2)]$. This says that the input which maximizes this entropy is $X \sim \mathcal{N}(0,P)$. We are now ready to upper-bound the mutual information

$$\begin{aligned} I(X;Y) &= h(Y) - h(Z) && (32)\\ &\leq \frac{1}{2}\log[2\pi e(P+\sigma^2)] + \frac{1}{2}\log(2\pi e\sigma^2) && (33)\\ &= \frac{1}{2}\log\left(1+\frac{P}{\sigma^2}\right) && (34)\end{aligned}$$

By using Eq. 26, we finally get that the information capacity of the AWGN channel is

$$C = \max_{f(x):\mathbb{E}[X^2]\leq P} I(X;Y) = \frac{1}{2}\log\left(1+\frac{P}{\sigma^2}\right), \quad (35)$$

and this maximum is achieved when $X \sim \mathcal{N}(0,P)$, i.e. $f(x) = \frac{1}{\sqrt{2\pi P}}e^{-\frac{x^2}{2}}$. In communication theory, the ratio $\frac{P}{\sigma^2}$ is often called signal-to-noise ratio (SNR). In the next subsection, we show that the capacity that we just computed is also the supremum of all achievable rates of the channel, i.e. we give the above equation its *operational* meaning.

3.4 *Operational* Meaning of the Capacity of the AWGN Channel

In this subsection, we first show the capacity in Eq. 35 can be achieved using an argument based on the Weak Law of Large Numbers (WLLN) and the sphere-packing method. Then we formalize our proof and show that indeed, Eq. 35 is also the supremum of the achievable rates.

3.4.1 Sphere Packing method

The idea that will be raised in this paragraph is rather a plausibility argument rather than a formal proof. It emanates from the following question: *Given Eq. 35, and for n uses of the channel, will we be able to send* $\frac{n}{2}\log\left(1+\frac{P}{\sigma^2}\right)$ *bits with low probability of error?* The answer turns out to be yes, as it is outlined next.

Suppose that the sequence $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is transmitted over n symbol durations, where the X_i's are i.i.d. $\sim \mathcal{N}(0,P)$. Using the WLLN, we can show that, with high probability, $||\mathbf{X}||^2 = \sum_{i=1}^n X_i^2 \approx nP$. This implies that, with high probability, the transmitted codeword $\mathbf{x}$ lies within an n-dimensional sphere of radius $\approx \sqrt{nP}$. Note that high-dimensional spheres have almost all their volume concentrated in their shell, which means that the typical set of $\mathbf{X}$ lies in the shell of the sphere of radius $\approx \sqrt{nP}$ (see Figure 3).

Furthermore, recall that the noise $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_n)$ has i.i.d. components Z_i that are drawn according to a zero-mean Gaussian distribution with variance σ^2, i.e. $Z_i \sim \mathcal{N}(0,\sigma^2)$. The WLLN asserts that, with high probability, $||\mathbf{Z}||^2 = \sum_{i=1}^n Z_i^2 \approx n\sigma^2$. This says that, given a specific codeword $\mathbf{x}$, the received signal lies on the shell of a sphere of radius $\sqrt{n\sigma^2}$, and centered at $\mathbf{x}$ (see Figure 3). Therefore, the received sequences $\mathbf{y}$ lie within a sphere of radius $\approx \sqrt{nP} + \sqrt{n\sigma^2}$.

When we encode our sequences, we want the "noise" spheres (i.e. the spheres centered around the codewords $\mathbf{x}$, and whose radius is approximately equal to $\sqrt{n\sigma^2}$) to be more or less disjoint, so that we can decode with low probability of error. The *sphere packing* or *Kepler conjecture* problem answers the following question: *How many such balls can we pack such that all of them are disjoint?* Roughly, the answer turns out to be

$$\text{Number of balls} \leq \frac{\text{Volume of big ball}}{\text{Volume of small ball}} \quad (36)$$

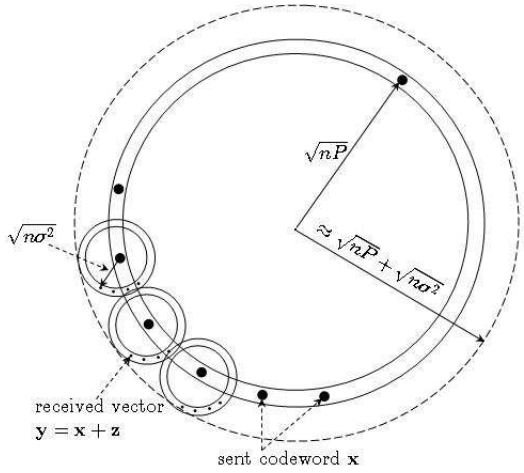

Figure 3: Sphere packing for the AWGN channel

$$= \frac{(\sqrt{nP} + \sqrt{n\sigma^2})^n}{(\sqrt{n\sigma^2})^n} \tag{37}$$

$$\approx \left(\frac{\sqrt{nP}}{\sqrt{n\sigma^2}}\right)^n \tag{38}$$

$$= 2^{\frac{n}{2}\log\frac{P}{\sigma^2}}, \tag{39}$$

where Eq. 38 uses the fact that the ratio $\frac{\sigma^2}{P}$ is assumed to be very small. In a nutshell, this very rough plausibility argument shows that the rate of the code is approximately $\frac{1}{2}\log\left(\frac{P}{\sigma^2}\right)$. Moreover, it indicates that we cannot hope to transmit data at rates greater than C with low probability of error. Yet, a more formal and cleaner proof of the operational meaning of capacity is derived in what follows.

3.4.2 Channel Coding Theorem for AWGN Channels

In this paragraph, we will formally prove that the capacity of an AWGN channel with power constraint P and noise variance σ^2 is the same as the information capacity defined in Eq. 35. But first, let's start by stating some definitions.

Definition 5 ((M,n) code for an AWGN channel) *A (M, n) code for the AWGN channel with power constraint P consists of the following:*

- *An message space $\{1, 2, \ldots, M\}$, where $M = 2^{nR}$, and R is the rate of the (M, n) code (in bits per transmission).* [2]
- *An encoding function $x : \{1, 2, \ldots, M\} \to \mathcal{X}^n$, yielding codewords $\mathbf{x}(m) = (x_1(m), x_2(m), \ldots, x_n(m))$. The codewords have i.i.d. components that satisfy the power constraint $X_i(m) \sim \mathcal{N}(0, P - \epsilon)$.*
- *A decoding function $g : \mathcal{Y}^n \to \{1, 2, \ldots, M\}$ that operates on the received sequence $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ as follows: if $\exists$ a unique m such that*

$$\sum_{i=1}^{n} |x_i(m) - y_i|^2 \leq n(\sigma^2 + \epsilon), \tag{40}$$

then output message m (i.e. choose nearest ball). Otherwise, declare an error.

[2] More precisely, $M = \lceil 2^{nR} \rceil$. However, we drop the ceiling function to simplify the notation.

Definition 6 (Achievable rate) [3] *A rate R is said to be achievable for an AWGN channel with 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 $\lambda^{(n)}$ tends to zero. The capacity of the channel is the supremum of the achievable rates*

Theorem 4 (Capacity of an AWGN Channel) *The capacity of an AWGN channel with power constraint P and noise variance σ^2 is*

$$C = \frac{1}{2}\log\left(1 + \frac{P}{\sigma^2}\right) \quad \textit{bits per tranmission} \tag{41}$$

Proof [Achievability] We begin by analyzing the probability of error. First, let's define the following events (assuming codeword m was transmitted):

$$E_0 = ||\mathbf{X}(m)||^2 > nP \quad \text{(i.e. too much signal power)} \tag{42}$$

$$E_1 = ||\mathbf{Z}||^2 > n(\sigma^2 + \epsilon) \quad \text{(i.e. too much noise variance)} \tag{43}$$

$$E_2 = \bigcup_{m' \neq m} E_2(m') : \sum_i ||\mathbf{y}_i - \mathbf{x}_i(m')||^2 \leq n(\sigma^2 + \epsilon) \tag{44}$$

By denoting $\mathbf{P}_e$ as the probability of error, we get

$$\mathbf{P}_e = \mathbf{P}(\text{encoding + decoding error}) = \mathbf{P}(E_0) + \mathbf{P}(E_1) + \mathbf{P}(E_2) \tag{45}$$

By the WLLN, $\mathbf{P}(E_0) \to 0$, and $\mathbf{P}(E_1) \to 0$, as $n \to \infty$. Hence, what remains is to analyze the probability of event E_2; more specifically, we want to analyze the probability of $E_2(m')$. We can define the probability that event $E_2(m')$ occurs as follows:

$\mathbf{P}(E_2(m')) = \mathbf{P}[\mathbf{X}(m')$ and $\mathbf{Y}$ are independent but jointly typical for the distribution of (X,Y)].

Now consider $\{(X_i, Y_i)\}_{i=1}^n$ and $\{(\widetilde{X}_i, \widetilde{Y}_i)\}_{i=1}^n$ each to be i.i.d. and drawn $\sim (X, Y)$. Then, it follows that the typical set for $\{(X_i, Y_i)\}_{i=1}^n$ has size $\approx 2^{nH(X,Y)}$, and that the probability that $\{(\widetilde{X}_i, \widetilde{Y}_i)\}_{i=1}^n$ is in the typical set is $\leq 2^{-nI(X;Y)}$. Therefore, $\mathbf{P}(E_2(m')) \leq 2^{-nI(X;Y)}$. By using the union bound, we get that

$$\mathbf{P}_e \approx \mathbf{P}(E_2) \tag{46}$$

$$= \mathbf{P}\left(\bigcup_{m' \neq m} E_2(m')\right) \tag{47}$$

$$\leq 2^{nR} 2^{-nI(X;Y)} \tag{48}$$

$$= 2^{-n(I(X;Y)-R)} \tag{49}$$

For n sufficiently large, and $R < I(X;Y)$, the probability of error goes to zero, which proves the existence of a good $(2^{nR}, n)$ code. Therefore, the forward part of the theorem is proved. We will prove the converse part in the next lecture. ■

[3] Cf. Cover and Thomas, p. 242.