ECE 771
Lecture 10 – The Gaussian channel

Objective: In this lecture we will learn about communication over a channel of practical interest, in which the transmitted signal is subjected to additive white Gaussian noise. We will derive the famous capacity formula.

1 The Gaussian channel

Suppose we send information over a channel that is subjected to additive white Gaussian noise. Then the output is

$$Y_i = X_i + Z_i$$

where Y_i is the channel output, X_i is the channel input, and Z_i is zero-mean Gaussian with variance N: $Z_i \sim \mathcal{N}(0, N)$. This is different from channel models we saw before, in that the output can take on a continuum of values. This is also a good model for a variety of practical communication channels.

We will assume that there is a constraint on the input power. If we have an input codeword $(x_1, x_2, \ldots, x_n)$, we will assume that the **average power** is constrained so that

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 \leq P$$

Let is consider the probability of error for binary transmission. Suppose that we can send either $+\sqrt{P}$ or $-\sqrt{P}$ over the channel. The receiver looks at the received signal amplitude and determines the signal transmitted using a threshold test. Then

$$\begin{aligned}
P_e &= \frac{1}{2}P(Y < 0|X = +\sqrt{P}) + \frac{1}{2}P(Y > 0|X = -\sqrt{P}) \\
&= \frac{1}{2}P(Z < -\sqrt{P}|X = +\sqrt{P}) + \frac{1}{2}P(Z > \sqrt{P}|X = -\sqrt{P}) \\
&= P(Z > \sqrt{P}) \\
&= \int_{\sqrt{P}}^{\infty} \frac{1}{\sqrt{2\pi N}} e^{-x^2/2N} dx \\
&= Q(\sqrt{P/N}) = 1 - \Phi(\sqrt{P/N})
\end{aligned}$$

where

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^{\infty} e^{-x^2/2} dx$$

or

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-x^2/t} dx$$

Definition 1 The **information capacity** of the Gaussian channel with power constraint is

$$C = \max_{p(x):EX^2 \leq P} I(X;Y).$$

□

We can compute this as follows:

$$\begin{aligned} I(X;Y) &= h(Y) - h(Y|X) \\ &= h(Y) - h(X+Z|X) \\ &= h(Y) - h(Z|X) \\ &= h(Y) - h(Z) \\ &\leq \frac{1}{2}\log 2\pi e(P+N) - \frac{1}{2}\log 2\pi eN \\ &= \frac{1}{2}\log(1+P/N) \end{aligned}$$

since $EY^2 = P + N$ and the Gaussian is the maximum-entropy distribution for a given variance. So

$$C = \frac{1}{2}log(1+P/N),$$

bits per channel use. The maximum is obtained *when X is Gaussian distributed.* (How do we make the input distribution look Gaussian?)

Definition 2 An (M, n) code for the Gaussian channel with power constraint P consists of the following:

1. An index set $\{1, 2, \ldots, M\}$
2. An encoding function $x : \{1, \ldots, M\} \to \mathcal{X}^n$, which maps an input index into a sequence that is n elements long, $x^n(1), x^n(2), \ldots, x^n(M)$, such that the average power constraints is satisfied:

$$\sum_{i=1}^{n} (x_i^n(w))^2 \leq nP$$

   for $w = 1, 2, \ldots, M$.
3. A decoding function $g : \mathcal{Y}^n \to \{1, 2, \ldots, M\}$.

□

Definition 3 A rate R is said to be *achievable* for a 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 $\lambda^{(n)} \to 0$. The **capacity** of the channel is the supremum of the achievable rates. □

Theorem 1 *The capacity of a Gaussian channel with power constraint P and noise variance N is*

$$C = \frac{1}{2}\log\left(1 + \frac{P}{N}\right) \text{ bits per transmission.}$$

Geometric plausibility For a codeword of length n, the received vector (in n space) is normally distributed with mean equal to the true codeword. With high probability, the received vector is contained in sphere about the mean of radius $\sqrt{n(N+\epsilon)}$. Why? Because with high probability, the vector falls within one standard deviation away from the mean in each direction, and the total distance away is the Euclidean sum:

$$E[z_1^2 + z_2^2 + \cdots z_n^2] = nN.$$

This is the square of the expected distance within which we expect to fall. If we assign everything within this sphere to the given codeword, we misdetect only if we fall outside this codeword.

Other codewords will have other spheres, each with radius approximately $\sqrt{n(N+\epsilon)}$. The received vectors a limited in energy by P, so they all must lie in a sphere of radius $\sqrt{n(P+N)}$. The number of (approximately) nonintersecting decoding spheres is therefore

$$\text{number of spheres} \approx \frac{\text{volume of sphere in } n\text{-space with radius } r = \sqrt{n(P+N)}}{\text{volume of sphere in } n\text{-space with radius } r = \sqrt{n(N+\epsilon)}}$$

The volume of a sphere of radius r in n space is proportional to r^n. Substituting in this fact we get

$$\text{number of spheres} \approx \frac{(n(P+N))^{n/2}}{(n(N+\epsilon))^{n/2}} \approx 2^{\frac{n}{2}(1+\frac{P}{N})}$$

Proof We will follow essentially the same steps as before.

1. First we generate a codebook *at random*. This time we generate the codebook according to the Gaussian distribution: let $X_i(w), i = 1, 2, \ldots, n$ be the code sequence corresponding to input index w, where each $X_i(w)$ is selected at random i.i.d. according to $\mathcal{N}(0, P-\epsilon)$. (With high probability, this has average power P.) The codebook is known by both transmitter and receiver.

2. Encode as described above.

3. The receiver gets a Y^n, and looks at the list of codewords $\{X^n(w)\}$ and searches for one which is *jointly typical* with the received vector. If there is only one such vector, it is declared as the transmitted vector. If there is more than one such vector, an error is declared. An error is also declared if the chosen codeword does not satisfy the power constraint.

For the probability of error, assume w.o.l.o.g. that codeword 1 is sent:

$$Y^n = X^n(1) + Z^n$$

Define the following events:

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

(the event that the codeword exceeds the power constraint) and

$$E_i = \{(X^n(i), Y^n) \text{ is in } A_\epsilon^{(n)}\}$$

The probability of error is then

$$\begin{aligned} P(\mathcal{E}) &= P(E_0 \cup E_1^c \cup E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}}) \\ &\le P(E_0) + P(E_1^c) + \sum_{i=2}^{2^{nR}} P(E_i) \qquad \text{union bound} \end{aligned}$$

By LLN, $P(E_0) \to 0$. By joint AEP, $P(E_1^c) \to 0$, so $P(E_1^c) \le \epsilon$ for n sufficiently large. By the code generation process, $X^n(1)$ and $X^n(i)$ are independent, so are

Y^n and $X^n(i), i \neq 1$. So the probability that $X^n(1)$ and Y^n are jointly typical is $\leq 2^{-n(I(X;Y)-3\epsilon)}$ by joint AEP. So

$$\begin{aligned} P_e^{(n)} &\leq \epsilon + \epsilon + \sum_{i=2}^{2^{nR}} 2^{-n(I(X;Y)-3\epsilon)} \\ &\leq 2\epsilon + (2^{nR}-1)2^{-n(I(X;Y)-3\epsilon)} \\ &=\leq 2\epsilon + 2^{nR}2^{-n(I(X;Y)-3\epsilon)} \leq 3\epsilon \end{aligned}$$

for n sufficiently large, if $R < I(X;Y) - 3\epsilon$.

This gives the average probability of error: we then go through the same kinds of arguments as before to conclude that the maximum probability of error also must go to zero.

□

The converse is that rate $R > C$ are not achievable, or, equivalently, that if $P_e^{(n)} \to 0$ then it must be that $R \leq C$.

Proof The proof starts with Fano's inequality:

$$H(W|Y^n) \leq 1 + nRP_e^{(n)} = n\epsilon_n$$

where

$$\epsilon_n = \frac{1}{n} + RP_e^{(n)}$$

and $\epsilon_n \to 0$ as $n \to \infty$.

The proof is a string of inequalities:

$$\begin{aligned} nR = H(W) &= I(W;Y^n) + H(W|Y^n) \qquad \text{uniform } W\text{; definition of } I \\ &\leq I(W;Y^n) + n\epsilon_n \qquad \text{Fano's inequality} \\ &= h(Y^n) - h(Y^n|X^n) + n\epsilon_n \\ &= h(Y^n) - h(Z^n) + n\epsilon_n \\ &\leq \sum_{i=1}^n h(Y_i) - h(Z^n) + n\epsilon_n \\ &= \sum_{i=1}^n h(Y_i) - \sum_{i=1}^n h(Z_i) + n\epsilon_n \\ &\leq \sum_{i=1}^n \frac{1}{2}\log 2\pi e(P_i+N) - \frac{1}{2}\log 2\pi eN + n\epsilon_n \qquad \text{entropies of } Y \text{ and } Z\text{; power constraint} \\ &= \sum_{i=1}^n \frac{1}{2}\log(1+P_i/N) + n\epsilon_n \\ &= n\left(\frac{1}{n}\sum_{i=1}^n \log(1+P_i/N)\right) + n\epsilon_n \\ &\leq n\log(1+\frac{1}{n}\sum_{i=1}^n P_i/N) + n\epsilon_n \qquad \text{Jensen's} \\ &\leq n\frac{1}{2}\log(1+P/N) + n\epsilon_n. \end{aligned}$$

Dividing through by n,

$$R \leq \frac{1}{2}\log(1+P/N) + \epsilon_n.$$

□

2 Band-limited channels

We now come to the first time in the book where the information is actually carried by a *time-waveform*, instead of a random variable. We will consider transmission over a band-limited channel (such as a phone channel). A key result is the sampling theorem:

Theorem 2 *If $f(t)$ is bandlimited to W Hz, then the function is completely determined by samples of the function taken every $\frac{1}{2W}$ seconds apart.*

This is the classical Nyquist sampling theorem. However, Shannon's name is also attached to it, since he provided a proof and used it. A representation of the function $f(t)$ is

$$f(t) = \sum_n f(\frac{n}{2W}) \operatorname{sinc}(t - \frac{n}{2W})$$

where

$$\operatorname{sinc}(t) = \frac{\sin(2\pi W t)}{2\pi W t}$$

From this theorem, we conclude (the dimensionality theorem) that a *bandlimited function has only $2W$ degrees of freedom per second.*

For a signal which has "most" of the energy in bandwidth W and "most" of the energy in a time T, then there are about $2WT$ degrees of freedom, and the time- and band-limited function can be represented using $2WT$ orthogonal basis functions, known as the *prolate spheroidal* functions. We can view band- and time-limited functions as vectors in a $2TW$ dimensional vector space.

Assume that the noise power-spectral density of the channel is $N_0/2$. Then the noise power is $(N_0/2)(2W) = N_0 W$. Over the time interval of T seconds, the energy per sample (per channel use) is

$$\frac{PT}{2WT} = \frac{P}{2W}.$$

Use this information in the capacity:

$$\begin{aligned} C &= \frac{1}{2}\log(1 + \frac{P}{N}) \text{ bits per channel use} \\ &= \frac{1}{2}\log(1 + \frac{P}{N_0 W}) \text{ bits per channel use.} \end{aligned}$$

There are $2W$ samples each second (channel uses), so the capacity is

$$C = (2W)\frac{1}{2}\log(1 + \frac{P}{N_0 W}) \text{ bits/second}$$

or

$$\boxed{C = W\log(1 + \frac{P}{N_0 W})}$$

This is the famous and key result of information theory.

As $W \to \infty$, we have to do a little calculus to find that

$$C = \frac{P}{N_0}\log_2 e \text{ bits per second.}$$

This is interesting: even with infinite bandwidth, the capacity is not infinite, but grows linearly with the power.

Example 1 For a phone channel, take $W = 3300$ Hz. If the SNR is $P/N_0W = 40\text{dB} = 10000$, we get

$$C = 43850 \text{ bits per second.}$$

If $P/WN_0 = 20\text{dB} = 100$ we get

$$C = 21972 \text{ bits/second.}$$

(The book is dated.) □

We cannot do better than capacity!

3 Kuhn-Tucker Conditions

Before proceeding with the next section, we need a result from constrained optimization theory known as the Kuhn-Tucker condition.

Suppose we are minimizing some convex objective function $L(x)$,

$$\min L(x)$$

subject to a constraint

$$f(x) \leq 0.$$

Let the optimal value of x be x_0. Then either the constraint is inactive, in which case we get

$$\left.\frac{\partial L}{\partial x}\right|_{x_0} = 0$$

or, if the constraint is active, it must be the case that the objective function increases for all *admissible* values of x:

$$\frac{\partial L}{\partial x}_{x\in\mathcal{A}} \geq 0$$

where $\mathcal{A}$ is the set of admissible values, for which

$$\frac{\partial f}{\partial y} \leq 0.$$

(Think about what happens if this is not the case.) Thus,

$$\operatorname{sgn}\frac{\partial L}{\partial x} = -\operatorname{sgn}\frac{\partial f}{\partial x}$$

or

$$\frac{\partial L}{\partial x} + \lambda\frac{\partial f}{\partial x} = 0 \qquad \lambda \geq 0. \tag{1}$$

We can create a new objective function

$$J(x,\lambda) = L(x) + \lambda f(x),$$

so the necessary conditions become

$$\frac{\partial J}{\partial x} = 0$$

and

$$f(x) \leq 0$$

where

$$\lambda \begin{cases} \geq 0 & f(y) = 0 \qquad \text{constraint is active} \\ = 0 & f(y) < 0 \qquad \text{constraint is inactive.} \end{cases}$$

For a vector variable $\mathbf{x}$, then the condition (1) means:

$$\frac{\partial L}{\partial x} \text{ is parallel to } \frac{\partial f}{\partial x} \text{ and pointing in opposite directions,}$$

where $\frac{\partial L}{\partial x}$ is interpreted as the gradient.

In words, what condition (1) says is: *the gradient of L with respect to x at a minimum must be pointed in such a way that decrease of L can only come by violating the constraints.* Otherwise, we could decrease L further. This is the essence of the Kuhn-Tucker condition.

4 Parallel Gaussian channels

Parallel Gaussian channels are used to model bandlimited channels with a non-flat frequency response. We assume we have k Gaussian channels,

$$Y_j = X_j + Z_j, \qquad j = 1, 2, \ldots, k.$$

where

$$Z_j \sim \mathcal{N}(0, N_j)$$

and the channels are independent. The total power used is constrained:

$$E \sum_{j=1}^{k} X_j^2 \leq P.$$

One question we might ask is: how do we distribute the power across the k channels to get maximum throughput.

We can find the maximum mutual information (the information channel capacity) as

$$\begin{aligned} I(X_1, \ldots, X_k; Y_1, \ldots, Y_k) &= h(Y_1, \ldots, Y_k) - h(Y_1, \ldots, Y_k | X_1, \ldots, X_k) = h(Y_1, \ldots, Y_k) - h(Z_1, \ldots, Z_k) \\ &= h(Y_1, \ldots, Y_k) - \sum_{i=1}^{k} h(Z_i) \\ &\leq \sum_{i=1}^{k} h(Y_i) - h(Z_i) \\ &\leq \sum_{i} \frac{1}{2} \log(1 + P_i/N_i) \end{aligned}$$

Equality is obtained when the Xs are independent normally distributed. We want to distribute the power available among the various channels, subject to not exceeding the power constraint:

$$J(P_1, \ldots, P_k) = \sum_{i} \frac{1}{2} \log(1 + \frac{P_i}{N_i}) + \lambda \sum_{i=1}^{k} P_i$$

with a side constraint (not shown) that $P_i \geq 0$. Differential w.r.t. P_j to obtain

$$\frac{1}{P_j + N_j} + \lambda \geq 0.$$

with equality only if all the constraints are inactive. After some fiddling, we obtain

$$P_j = \nu - N_j$$

(since λ is a constant). However, we must also have $P_j \geq 0$, so we must ensure that we don't violate that if $N_j > \nu$. Thus, we let

$$P_j = (\nu - N_j)^+$$

where

$$(x)^+ = \begin{cases} x & x \geq 0 \\ 0 & x < 0 \end{cases}$$

and ν is chosen so that

$$\sum_{i=1}^{n} (\nu - N_i)^+ = P$$

Draw picture; explain "water filling."

5 Channels with colored Gaussian noise

We will extend the results of the previous section now to channels with non-white Gaussian noise. Let K_z be the covariance of the noise K_x the covariance of the input, with the input constrained by

$$\frac{1}{n} \sum_i E X_i^2 \leq P$$

which is the same as

$$\frac{1}{n} \operatorname{tr}(K_X) \leq P.$$

We can write

$$I(X_1, \ldots, X_n; Y_1, \ldots, Y_n) = h(Y_1, \ldots, Y_n) - h(Z_1, \ldots, Z_n)$$

where

$$h(Y_1, \ldots, Y_n) \leq \frac{1}{2} \log((2\pi e)^n |K_x + K_z|)$$

Now how do we choose K_x to maximize $K_x + K_z$, subject to the power constraint? Let

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

then

$$\begin{aligned} |K_x + K_z| &= |K_x + Q \Lambda Q^T| = |Q||Q^T K_x Q + \Lambda||Q^T| \\ &= |Q^T K_x Q + \Lambda| = |A + \lambda| \end{aligned}$$

where $A = Q^T K_x Q$. Observe that

$$\mathrm{tr}(A) = \mathrm{tr}(Q^T K_x Q) = \mathrm{tr}(Q^T Q K_x) = \mathrm{tr}(K_x)$$

So we want to maximize $|A + \Lambda|$ subject to $\mathrm{tr}(A) \leq nP$. The key is to use an inequality, in this case Hadamard's inequality. Hadamard's inequality follows directly from the "conditioning reduces entropy" theorem:

$$h(X_1, \ldots, X_n) \leq \sum h(X_i).$$

Let $\mathbf{X} \sim \mathcal{N}(0, K)$. Then

$$h(\mathbf{X}) = \frac{1}{2} \log(2\pi e)^n |K|$$

and

$$h(X_i) = \frac{1}{2} \log(2\pi e) K_{ii}$$

Substituting in and simplifying gives

$$|K| \leq \prod_i K_{ii}$$

with equality iff K is diagonal.

Getting back to our problem,

$$|A + \Lambda| \leq \prod_i (A_{ii} + \Lambda_{ii})$$

with equality iff A is diagonal. We have

$$\frac{1}{n} \sum_i A_{ii} \leq P$$

(the power constraint), and $A_{ii} \geq 0$. As before, we take

$$A_{ii} = (\nu - \lambda_i)^+$$

where ν is chosen so that

$$\sum A_{ii} = nP.$$

Now we want to generalize to a continuous time system. For a channel with AWGN and covariance matrix $K_Z^{(n)}$, the covariance is Toeplitz. If the channel noise process is stationary, then the covariance matrix is Toeplitz, and the eigenvalues of the covariance matrix tend to a limit as $n \to \infty$. The density of the eigenvalues on the real line tends to the power spectrum of the stochastic process. That is, if $K_{ij} = K_{i-j}$ are the autocorrelation values and the power spectrum is

$$S(\omega) = \mathcal{F}[r_k]$$

then

$$\lim_{M \to \infty} \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_M}{M} = \frac{1}{2\pi} \int_{-\pi}^{\pi} S(\omega) d\omega.$$

In this case, the water filling translates to water filling in the spectral domain. The capacity of the channel with noise spectrum $N(f)$ can be shown to be

$$C = \int \frac{1}{2} \log(1 + \frac{(\nu - N(f))^+}{N(f)}) df$$

where ν is chosen so that

$$\int (\nu - N(f))^+ df = P$$