Lecture 6: Gaussian Channels
Definition 18. The (joint) differential entropy of a continuous random vector $X^n \sim p_X^n(x)$ over $\mathbb{R}$ is:

$$h(X^n) = -\int p_X^n(x) \log p_X^n(x) dx = -\mathbb{E} \left[ \log p_X^n(X^n) \right]$$

- $h(X^n)$ is a concave function of the pdf $p_X^n(x)$, but it is not necessarily non-negative.

- Sometimes we write $h(p_X)$ when we wish to stress that $h(\cdot)$ is a function of the pdf $p_X$.

- Translation: $h(X^n + a) = h(X^n)$.

- Scaling: $h(A X^n) = h(X^n) + \log |A|$. 
Differential entropy (2)

Example: Uniform distribution:

\[ p_X(x) = \frac{1}{a} 1\{x \in [0, a]\} \]

Then, we have

\[ h(X) = \int_0^a \frac{1}{a} \log a \, dx = \log a \]

... more in general, for a uniform distribution in \( \mathbb{R}^n \) over a compact domain \( \mathcal{D} \) with volume \( \text{Vol}(\mathcal{D}) \) we have

\[ h(X^n) = \log \text{Vol}(\mathcal{D}) \]
Gaussian distribution

- The standard normal distribution $X \sim \mathcal{N}(0, 1)$ has density:

$$p_X(x) = \frac{1}{\sqrt{2\pi}\sigma^2} \exp \left( -\frac{x^2}{2\sigma^2} \right)$$

- The real multi-variate Gaussian distribution $X^n \sim \mathcal{N}(\mu, \Sigma)$ has density:

$$p_{X^n}(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp \left( -\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu) \right)$$
Real Gaussian random vectors

- We notice that, for Gaussian densities, $\log p_{X^n}(x)$ is a quadratic function of $x$. E.g., for the real case we have

$$
\log p_{X^n}(x) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma| - \frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \log e
$$

- Since $h(X^n)$ is invariant by translations, it must be independent of $\mu$. Hence, there is no loss of generality in assuming $\mu = 0$.

- Taking expectation and changing sign, noticing that $x^T \Sigma^{-1} x = \text{tr}(xx^T \Sigma^{-1})$ we have

$$
h(X^n) = \frac{n}{2} \log(2\pi) + \frac{1}{2} \log |\Sigma| + \frac{1}{2} \text{tr}(\mathbb{E}[xx^T] \Sigma^{-1}) \log e
= \frac{1}{2} \log((2\pi e)^n |\Sigma|)
$$
The complex circularly-symmetric Gaussian distribution $X^n \sim \mathcal{CN}(\mu, \Sigma)$, where $X^n = X^n_R + jX^n_I$, where $X^n_R \sim \mathcal{N}(\text{Re}\{\mu\}, \frac{1}{2}\Sigma)$, $X^n_I \sim \mathcal{N}(\text{Im}\{\mu\}, \frac{1}{2}\Sigma)$ and $X^n_R, X^n_I$ are statistically independent, has density:

$$p_{X^n}(x) = \frac{1}{\pi^n|\Sigma|} \exp\left(- (x - \mu)^H \Sigma^{-1} (x - \mu)\right)$$

Repeating the arguments of before, for the complex circularly symmetric case we have:

$$h(X^n) = \log((\pi e)^n|\Sigma|)$$
Conditional differential entropy

**Definition 19.** For two jointly distributed continuous random vectors $X^n, Y^m$ over $\mathbb{R}$, respectively, with joint pdf $p_{X^n,Y^m}(x,y)$, the conditional differential entropy of $X^n$ given $Y^m$ is:

$$h(X^n|Y^m) = -\int \int p_{X^n,Y^m}(x,y) \log p_{X^n|Y^m}(x|y) \, dx \, dy$$

$$= -\mathbb{E} \left[ \log p_{X^n|Y^m}(X^n|Y^m) \right]$$
Divergence and mutual information

**Definition 20.** Let \( p_X \) and \( q_X \) be pdfs defined on \( \mathbb{R} \). The divergence (Kullback-Leibler distance) between \( p_X \) and \( q_X \) is defined as

\[
D(p_X \| q_X) = \int p_X(x) \log \frac{p_X(x)}{q_X(x)} \, dx
\]

Notice that \( D(p_X \| q_X) < \infty \) if \( \text{supp}(p_X) \subseteq \text{supp}(q_X) \).

**Definition 21.** The mutual information of \( X, Y \sim p_{X,Y} \) is defined by

\[
I(X; Y) = \int p_{X,Y}(x, y) \log \frac{p_{X,Y}(x, y)}{p_X(x)p_Y(y)} \, dxdy
\]
Relations (similar to the discrete case)

- We have the following expressions:

\[
I(X; Y) = \mathbb{E} \left[ \log \frac{p_{X,Y}(X,Y)}{p_X(X)p_Y(Y)} \right]
\]

\[
= \mathbb{E} \left[ \log \frac{p_{X|Y}(X|Y)}{p_X(X)} \right] = h(X) - h(X|Y)
\]

\[
= \mathbb{E} \left[ \log \frac{p_{Y|X}(Y|X)}{p_Y(Y)} \right] = h(Y) - h(Y|X)
\]

\[
= D \left( p_{Y|X} \parallel p_Y \middle| p_X \right)
\]

\[
= D \left( p_{Y|X} \parallel q_Y \middle| p_X \right) - D(p_Y \parallel q_Y)
\]

where the last line holds for any pdf \( q_Y \) for which the divergences are finite.
Properties (1)

- Non-negativity and convexity of $D(p_X \| q_X)$ hold verbatim as in the discrete case.

- $I(X; Y) \geq 0$, therefore $h(Y|X) \leq h(Y)$ (Conditioning reduces differential entropy).

- The chain rule for entropy holds for differential entropy as well:

  $$h(X^n) = \sum_{i=1}^{n} h(X_i|X^{i-1})$$

- Independence upper bound

  $$h(X^n) \leq \sum_{i=1}^{n} h(X_i)$$
Properties (2)

• As an application of the independence bound for Gaussian differential entropies, we get Hadamard inequality:

\[
\frac{1}{2} \log ((2\pi e)^n |\Sigma|) \leq \sum_{i=1}^n \frac{1}{2} \log (2\pi e \Sigma_{i,i}) \Rightarrow |\Sigma| \leq \prod_{i=1}^n \Sigma_{i,i}
\]

• The chain rule for mutual information holds as well:

\[
I(X; Y^n) = \sum_{i=1}^n I(X; Y_i | Y^{i-1})
\]

where the conditional mutual information is \( I(X; Y | Z) = h(Y | Z) - h(Y | X, Z) \).
Extremal of differential entropy

**Theorem 15.** *Differential entropy maximizer:* Let $X^n$ be a random vector with mean zero and covariance matrix $\mathbb{E}[X^n(X^n)^\top] = \text{Cov}(X^n) = \Sigma$. Then,

$$h(X^n) \leq \frac{1}{2} \log \left((2\pi e)^n \mid \Sigma \right)$$

Let $X^n$ be a complex random vector with mean zero and covariance matrix $\mathbb{E}[X^n(X^n)^\mathbb{H}] = \text{Cov}(X^n) = \Sigma$. Then,

$$h(X^n) \leq \log ((\pi e)^n \mid \Sigma)$$
Proof:

Let $g(x)$ denote the Gaussian pdf with mean zero and covariance matrix $\Sigma$. Then

$$0 \leq D(p_{X^n} \parallel g)$$
$$= \int p_{X^n}(x) \log \frac{p_{X^n}(x)}{g(x)} \, dx$$
$$= -h(X^n) - \int p_{X^n}(x) \log g(x) \, dx$$
$$= -h(X^n) - \int g(x) \log g(x) \, dx$$
$$= -h(X^n) + h(g)$$
$$= -h(X^n) + \frac{1}{2} \log ((2\pi e)^n |\Sigma|)$$
Some consequences

**Theorem 16. Differential entropy and MMSE estimation:** Let $X$ be a continuous random variable and $\hat{X}$ an estimator of $X$. Then

$$\mathbb{E} \left[ |X - \hat{X}|^2 \right] \geq \frac{1}{2\pi e} e^{2h(X)}$$

with equality if and only if $X$ is Gaussian and $\hat{X} = \mathbb{E}[X]$.

**Proof:**

We have:

$$\mathbb{E} \left[ |X - \hat{X}|^2 \right] \geq \min_{\hat{X}} \mathbb{E} \left[ |X - \hat{X}|^2 \right]$$

$$= \mathbb{E} \left[ |X - \mathbb{E}[X]|^2 \right]$$

$$= \text{Var}(X)$$

$$\geq \frac{1}{2\pi e} e^{2h(X)}$$
where we have used the fact that the Minimum Mean-Square Error (MMSE) estimator for $X$ is its mean.

**Corollary 8.** Let $X$ and $Y$ be jointly distributed continuous random variables. For any estimator function $\hat{x}(y)$ of $X$ from $Y$, we have

$$
\mathbb{E} \left[ |X - \hat{x}(Y)|^2 \right] \geq \frac{1}{2\pi e} e^{2h(X|Y)}
$$

(Proof: HOMEWORK)
Corollary 9. Let $X^n$ and $Y^m$ be jointly distributed continuous random vectors, and let

$$\hat{x}(Y^m) = \mathbb{E}[X^n|Y^m], \quad E^n = X^n - \hat{x}(Y^m)$$

denote the MMSE estimator of $X^n$ given $Y^m$ and the corresponding estimation error sequence. Let $\Sigma_e = \text{Cov} (X^n - \hat{x}(Y^m))$ denote the covariance matrix of the MMSE error. Then,

$$h(X^n|Y^m) \leq \frac{1}{2} \log ((2\pi e)^n |\Sigma_e|)$$

with equality iff $X^n$ and $Y^m$ are jointly Gaussian.

(Proof: HOMEWORK)
• When $X^n, Y^m$ are jointly zero-mean Gaussian, then (sequences are considered as column vectors here), the MMSE estimator is given by:

$$\hat{x}(Y^m) = \Sigma_{xy} \Sigma_y^{-1} Y^m$$

$$\Sigma_e = \Sigma_x - \Sigma_{xy} \Sigma_y^{-1} \Sigma_{yx}$$

where $\Sigma_x = \text{Cov}(X^n)$, $\Sigma_y = \text{Cov}(Y^m)$, $\Sigma_{xy} = \mathbb{E}[X^n (Y^m)^\top]$ and $\Sigma_{yx} = \Sigma_{xy}^\top$.

• For two one-variate jointly Gaussian zero-mean random variables $X$ and $Y$, we have

$$h(X|Y) = \frac{1}{2} \log \left( \frac{\sigma_x^2 \sigma_y^2 - \sigma_{xy}^2}{\sigma_y^2} \right)$$

where $\sigma_x^2, \sigma_y^2$ are the variances and $\sigma_{xy} = \mathbb{E}[XY]$. 

Copyright G. Caire (Sample Lectures)
From discrete to continuous

- Let $\mathcal{X} \subseteq \mathbb{R}$ be the range of $X$ (you can generalize this to random vectors in $\mathbb{R}^n$, of course).

- A partition $\mathcal{P}$ of $\mathcal{X}$ is a finite collection of disjoint sets $\mathcal{P}_i$ such that $\bigcup_i \mathcal{P} = \mathcal{X}$.

- We denote by $[X]_\mathcal{P}$ the quantization of $X$ by the partition $\mathcal{P}$, such that

$$\mathbb{P}([X]_\mathcal{P} = i) = \mathbb{P}(X \in \mathcal{P}_i) = \int_{\mathcal{P}_i} p_X(x)dx$$

- For two random variables $X, Y$ with joint pdf $p_{X,Y}$, let $[X]_\mathcal{P}$ and $[Y]_\mathcal{Q}$ denote their discretized versions and define the DMC transition probability

$$P_{[Y]_\mathcal{Q}|[X]_\mathcal{P}}(s|r) = \mathbb{P}(Y \in \mathcal{Q}_s|X \in \mathcal{P}_r) = \frac{\int_{\mathcal{P}_r} \int_{\mathcal{Q}_s} p_{X,Y}(x,y)dydx}{\int_{\mathcal{P}_r} p_X(x)dx}$$
Continuous memoryless channels

- As an application of the data processing inequality we have:

$$I(X; Y) = \sup_{\mathcal{P}, \mathcal{Q}} I([X]_{\mathcal{P}}; [Y]_{\mathcal{Q}})$$

- The proof of the channel coding theorem for continuous-input, continuous-output, discrete-time channels, with transition density $p_{Y|X}(y|x)$, follows by defining a suitable sequence of discretized inputs (input partitions) and discretized output (output partitions), and taking the sup over all possible finite partitions.

- This step is standard, and applies to any well-behaved channel with well-behaved transition probability density. Hence, from now on, we will just **use** the channel coding theorem without further questions.
Gaussian channels

- Discrete-time, continuous-alphabet channel $\mathcal{X} = \mathcal{Y} = \mathbb{R}$, $Z \sim \mathcal{N}(0, N_0/2)$.

- Input-output relation: $Y = x + Z$.

- Average transmit power constraint:

$$\frac{1}{n} \sum_{i=1}^{n} |x_i(m)|^2 \leq E_s, \quad \forall \ m = 1, \ldots, 2^{nR}$$

- Codewords are points in $\mathbb{R}^n$ contained in a sphere of radius $\sqrt{nE_s}$. 
Theorem 17. The capacity of the AWGN channel with received Signal to Noise Ratio \( SNR = 2E_s/N_0 \) is given by

\[
C(SNR) = \frac{1}{2} \log(1 + SNR)
\]

Proof:

- From the channel capacity formula with input cost, we have:

\[
C(E_s) = \sup_{p_X : \mathbb{E}[|X|^2] \leq E_s} I(X; Y)
\]
Then:

\[
C(E_s) = \sup_{p_X : \mathbb{E}[|X|^2] \leq E_s} h(Y) - h(Y|X)
\]

\[
= \sup_{p_X : \mathbb{E}[|X|^2] \leq E_s} h(Y) - h(X + Z|X)
\]

\[
= \sup_{p_X : \mathbb{E}[|X|^2] \leq E_s} h(Y) - h(Z)
\]

\[
= \sup_{p_X : \mathbb{E}[|X|^2] \leq E_s} h(X + Z) - \frac{1}{2} \log(\pi e N_0)
\]

\[
= \frac{1}{2} \log(2\pi e (N_0/2 + E_s)) - \frac{1}{2} \log(\pi e N_0)
\]

\[
= \frac{1}{2} \log(1 + 2E_s/N_0)
\]

where the maximum is achieved by letting \( X \sim \mathcal{N}(0, E_s) \).
• For the converse, we consider a sequence of codes \( \{C_n\} \) of rate \( R \), with 
\[
P_e(n) \rightarrow 0 \quad \text{as} \quad n \rightarrow \infty
\]
and the relaxed (average) input cost constraint 
\[
2^{-nR} \sum_{m=1}^{2^n} \frac{1}{n} \sum_{i=1}^{n} |x_i(m)|^2 \leq E_s.
\]
Using Fano and data processing inequalities, we have

\[
nR = H(M) \leq I(X^n; Y^n) + n\epsilon_n = h(Y^n) - h(Z^n) + n\epsilon_n
\]

\[
= h(Y^n) - \frac{n}{2} \log(\pi e N_0) + n\epsilon_n
\]

\[
\leq \sum_{i=1}^{n} \frac{1}{2} \log \left( 2\pi e \left( \frac{N_0}{2} + \text{Var}(X_i) \right) \right) - \frac{n}{2} \log(\pi e N_0) + n\epsilon_n
\]

\[
\leq \frac{n}{2} \log \left( 2\pi e \left( \frac{N_0}{2} + \frac{1}{n} \sum_{i=1}^{n} \text{Var}(X_i) \right) \right) - \frac{n}{2} \log(\pi e N_0) + n\epsilon_n
\]

\[
\leq \frac{n}{2} \log \left( 2\pi e \left( \frac{N_0}{2} + E_s \right) \right) - \frac{n}{2} \log(\pi e N_0) + n\epsilon_n
\]
Geometric interpretation: sphere packing

- The volume of a sphere in $\mathbb{R}^n$ of radius $r$ is given by $V_n r^n$, where $V_n$ is the volume of the unit sphere (known formula, irrelevant in this context).

- The received random vector $Y^n = X^n + Z^n$, with high probability (law of large numbers) is inside a sphere of radius $\sqrt{n(N_0/2 + E_s)}$.

- The noise vector $Z^n$, with high probability, is inside a sphere of radius $\sqrt{nN_0/2}$.

- For large $n$, it is possible to pack “noise” spheres into the “typical output” sphere, with negligible empty space in between.

$$\text{number of spheres} = \frac{V_n n^\frac{n}{2} (N_0/2 + E_s)^\frac{n}{2}}{V_n n^\frac{n}{2} (N_0/2)^{\frac{n}{2}}} = (1 + 2E_s/N_0)^{\frac{n}{2}}$$
The resulting rate is:

\[ R = \frac{1}{n} \log \text{(number of spheres)} = \frac{1}{2} \log (1 + 2E_s/N_0) \]
An informal argument: consider a base-band waveform channel with single-sided bandwidth $W$ Hz.

Shannon’s sampling theorem: any (infinite duration) base-band signal can be represented uniquely from its samples, taken at the so-called Nyquist rate $2W$.  

Copyright G. Caire (Sample Lectures)
• Noise power

\[ N = \frac{N_0}{2} \times 2W = N_0W \]

• Signal power

\[ P_x = E_s \times 2W \]

• Signal to Noise Ratio (SNR):

\[ \text{SNR} = \frac{P_x}{N} = \frac{2E_s}{N_0} \]

• Capacity in bit/s: multiply the capacity in bit/symbol times the number of symbols/s:

\[ C(\text{SNR}) = 2W \times \frac{1}{2} \log \left( 1 + \frac{2E_s}{N_0} \right) = W \log(1 + \text{SNR}) \]
Continuous-time passband channels

- Consider a passband waveform channel with carrier frequency modulation $f_0 \gg W$ and single-sided bandwidth $W$ Hz centered around $f_0$.

- The complex baseband equivalent system has single sided bandwidth $W/2$, complex signal $X$ and complex circularly symmetric noise $Z \sim \mathcal{CN}(0, N_0)$.

- Noise power
  \[ N = N_0 \times W \]

- Signal power (optimal input $X \sim \mathcal{CN}(0, E_s)$).
  \[ P_x = E_s \times W \]
• Signal to Noise Ratio (SNR):

\[ \text{SNR} = \frac{P_x}{N} = \frac{E_s}{N_0} \]

• Repeating the capacity calculation for complex circularly symmetric input and output, we find

\[ C(\text{SNR}) = \log(1 + \text{SNR}) \]

• Capacity in bit/s: multiply the capacity in bit/symbol times the number of symbols/s:

\[ C(\text{SNR}) = W \times \log \left( 1 + \frac{E_s}{N_0} \right) = W \log(1 + \text{SNR}) \]
Consider the parallel complex circularly symmetric Gaussian channel defined by

\[ Y = Gx + Z \]

where \( x, Y, Z \in \mathbb{C}^K \), \( G = \text{diag}(G_1, \ldots, G_K) \) is a diagonal matrix of fixed coefficients, and \( Z \sim \mathcal{CN}(0, N_0 I_K) \).

Notice that the input and output alphabets are the \( K \)-dimensional vector space \( \mathbb{C}^K \), and codewords are sequences of vectors \( \mathbf{x}(m) = (x_1(m), \ldots, x_n(m)) \).

Input constraint: codewords \( \mathbf{x}(m) \), must satisfy

\[
\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} |x_{k,i}(m)|^2 \leq E_s
\]
The capacity is given by

\[
C(E_s) = \max_{p_X: \mathbb{E}[\|X\|^2] \leq E_s} I(X; Y)
\]

\[
= \max_{p_X: \mathbb{E}[\|X\|^2] \leq E_s} h(Y) - h(Z)
\]

\[
= \max_{p_X: \mathbb{E}[\|X\|^2] \leq E_s} h(Y) - K \log((\pi e)N_0)
\]

\[
= \max_{p_X: \mathbb{E}[\|X\|^2] \leq E_s} \sum_{k=1}^{K} \log \left( (\pi e)(N_0 + |G_k|^2 E_{s,k}) \right) - K \log((\pi e)N_0)
\]

\[
= \max_{\sum_{k=1}^{K} E_{s,k} \leq E_s} \sum_{k=1}^{K} \log \left( 1 + \frac{|G_k|^2 E_{s,k}}{N_0} \right)
\]
We must maximize the mutual information with respect to the power allocation over the parallel channels:

\[
\text{maximize} \quad \sum_{k=1}^{K} \log \left( 1 + \frac{|G_k|^2 E_{s,k}}{N_0} \right)
\]

subject to

\[
\sum_{k=1}^{K} E_{s,k} \leq E_s
\]

\[
E_{s,k} \geq 0, \quad \forall \ k
\]
The Lagrangian function is

$$\mathcal{L}(\{E_{k,s}\}, \lambda) = \sum_{k=1}^{K} \log \left(1 + \frac{|G_k|^2 E_{s,k}}{N_0}\right) - \lambda \left(\sum_{k=1}^{K} E_{s,k} - E_s\right)$$

We impose the KKT conditions

$$\frac{\partial \mathcal{L}}{\partial E_{s,k}} = \frac{|G_k|^2}{N_0} \frac{1}{1 + \frac{|G_k|^2 E_{s,k}}{N_0}} - \lambda \leq 0$$

with equality for all $k$ such that $E_{s,k}^* > 0$.

Solving for $E_{s,k}$ we find

$$E_{s,k}^* = \frac{1}{\lambda} - \frac{N_0}{|G_k|^2}$$
• We let $1/\lambda = \nu$, and argue that the KKT conditions are satisfied if we let

$$E_{s,k}^* = \left[ \nu - \frac{N_0}{|G_k|^2} \right]_+$$

In fact, if $\nu \geq \frac{N_0}{|G_k|^2}$ then $E_{s,k}^* > 0$ and $\frac{\partial L}{\partial E_{s,k}} = 0$. Otherwise, if $\nu < \frac{N_0}{|G_k|^2}$ then $E_{s,k}^* = 0$ and $\frac{\partial L}{\partial E_{s,k}} < 0$.

• This optimal power allocation is called “Waterfilling”.

• Intuitively, we spend more power over the better channels (the ones with higher gain).

• Replacing the water filling solution into the mutual information expression, we obtain

$$C(E_s) = \sum_{k=1}^{K} \left[ \log \left( \frac{\nu |G_k|^2}{N_0} \right) \right]_+$$

Copyright G. Caire (Sample Lectures)
End of Lecture 6