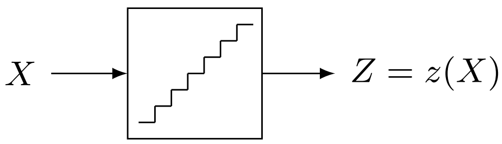
Information Theory

Lecture 5

- Continuous variables and Gaussian channels: CT8–9
  - Differential entropy: CT8
  - Capacity and coding for Gaussian channels: CT9

"Entropy" of a Continuous Variable

- A continuous random variable, X, with pdf $f(x)$.
- A quantizer $z(X)$, with quantizer interval Δ



where

$$i\Delta \leq X < (i+1)\Delta \implies Z = z(X) = x_i$$

for some $x_i \in [i\Delta, (i+1)\Delta]$.

- The variable Z has entropy

$$H(Z) = -\sum_i p(i) \log p(i),$$

where $p(i) = \Pr\big(i\Delta \leq X < (i+1)\Delta\big)$.

- Notice that

$$p(i) = \int_{i\Delta}^{(i+1)\Delta} f(x)dx = f(x_i)\Delta$$

for some $x_i \in [i\Delta, (i+1)\Delta]$. Hence for small Δ, we get

$$\begin{aligned} H(Z) &= -\sum_i f(x_i)\Delta \log\big(f(x_i)\Delta\big) \\ &= -\sum_i f(x_i)\Delta \log f(x_i) - \log\Delta \\ &\approx -\int_{-\infty}^{\infty} f(x)\log f(x)dx - \log\Delta \end{aligned}$$

(if $f(x)$ is Riemann integrable).

- Define the *differential entropy* $h(X)$, or $h(f)$, of X as

$$h(X) \triangleq -\int f(x)\log f(x)dx$$

(if the integral exists).

- Then for small Δ

$$H(Z) + \log\Delta \approx h(X)$$

- Note that $H(Z) \to \infty$, in general, even if $h(X)$ exists and is finite;
  - $h(X)$ is *not* "entropy," and $H(Z) \to h(X)$ does *not* hold!

- *Maximum differential entropy*:
  For any random variable X with pdf $f(x)$ such that

$$E[X^2] = \int x^2 f(x) dx = P$$

  it holds that

$$h(X) \leq \frac{1}{2} \log 2\pi e P$$

  with equality iff $f(x) = \mathcal{N}(0, P)$.

Typical Sets for Continuous Variables

- A discrete-time continuous-amplitude i.i.d. process $\{X_m\}$, with marginal pdf $f(x)$ of support $\mathcal{X}$.
- It holds that

$$-\lim_{n\to\infty} \frac{1}{n} \log f(X_1^n) = -E \log f(X_1) = h(f) \quad \text{a.s.}$$

- Define the *typical set* $A_\varepsilon^{(n)}$, with respect to $f(x)$, as

$$A_\varepsilon^{(n)} = \left\{ x_1^n \in \mathcal{X}^n : \left| -\frac{1}{n} \log f(x_1^n) - h(f) \right| \leq \varepsilon \right\}$$

- For $A \subset \mathbb{R}^n$, define

$$\mathrm{Vol}(A) \triangleq \int_A dx_1^n$$

- For n sufficiently large

$$\Pr\big(X_1^n \in A_\varepsilon^{(n)}\big) = \int_{A_\varepsilon^{(n)}} f(x_1^n) dx_1^n > 1 - \varepsilon$$

  and

$$\operatorname{Vol}\big(A_\varepsilon^{(n)}\big) \geq (1-\varepsilon) 2^{n(h(f)-\varepsilon)}$$

- For all n

$$\operatorname{Vol}\big(A_\varepsilon^{(n)}\big) \leq 2^{n(h(f)+\varepsilon)}$$

- Since $\operatorname{Vol}\big(A_\varepsilon^{(n)}\big) \approx 2^{nh(f)} = \big(2^{h(f)}\big)^n$, $h(f)$ is the logarithm of the side-length of a hypercube with the same volume as $A_\varepsilon^{(n)}$.
  - *Low* $h(f) \implies X_1^n$ typically lives in a *small subset* of $\mathbb{R}^n$.
- *Jointly typical sequences*: Straightforward extension.

Relative Entropy and Mutual Information

- Define the *relative entropy* between the pdfs f and g as

$$D(f\|g) = \int f(x) \log \frac{f(x)}{g(x)} dx$$

  and the *mutual information* between $(X, Y) \sim f(x, y)$ as

$$\begin{aligned} I(X;Y) &= D\big(f(x,y)\|f(x)f(y)\big) \\ &= \iint f(x,y) \log \frac{f(x,y)}{f(x)f(y)} dxdy \end{aligned}$$

- While $h(X)$, for a continuous real-valued X, does not have an interpretation as "entropy," both $D(f\|g)$ and $I(X;Y)$ have equivalent interpretations as in the discrete case.

- In fact, both relative entropy and mutual information exist, and their operational interpretations stay intact, under very general conditions.
- Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ be random variables (or "measurable functions") defined on a common abstract probability space $(\Omega, \mathcal{B}, P)$. Let $q(x)$ and $r(y)$ be "quantizers" that map X and Y, respectively, into real-valued discrete versions $q(X)$ and $r(Y)$. Then, mutual information is defined as

$$I(X;Y) \triangleq \sup I\big(q(X); r(Y)\big),$$

over all quantizers q and r. (The two previous definitions of $I(X;Y)$ are then special cases of this general definition.)

The Gaussian Channel

- A *continuous-alphabet memoryless channel* $(\mathcal{X}, f(y|x), \mathcal{Y})$ maps a continuous real-valued channel input $X \in \mathcal{X}$ to a continuous real-valued channel output $Y \in \mathcal{Y}$, in a stochastic and memoryless manner as described by the conditional pdf $f(y|x)$.
- A *memoryless Gaussian channel* (with noise variance σ^2) is defined as $\mathcal{X} = \mathcal{Y} = \mathbb{R}$, and

$$f(y|x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}(y-x)^2\Big).$$

That is, for a given $X = x$ the channel adds zero mean Gaussian "noise" Z, of variance σ^2, such that the variable $Y = x + Z$ is measured at its output.

- *Coding for a continuous* $\mathcal{X}$: if $\mathcal{X}$ is very large, or even $\mathcal{X} = \mathbb{R}$, coding needs to be defined *subject to a power constraint*.
- An (M, n) code with an average power constraint P:
  1. An *index set* $\mathcal{I}_M \triangleq \{1, \dots, M\}$.
  2. An *encoder mapping* $\alpha : \mathcal{I}_M \longmapsto \mathcal{X}^n$, which defines the *codebook*

$$\mathcal{C}_n \triangleq \left\{ x_1^n : x_1^n = \alpha(i),\ \forall\, i \in \mathcal{I}_M \right\} = \left\{ x_1^n(1), \dots, x_1^n(M) \right\},$$

  subject to

$$\frac{1}{n} \sum_{m=1}^{n} x_m^2(i) \le P, \quad \forall\, i \in \mathcal{I}_M.$$

  3. A *decoder mapping* $\beta : \mathcal{Y}^n \longmapsto \mathcal{I}_M$.

- A *rate*

$$R \triangleq \frac{\log M}{n}$$

is *achievable* (subject to the power constraint P) if there exists a sequence of $(\lceil 2^{nR} \rceil, n)$ codes with codewords satisfying the power constraint, and such that the maximal probability of error

$$\lambda^{(n)} = \max_i \Pr\big(\beta(Y_1^n) \neq i \mid X_1^n = x_1^n(i)\big)$$

tends to 0 as $n \to \infty$.

The *capacity* C is the *supremum of all rates that are achievable over the channel.*

Memoryless Gaussian Channel: Lower Bound for C

- *Gaussian random code design*: Fix the distribution

$$f(x) = \frac{1}{\sqrt{2\pi(P-\varepsilon)}} \exp\left(-\frac{x^2}{2(P-\varepsilon)}\right)$$

  and draw

$$\mathcal{C}_n = \left\{X_1^n(1), \dots, X_1^n(M)\right\}$$

  i.i.d. according to

$$f(x_1^n) = \prod_m f(x_m).$$

- *Encoding*: A message $\omega \in \mathcal{I}_M$ is encoded as $X_1^n(\omega)$

- *Transmission*: Received sequence

$$Y_1^n = X_1^n(\omega) + Z_1^n$$

  where $\{Z_m\}$ are i.i.d. zero-mean Gaussian with $E[Z_m^2] = \sigma^2$.

- *Decoding*: Declare $\hat{\omega} = \beta(Y_1^n) = i$ if $X_1^n(i)$ is the only codeword such that

$$(X_1^n(i), Y_1^n) \in A_\varepsilon^{(n)}$$

  and in addition $\frac{1}{n}\sum_{m=1}^n X_m^2(i) \le P$, otherwise set $\hat{\omega} = 0$.

- *Average probability of error*:

$$\pi_n = \Pr(\hat{\omega} \neq \omega) = \left\{\text{symmetry}\right\} = \Pr(\hat{\omega} \neq 1 | \omega = 1)$$

  with "Pr" over the random codebook and the noise.

- Let
$$E_0 = \left\{ \frac{1}{n} \sum\nolimits_m X_m^2(1) > P \right\}$$
and
$$E_i = \left\{ \left(X_1^n(i),\, X_1^n(1) + Z_1^n \right) \in A_\varepsilon^{(n)} \right\}$$
then
$$\begin{aligned} \pi_n &= P(E_0 \cup E_1^c \cup E_2 \cup \cdots \cup E_M) \\ &\leq P(E_0) + P(E_1^c) + \sum\nolimits_{i=2}^{M} P(E_i) \end{aligned}$$
- Fix a small $\varepsilon > 0$:
  - Law of large numbers: $P(E_0) < \varepsilon$ for sufficiently large n, since $\frac{1}{n}\sum_{m=1}^{n} X_m^2(1) \to P - \varepsilon$ a.s.
  - Joint AEP: $P(E_1^c) < \varepsilon$ for sufficiently large n.
  - Definition of joint typicality:
$$P(E_i) \leq 2^{-n(I(X;Y)-3\varepsilon)}, \quad i = 2, \dots, M.$$

- For sufficiently large n, we thus get
$$\pi_n \leq 2\varepsilon + 2^{-n(I(X;Y)-R-3\varepsilon)}$$
with
$$I(X;Y) = \iint f(y|x)f(x) \log \frac{f(y|x)}{\int f(y|x)f(x)dx} dxdy$$
where $f(x) = \mathcal{N}(0, P - \varepsilon)$ generated the codebook and $f(y|x)$ is given by the channel. Since $f(y|x) = \mathcal{N}(x, \sigma^2)$
$$I(X;Y) = \frac{1}{2} \log \left(1 + \frac{P - \varepsilon}{\sigma^2}\right)$$

- As long as $R < I(X;Y) - 3\varepsilon$, $\pi_n \to 0$ as $n \to \infty \implies$ exists at least one code, say $\mathcal{C}_n^*$, with $P_e^n \to 0$ for $R < I(X;Y) - 3\varepsilon$
- Throw away worst half of the codewords in $\mathcal{C}_n^*$ to strengthen from $P_e^{(n)}$ to $\lambda^{(n)}$ (the worst half has the codewords that do not satisfy the power constraint, i.e., $\lambda_i = 1$) $\implies$ all

$$R < \frac{1}{2} \log \left(1 + \frac{P - \varepsilon}{\sigma^2}\right)$$

are achievable for all $\varepsilon > 0 \implies$

$$C \geq \frac{1}{2} \log \left(1 + \frac{P}{\sigma^2}\right)$$

Memoryless Gaussian Channel: An Upper Bound for C

- Consider any sequence of codes that can achieve the rate R, that is $\lambda^{(n)} \to 0$ and $\frac{1}{n}\sum_{m=1}^{n} x_m^2(i) \leq P$, $\forall n$.
- Assume $\omega \in \mathcal{I}_M$ equally likely. Fano $\implies$

$$R \leq \frac{1}{n} \sum_{m=1}^{n} I(x_m(\omega); Y_m) + \epsilon_n$$

where $\epsilon_n = \frac{1}{n} + RP_e^{(n)} \to 0$ as $n \to \infty$, and where

$$\begin{aligned} I(x_m(\omega); Y_m) &= h(Y_m) - h(Z_m) \\ &= h(Y_m) - \frac{1}{2} \log 2\pi e \sigma^2 \end{aligned}$$

- Since $E[Y_m^2] = P_m + \sigma^2$ where $P_m = \frac{1}{M}\sum_{i=1}^{M} x_m^2(i)$ we get

$$h(Y_m) \leq \frac{1}{2}\log 2\pi e(\sigma^2 + P_m)$$

and hence $I(x_m(\omega); Y_m) \leq \frac{1}{2}\log(1 + \frac{P_m}{\sigma^2})$. Thus,

$$\begin{aligned} R &\leq \frac{1}{n}\sum_{m=1}^{n}\frac{1}{2}\log\left(1 + \frac{P_m}{\sigma^2}\right) + \epsilon_n \\ &\leq \frac{1}{2}\log\left(1 + \frac{\frac{1}{n}\sum_m P_m}{\sigma^2}\right) + \epsilon_n \\ &\leq \frac{1}{2}\log\left(1 + \frac{P}{\sigma^2}\right) + \epsilon_n \ \to \ \frac{1}{2}\log\left(1 + \frac{P}{\sigma^2}\right) \quad \text{as } n \to \infty \end{aligned}$$

for all achievable R, due to Jensen's inequality and the power constraint $\implies$

$$C \leq \frac{1}{2}\log\left(1 + \frac{P}{\sigma^2}\right)$$

The Coding Theorem for a Memoryless Gaussian Channel

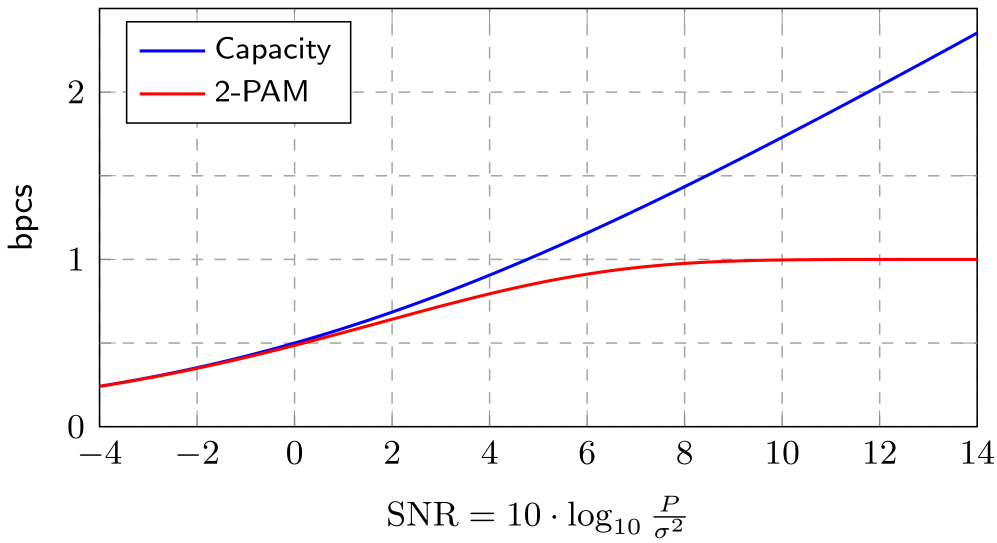
Theorem

A memoryless Gaussian channel with noise variance σ^2 and power constraint P has capacity

$$C = \frac{1}{2}\log\left(1 + \frac{P}{\sigma^2}\right)$$

That is, all rates $R < C$ and no rates $R > C$ are achievable.

AWGN Capacity vs. Simple Binary Scheme



Simple binary scheme:

- Two possible input values: $X \in \{-\sqrt{P}, \sqrt{P}\}$
- Continuous output (soft decoder): $Y = X + Z \in \mathbb{R}$
- Rate: $I(X;Y) = h(X+Z) - h(Z)$

Parallel Gaussian Channels

- Consider the scenario where there are K available channels

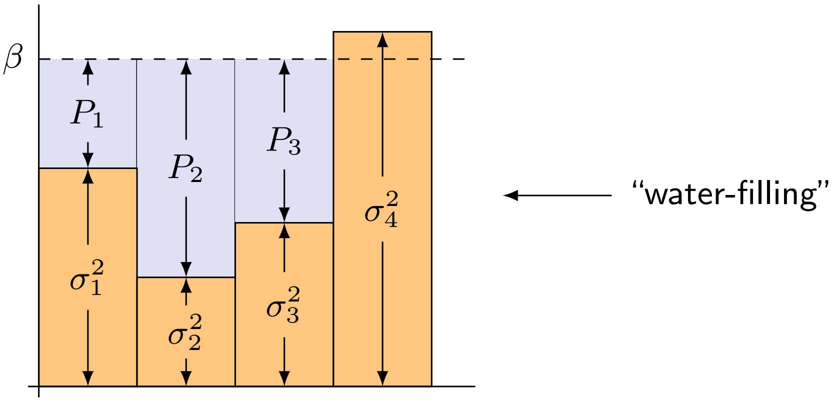
$$Y_k = X_k + Z_k, \quad k = 1, \dots K,$$

  that can be used simultaneously. Here we assume that Z_k are zero-mean independent Gaussian, with $E[Z_k^2] = \sigma_k^2$.

- The capacity of the equivalent "super-channel" is obtained by signaling independently with powers $P_k = E[X_k^2]$ determined as

$$P_k = \begin{cases} \beta - \sigma_k^2, & \sigma_k^2 < \beta \\ 0, & \sigma_k^2 \geq \beta \end{cases}$$

  where β is chosen such that $\sum_k P_k = P$, the total transmit power.



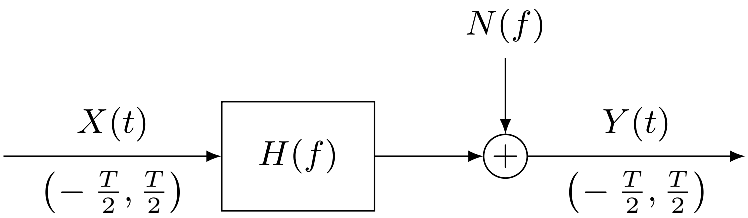
- The total capacity is then the sum of the capacities of the individual sub-channels

$$C = \frac{1}{2}\sum_{k=1}^{K} \log\left(1 + \frac{P_k}{\sigma_k^2}\right),$$

where P_k was defined previously.

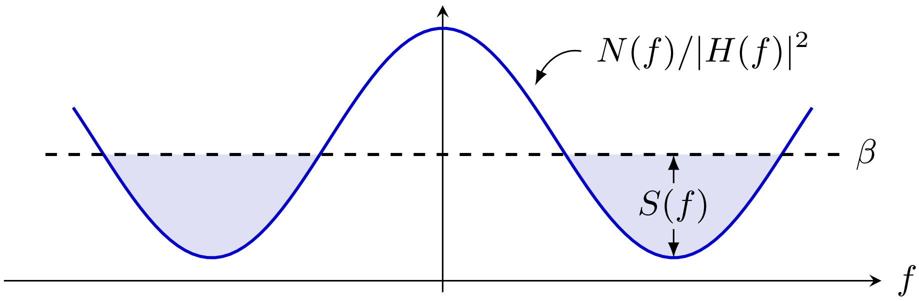
- All channels "linearly related" to a set of parallel Gaussian channels can be handled using the above results!

Gaussian Waveform Channel



- Linear-filter waveform channel with Gaussian noise
  - Independent Gaussian noise with spectral density $N(f)$
  - Linear filter $H(f)$
  - Input and output confined to time interval $(-\frac{T}{2}, \frac{T}{2})$
  - Power constraint

$$\frac{1}{T}\int_{-T/2}^{T/2} E[X^2(t)]dt \leq P$$



- This channel has capacity (in bits per second) given by

$$C = \frac{1}{2}\int_{\mathcal{F}(\beta)} \log \frac{|H(f)|^2 \cdot \beta}{N(f)} df$$
$$P = \int_{\mathcal{F}(\beta)} \left[\beta - \frac{N(f)}{|H(f)|^2}\right] df$$

  where

$$\mathcal{F}(\beta) = \left\{f : N(f) \cdot |H(f)|^{-2} \le \beta\right\}$$

  and where different possible pairs (C, P) correspond to different values of $\beta \in (0, \infty)$.

- That is, there exists a code (set of M possible input waveforms) such that arbitrarily low error probability is possible as long as

$$R = \frac{\log M}{T} < C$$

  and as $T \to \infty$. For $R > C$ the error probability is > 0.

- The famous special case of a band-limited AWGN channel:
  - Perfect low-pass filter of bandwidth W

$$H(f) = \begin{cases} 1 & |f| \le W \\ 0 & |f| > W \end{cases}$$

  - White Gaussian noise, with $N(f) = N_0/2$
  - The capacity of this channel is (Shannon '48):

$$C = W \cdot \log\left(1 + \frac{P}{WN_0}\right) \quad \text{[bits per second]}$$