I232 Information Theory
Chapter 9: Differential Entropy and the Gaussian Channel

Brian Kurkoski

Japan Advanced Institute of Science and Technology

2022 May

Do Pop Quiz 9 on the LMS

Now, Consider Continuous Random Variables

- **Discrete:** Sample set is discrete, e.g. $\mathcal{X} = \{1, 2, \ldots, |\mathcal{X}|\}$, so

$$\sum_{x \in \mathcal{X}} p(x) = 1.$$

- **Continuous:** Sample set is continuous: $\mathcal{X} \subseteq \mathbb{R}$ (real numbers) so:

$$\int_{\mathcal{X}} p(x) dx = 1.$$

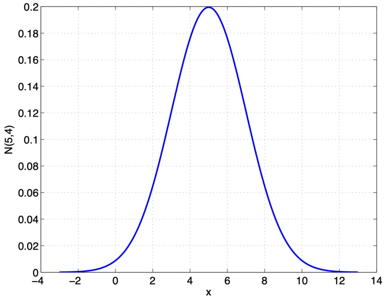
and $p(x) \geq 0$ in both cases.

Gaussian Distribution

Gaussian probability distribution with mean m and variance σ^2:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

denoted as $\mathcal{N}(m, \sigma^2)$.



Contents

9.1 Differential Entropy

Definition (Definition 9.1)

The *differential entropy*[1] $h(\mathsf{X})$ of a continuous random variable X with probability distribution $p_{\mathsf{X}}(x)$ is:

$$
\begin{aligned}
h(\mathsf{X}) &= -\int_{\mathcal{X}} p_{\mathsf{X}}(x) \log p_{\mathsf{X}}(x) dx \quad \text{in bits} \\
h(\mathsf{X}) &= -\int_{\mathcal{X}} p_{\mathsf{X}}(x) \ln p_{\mathsf{X}}(x) dx \quad \text{in nats}
\end{aligned}
$$

[1] Usually, $h(\mathsf{X})$ is used for differential entropy. The lecture notes use $H(\mathsf{X})$ for differential entropy, and these slides mostly use $h(\mathsf{X})$, but may be inconsistent in some places. Alternatively, sometimes $h(f)$ or $H(f)$ is written, if the probability distribution is $f(x)$.

Differential Entropy of Uniform Distributed X

For X $\sim$ uniform distribution, i.e., $p(x) = 1/a,\ 0 \leq x \leq a$:

$$h(X) = -\int_0^a \frac{1}{a} \log \frac{1}{a} dx = \log a.$$

Differential Entropy of Gaussian Distributed X

For X $\sim \mathcal{N}(0, \sigma^2)$, i.e., $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$:

$$h(X) = \tfrac{1}{2} \log 2\pi e \sigma^2.$$

Multivariable Differential Entropy

Definition (Definition 9.3)

The *differential entropy* of $\mathsf{X}_1, \mathsf{X}_2, \ldots, \mathsf{X}_n$ with joint density $p_{\mathbf{X}}(\mathbf{x})$

$$H(\mathsf{X}_1, \ldots, \mathsf{X}_n) = -\int \cdots \int p_{\mathbf{X}}(\mathbf{x}) \log p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}$$

Multivariable Differential Entropy

Definition (Definition 9.3)

The *differential entropy* of $X_1, X_2, \ldots, X_n$ with joint density $p_{\mathbf{X}}(\mathbf{x})$

$$H(X_1, \ldots, X_n) = -\int \cdots \int p_{\mathbf{X}}(\mathbf{x}) \log p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}$$

Definition (Definition 9.4)

If X, Y are jointly distributed as $p_{XY}(x, y)$, then *conditional differential entropy* is:

$$H(X|Y) = -\int \int p_{XY}(x, y) \log p_{X|Y}(x|y) dx dy$$

Properties of Differential Entropy

For a single variable X:

- For a constant c, $h(\mathsf{X} + c) = h(\mathsf{X})$.
- For a constant a, $h(a\mathsf{X}) = h(\mathsf{X}) + \log |a|$.

For multivariabfle differential entropy with $\boldsymbol{\mathsf{X}} = \mathsf{X}_1, \mathsf{X}_2, \ldots, \mathsf{X}_n$:

- For a constant $\mathbf{c}$, $h(\mathbf{X} + \mathbf{c}) = \mathbf{h}(\mathbf{X})$.
- For an invertible matrix $\mathbf{A}$, $h(\mathbf{AX}) = h(\mathbf{X}) + \log |\det(\mathbf{A})|$.

★1

Kullback-Leibler Divergence

Definition (Definition 9.5)

The *Kullback-Leibler divergence* $D\big(f(x)||g(x)\big)$ between distributions $f(x)$ and $g(x)$ is given by:

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

The relative entropy is a "distance" between the distance between a true density $f(x)$ and another density $g(x)$.

Proposition

$D(f||g) \geq 0$ with equality if and only if $f = g$. ★2

Mutual Information

Definition (Definition 9.6)

Let X and Y be continuous jointly distributed random variables. Then the *mutual information* $I(\mathsf{X};\mathsf{Y})$ between X and Y is:

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

★3

Proposition (Proposition 9.7)

$I(\mathsf{X};\mathsf{Y}) \geq 0$ with equality if and only if X and Y are independent. ★poll

KL Divergence and Mutual Information

Definition: The *Kullback-Leibler divergence* $D(f(x)||g(x))$ between distributions $f(x)$ and $g(x)$ is given by

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

Definition: Let X and Y be continuous jointly distributed random variables. The *mutual information* $I(X;Y)$ between X and Y is

$$I(X;Y) = h(X) - h(X|Y).$$

Outline

9.2 Differential Entropy of Gaussians

poll

Proposition (Proposition 9.9)

Among random variables with mean 0 and variance σ^2, the Gaussian random variable $\mathcal{N}(0, \sigma^2)$ maximizes differential entropy.

★4

Multivariate Gaussian

The *multivariate Gaussian distribution*[2] of an n-dimensional random vector $\mathbf{X} = [\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n}]^{\mathrm{T}}$ is:

$$p(\mathbf{x}) = \frac{\mathbf{1}}{(\mathbf{2}\pi)^{\mathbf{1/2}}|\det \mathbf{K}|^{\mathbf{1/2}}} \exp\left(-\frac{\mathbf{1}}{\mathbf{2}}(\mathbf{x}-\mathbf{m})^{\mathrm{T}}\mathbf{K}^{-\mathbf{1}}(\mathbf{x}-\mathbf{m})\right)$$

with mean vector $\mathbf{m}$ and covariance matrix $\mathbf{K}$:

$$\mathbf{K} = \mathrm{E}\big[(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^{\mathrm{T}}\big].$$

Example $n = 2$: X, Y are bivariate Gaussians with:

$$\mathbf{K} = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix} \text{ and } \mathbf{m} = \begin{bmatrix} m_x \\ m_y \end{bmatrix}$$

where m_x, σ_x^2 are the mean, variance of X. Likewise m_y, σ_y^2 are the mean, variance of Y. $-1 \leq \rho \leq 1$ is the correlation coefficient

[2] The Gaussian distribution is also called the normal distribution.

Example of Bivariate Gaussian

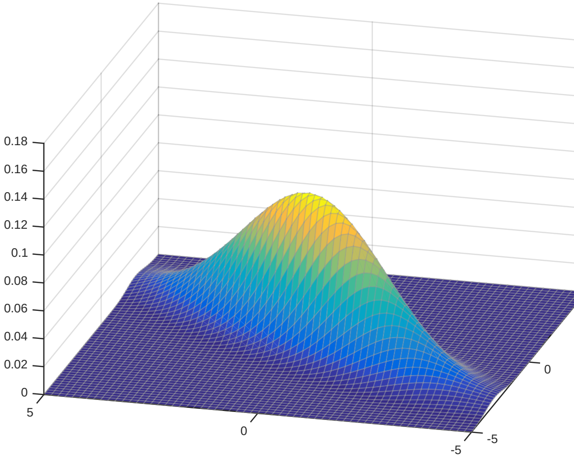
$p_{\mathsf{XY}}(x, y) =$

$$\frac{1}{(2\pi)^{1/2}|\det \mathbf{K}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{m})^{\mathrm{T}}\mathbf{K}^{-1}(\mathbf{x}-\mathbf{m})\right)$$

$$\mathbf{K} = \begin{bmatrix} \sigma_x^2 & \sigma_x\sigma_y\rho \\ \sigma_x\sigma_y\rho & \sigma_y^2 \end{bmatrix} \text{ and } \mathbf{m} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

where

$$\sigma_x^2 = 4, \sigma_y^2 = 2$$
$$\rho = 0.3$$



—

In general, what does $\rho = 0$ mean? $\rho = +1$ and $\rho = -1$?

Entropy of Multivariate Gaussian Distribution

Proposition: Let $\mathbf{X} \sim \mathcal{N}(\mathbf{m}, \mathbf{K})$ have a multivariate Gaussian distribution. Then:

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

Example

Let $(\mathsf{X}, \mathsf{Y}) \sim \mathcal{N}(0, \mathbf{K})$ where:

$$\mathbf{K} = \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix}$$

Find $h(\mathsf{X}), h(\mathsf{Y}), h(\mathsf{X}, \mathsf{Y})$ and $I(\mathsf{X}; \mathsf{Y})$. ★5

Differential vs Discrete: Similarities

	discrete	continuous
Entropy $H(\mathsf{X})$	$-\sum_{x\in\mathcal{X}} p_{\mathsf{X}}(x)\log p_{\mathsf{X}}(x)$	$-\int_{\mathcal{X}} p_{\mathsf{X}}(x)\log p_{\mathsf{X}}(x)dx$
Cond. Entropy $H(\mathsf{X}\vert\mathsf{Y})$	$-\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{\mathsf{XY}}(x,y)\log p_{\mathsf{X}\vert\mathsf{Y}}(x\vert y)$	$-\int_{\mathcal{X}}\int_{\mathcal{Y}} p_{\mathsf{XY}}(x,y)\log p_{\mathsf{X}\vert\mathsf{Y}}(x\vert y)dydx$
KL Divergence $D\big(f(x)\vert\vert g(x)\big)$	$\sum_{x\in\mathcal{X}} f(x)\log\frac{f(x)}{g(x)}$	$\int_{\mathcal{X}} f(x)\log\frac{f(x)}{g(x)}dx$
Mutual Information $I(\mathsf{X};\mathsf{Y})$	$\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{\mathsf{XY}}(x,y)\log\frac{p_{\mathsf{XY}}(x,y)}{p_{\mathsf{X}}(x)p_{\mathsf{Y}}(y)}$	$\int_{\mathcal{X}}\int_{\mathcal{Y}} p_{\mathsf{XY}}(x,y)\log\frac{p_{\mathsf{XY}}(x,y)}{p_{\mathsf{X}}(x)p_{\mathsf{Y}}(y)}dxdy$
Mutual Information	$I(\mathsf{X};\mathsf{Y}) = D\big(p_{\mathsf{XY}}(x,y)\vert\vert p_{\mathsf{X}}(x)p_{\mathsf{Y}}(y)\big)$	
Non-negativity	$D\big(p_{\mathsf{XY}}(x,y)\vert\vert p_{\mathsf{X}}(x)p_{\mathsf{Y}}(y)\big) \geq 0 \Rightarrow I(\mathsf{X};\mathsf{Y}) \geq 0$	
Conditioning reduces entropy	$H(\mathsf{X}\vert\mathsf{Y}) \leq H(\mathsf{X})$	
Chain rule	$H(\mathsf{X},\mathsf{Y}) = H(\mathsf{X}\vert\mathsf{Y}) + H(\mathsf{Y})$	

Differential vs Discrete: Differences

	discrete	continuous
Non-negativity of $H(\mathsf{X})$?	Yes: $H(\mathsf{X}) \geq 0$	$\boxed{\text{No: } H(\mathsf{X}) < 0 \text{ possible}}$
Entropy maximizing $p_{\mathsf{X}}(x)$	uniform $p_{\mathsf{X}}(x) = \frac{1}{\lvert\mathcal{X}\rvert}$	Gaussian $\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-m)^2}{2\sigma^2})$
Maximum value of $H(\mathsf{X})$?	$H(\mathsf{X}) \leq \log \lvert\mathcal{X}\rvert$	$H(\mathsf{X}) \leq \log 2\pi e\sigma^2$
Shift by constant c	$H(\mathsf{X}+c) = H(\mathsf{X})$	
Multiply by constant a	$H(a\mathsf{X}) = H(\mathsf{X})$	$\boxed{H(a\mathsf{X}) = H(\mathsf{X}) + \log \lvert a\rvert}$

Outline

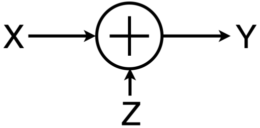
9.3 Coding for the Gaussian Channel

This section considers:

- Gaussian channel model, also called additive white Gaussian noise (AWGN) channel
  - Is a good model for various communication systems.
  - Has a new restriction, a **power constraint**
- Communication system for the AWGN channel,
  - Particularly the code satisfying the power constraint.

9.3.1 Gaussian Channel Model

The additive-white Gaussian noise (AWGN) channel model is:



X, Y and Z are continuous random variables.

- **Input:** $\mathsf{X} \sim p_{\mathsf{X}}(x)$, Transmit power constraint: $\frac{1}{n}\sum_{i=1}^{n} x_i^2 \le P$
- **Noise:** $\mathsf{Z} \sim \mathcal{N}(0, N)$, Gaussian with noise power N
- **Output:** $\mathsf{Y} = \mathsf{X} + \mathsf{Z}$

Signal-to-Noise Ratio

The power of a signal is the square of its value.

Signal power P is the channel input power constraint:

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

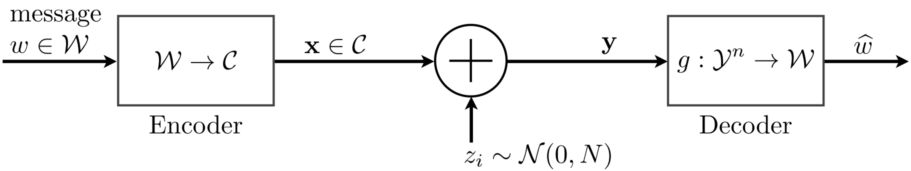
Noise power N is the variance of the zero-mean Gaussian noise:

$$N = \int z^2 p_{\mathsf{Z}}(z)dz = E[\mathsf{Z}^2]$$

Then, the *signal-to-noise* ratio is:

$$\mathrm{SNR} = \frac{P}{N}$$

9.3.2 Communication System Model



- Encoder: encodes message w to codeword $\mathbf{x}$:

$$\mathbf{x} = [x_1, x_2, \ldots, x_n]$$

- Channel: AWGN channel with noise power N, $\mathbf{z} = (z_1, \ldots, z_n)$, with output $\mathbf{y}$:

$$\mathbf{y} = \mathbf{x} + \mathbf{z}$$

- Decoder: decodes $\mathbf{y} = (y_1, \ldots, y_n)$ to estimated message $\widehat{w}$

Codebook with Power Constraint

The codebook is M points which satisfy the power constraint.

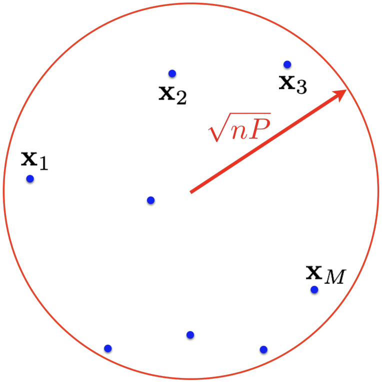
Definition

An (M, n) *Gaussian channel code* $\mathcal{C}$ with power constraint P consists of a message index set $\{1, 2, \ldots, M\}$, and codebook

$$\mathcal{C} = \{\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(M)\}$$

with $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, and each $\mathbf{x}(w)$ satisfies the power constraint P:

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



Encoder and Decoder

As in the DMC case, for the AWGN *encoding function* maps messages to codewords:

$$\{1, 2, \ldots, M\} \rightarrow \mathcal{C}$$

The *code rate* R is:

$$R = \frac{1}{n} \log M$$

The *decoding function* is a deterministic rule g:

$$\widehat{\mathsf{W}} = g(\mathsf{Y}^n).$$

If $\widehat{\mathsf{W}} = \mathsf{W}$, then there is no error. Three types of errors:

conditional probability of error	λ_w	$= \Pr\left(\widehat{\mathsf{W}} \neq w \vert \mathsf{W} = w\right)$
maximum probability of error	$\lambda^{(n)}$	$= \max_w \Pr\left(\widehat{\mathsf{W}} \neq w \vert \mathsf{W} = w\right)$
average probability of error	$P_e^{(n)}$	$= \frac{1}{M} \sum_{w \in \mathcal{W}} \Pr\left(\widehat{\mathsf{W}} \neq w \vert \mathsf{W} = w\right)$

Examples with $n = 1$

Example

Consider the $n = 1$ code with $\mathcal{C} = \{-3, -1, +1, +3\}$. Find the rate R. What power constraint does this code satisfy? ★6

Example

Consider the code with $\mathcal{C} = \{-1, +1\}$. This is transmitted over an AWGN channel with noise power N. The decoder decides -1 was transmitted if the received symbol y is less than 0, and decides $+1$ was transmitted if the received symbol $y > 0$. What is the probability of decoder error?

Outline

9.4 Gaussian Channel Capacity

Definition: The capacity C of the Gaussian channel with power constraint P:

$$C = \max_{p(x),\ \mathrm{E}[X^2] \le P} I(X;Y).$$

An **optimal** $p^*(x)$ is called the *capacity-achieving input distribution:*

$$p^*(x) = \arg \max_{p(x),\ \mathrm{E}[X^2] \le P} I(X;Y).$$

Gaussian Channel Capacity

Proposition: *Gaussian channel capacity* The capacity of a Gaussian channel with power contraint P and noise variance N is:

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

The capacity-achieving input distribution is $p^*(x) = \mathcal{N}(0, P)$.

The Shannon Statue

The channel capacity formula is so important, it is written on piece of paper held by Shannon's statue.

Shannon statue at University of California San Diego

An Upper Bound

We can easily prove an upper bound on the capacity:

$$C = \max_{p_{\mathsf{X}}(x): E\mathsf{X}^2 \leq P} I(\mathsf{X}; \mathsf{Y}) \leq \frac{1}{2} \log\big(1 + \frac{P}{N}\big).$$

If we choose $\mathsf{X} \sim \mathcal{N}(0, P)$, then this agrees with the Gaussian channel capacity. ★7

One proof of the channel capacity theorem uses the AEP.

9.4.2 Typical Set and AEP

Consider the typical set and AEP for continuous-valued random variables.

Definition: For $\epsilon > 0$ and any n, the typical set $\mathcal{T}_\epsilon^{(n)}$ is:

$$\mathcal{T}_\epsilon^{(n)} = \left\{ (x_1, \ldots, x_n) \in \mathcal{X}^n : \left| -\frac{1}{n} \log p(x_1, \ldots, x_n) - h(X) \right| \leq \epsilon \right\}.$$

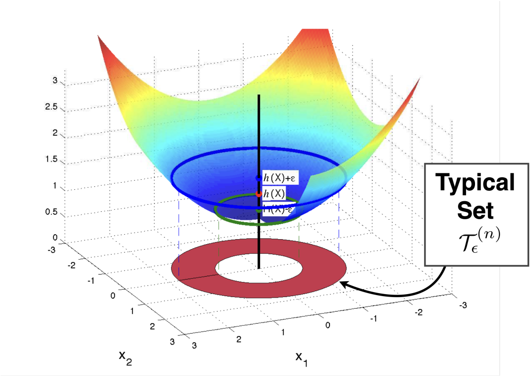
Note: $\mathcal{T}_\epsilon^{(n)} \subseteq \mathbb{R}^n$ is a region of space.

What does the typical set look like?

Consider $n = 2$ and $\mathcal{N}(0,1)$. The typical sequences are those:

$$h(X) - \epsilon \leq -\tfrac{1}{2} \log p(x_1, x_2) \leq h(X) + \epsilon$$

$$h(X) - \epsilon \leq \tfrac{x_1^2}{4} + \tfrac{x_2^2}{4} + \tfrac{1}{2} \ln 2\pi \leq h(X) + \epsilon$$



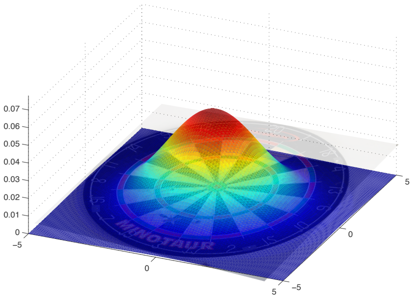
AEP for Continuous Random Variables

Theorem—Asymptotic Equipartition Property: if $X_1, \ldots, X_n$ are IID random variables with probability distribution $p(x)$ and entropy $h(X)$. Then,

$$\lim_{n \to \infty} \Pr[\mathbf{X} \in \mathcal{T}_\epsilon^{(\mathbf{n})}] = \mathbf{1}.$$

Shannon Plays Darts

A dartboard

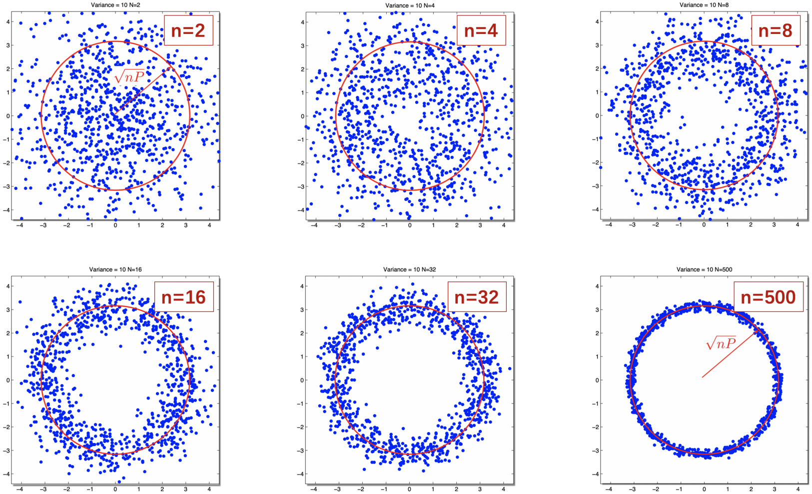


Throwing darts with a Gaussian distribution

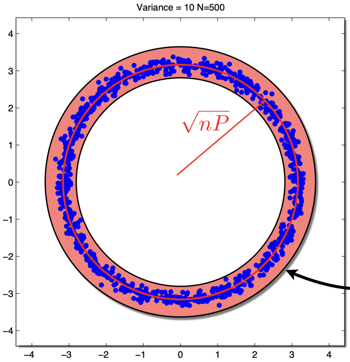
Picture is dimension $n = 2$. For high dimensions, visualize as follows, even if not precise:

- Generate random $(x_1, x_2, \ldots, x_n)$
- Plot point radius: the same
- Plot point angle: same as (x_1, x_2)

Graphical Illustration of IID Gaussian Distributed $\mathbf{X}$ – 1000 Samples



Graphical Illustration of IID Gaussian Distributed $\mathbf{X}$



Most points are in a doughnut-shaped typical set.

On average, it is hard to hit the bulls's eye!

Typical Set

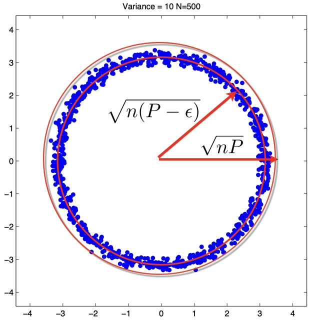
$\mathcal{T}_\epsilon^{(n)}$

9.4.3 Codebook Construction

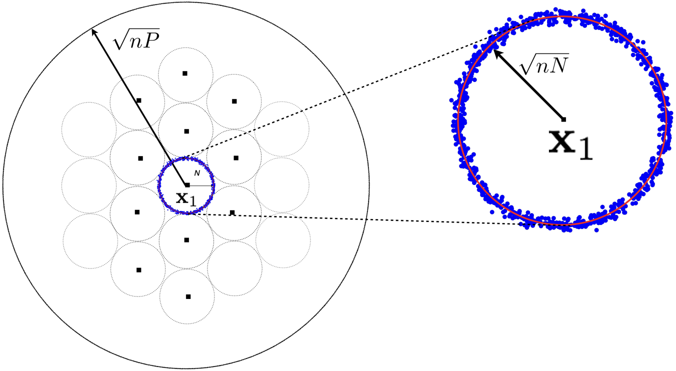
To prove the channel coding theorem, the codebook is constructed using an i.i.d. Gaussian source:

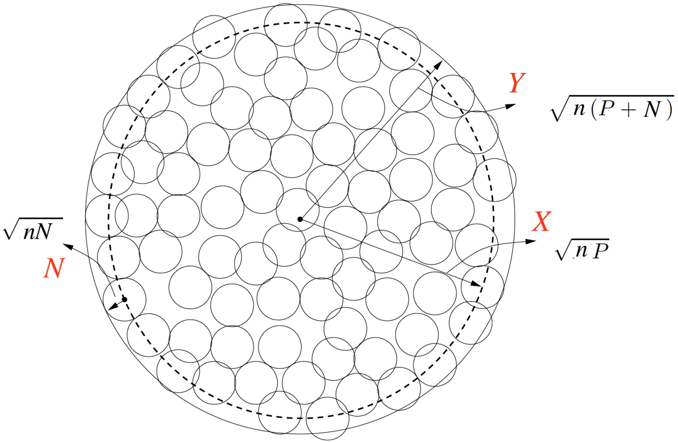
$$p^*_{\mathsf{X}}(x) = \mathcal{N}(0, P - \epsilon)$$

Codewords are near the surface of a sphere with radius $\sqrt{n(P-\epsilon)}$.



Noise is centered on codeword





Intuition: Correct decoding $\Leftrightarrow$ No overlap in small circles

$$M = \mathrm{vol}(Y)/\mathrm{vol}(N) = \left(\sqrt{n(P+N)}\right)^n / \left(\sqrt{nN}\right)^n = \left(1 + P/N\right)^{n/2}$$

$$\Rightarrow R = \frac{1}{n}\log M = \frac{1}{2}\log(1 + \frac{P}{N})$$

Gaussian Channel Coding Theorem

Proof outline — direct part

1. Random codebook constructed $\sim \mathcal{N}(0, P - \epsilon)$
2. Decoding is "jointly typical decoding"
3. Error occurs if transmitted $\mathbf{x}$ and received $\mathbf{y}$ are not jointly typical
4. Upper bound the probability of error $\to 0$

We omit the proof as it is similar as the discrete case.

Example — Binary-Input AWGN Channel Capacity

The binary-input additive white Gaussian noise (AWGN) channel has input $\mathsf{X} \in \{+1, -1\}$ with uniform input distribution $p_{\mathsf{X}}(x) = [\frac{1}{2}, \frac{1}{2}]$, AWGN Z distributed as $\mathcal{N}(0, \sigma^2)$ and channel output is Y:

$$\mathsf{Y} = \mathsf{Z} + \mathsf{X}.$$

The capacity can be achieved using $p^*_{\mathsf{X}}(x) = [\frac{1}{2}, \frac{1}{2}]$. Find the capacity C for this channel.

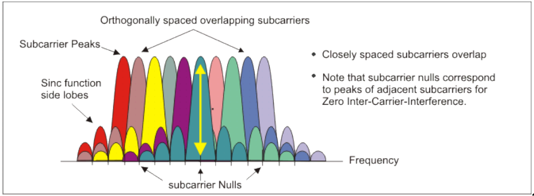
The capacity of this important channel does not have a closed form solution.

Outline

9.5 Parallel Gaussian Channels

In this section, consider parallel Gaussian channels.

As motivation, consider orthogonal frequency division multiplexing (OFDM) which is fundamental to 4G and 5G communications.



[3]

OFDM Signal Frequency Spectra

In order to reduce wireless interference, each users bandwidth is separated into a number of parallel channels.

[3] http://rfmw.em.keysight.com/

k parallel, independent Gaussian channels:

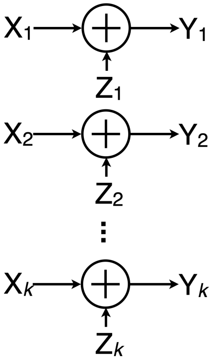
$$Y_j = X_j + Z_j, \quad j = 1, \ldots, k.$$

Each channel has:

$$\begin{aligned} \text{transmit power}: \quad & P_j = \mathrm{E}[X_j^2], \\ \text{IID}: \quad & Z_j \sim \mathcal{N}(0, N_j). \end{aligned}$$

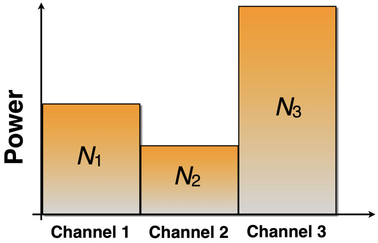
Constraint on the total power:

$$\mathrm{E}[X_1^2] + \cdots + \mathrm{E}[X_k^2] = \sum_i P_i \leq P.$$



We can allocate power to the various channels. Each channel has its own noise N_j. How to allocate?

How to Allocate Power P to Channels?



Capacity of Parallel Gaussian Channels

Goal: Allocate the transmit power among the channels to maximize the parallel channels' information capacity

Definition: The capacity of the parallel Gaussian channel is

$$C = \max_{p(\mathbf{x}): \sum \mathrm{E}[\mathbf{X_i^2}] \leq \mathbf{P}} I(X_1, \ldots, X_k; Y_1, \ldots, Y_k)$$

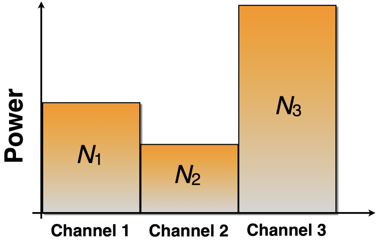
where $p(\mathbf{x})$ is the joint input distribution.

Capacity of Parallel Gaussian Channels

Proposition: *Waterfilling* For parallel Gaussian channels, the information capacity is achieved by $\{X_i \sim \mathcal{N}(0, P_i^*)\}$ with

$$P_i^* = (v - N_i)^\dagger$$

and $\sum P_i^* = P$.

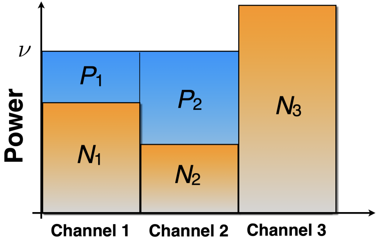


Waterfilling Achieves the Capacity

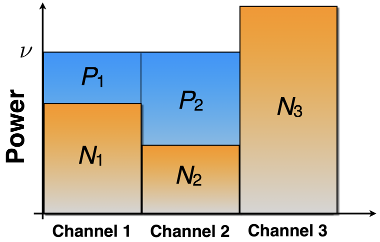
Proposition: *Waterfilling* For parallel Gaussian channels, the information capacity is achieved by $\{X_i \sim \mathcal{N}(0, P_i^*)\}$ with

$$P_i^* = (v - N_i)^\dagger$$

and $\sum P_i^* = P$.



Waterfilling Achieves the Capacity



Intuition: Assume the power is increased ΔP. Then, rate is increased

$$\begin{aligned} \Delta R &= \tfrac{1}{2}\log\left(1+\tfrac{P+\Delta P}{N}\right) - \tfrac{1}{2}\log\left(1+\tfrac{P}{N}\right) \\ &= \tfrac{1}{2}\log\left(1+\tfrac{\Delta P}{P+N}\right) \end{aligned}$$

Larger rate gain for channels with lower $P_i + N_i \Rightarrow$ Waterfilling!

Summary

- **Differential entropy** is the entropy of **continuous** random variables
- Many properties similar to discrete case but:
  - Differential entropy can be **negative**
  - Maximum value given by a **Gaussian distribution**
  - **Multiply** by a constant **changes entropy**
- The **AWGN** channel capacity is $C = 0.5 \log(1 + P/N)$
- **Parallel** Gaussian channels: **"waterfilling"** achieves the capacity

Class Info

- Next lecture: Monday, May 22. Rate-distortion theory. There will be a pop quiz.
- Tutorial Hours: Monday, May 22 at 13:30. Ask questions about homework.
- Homework 7 and 9 on LMS. Deadline: Monday, May 22 at 18:00