



Information Theory

Agenda



- ❖ Course Schedule
- ❖ Primers on Probability/Random Variable

E-mail

- ❖ Please send me a short e-mail that you are in the class
heungno@gist.ac.kr
- ❖ I will use e-mail for notes and special announcements.

Course Information

- ❖ Class hours: 9:00-10:30 am Monday, Wednesday
- ❖ Lecture room: B203
- ❖ Office hours:
 - 10:30am ~ 12:00am Monday,
 - 10:00am ~ 11:00am Tuesday.
 - Or make an appointment via e-mail.

Grade Distribution

- ❖ Midterm 1 (20%)
- ❖ Homework + Grading (20%)
- ❖ Paper Reading + Presentation (30%)
- ❖ Final exam (30%)

Tentative Schedule

<i>Weekly Course Schedule</i>		
<i>Calendar</i>	<i>Description</i>	<i>*Remarks</i>
<i>1st week</i>	<i>Introduction to Information Theory, Entropy</i>	
<i>2nd week</i>	Entropy, Relative Entropy and Mutual Information	
<i>3rd week</i>	Entropy, Relative Entropy and Mutual Information	
<i>4th week</i>	Asymptotic Equipartition Property	
<i>5th week</i>	Asymptotic Equipartition Property/Entropy Rates of a Stochastic Process	
<i>6th week</i>	Entropy rates of Markove Chain	
<i>7th week</i>	Data compression	Midterm 1
<i>8th week</i>	Channel capacity	
<i>9th week</i>	Channel capacity theorems/forward/reverse	
<i>10th week</i>	Differential entropy	
<i>11th week</i>	Gaussian channel capacity	
<i>12th week</i>	MIMO channel capacity theorem	
<i>13th week</i>	Multiple access channel capacity theorem	
<i>14th week</i>	Slepian Wolf	
<i>15th week</i>	Student Presentation	
<i>16th week</i>	Student Presentation	
	Final Exam	

Homework Policies

- ❖ Discussion and exchange of ideas are strongly encouraged.
- ❖ Must submit your own independent report.
- ❖ On each homework, a reviewer will be assigned (will take turns).
- ❖ The job of each reviewer is to
 - grade homework/project sets,
 - type up the best homework solution,
 - get an approval of the solution manual from me, and
 - distribute the graded homework and solution to the students within a week.

Selection of Journal for class presentation

- ❖ Purpose: See if you can apply the knowledge learned in class to expand the topics of your reading.
 - IEEE Information Theory, IEEE Trans. Comm., IEEE JSAC
- ❖ Procedure
 - Find the area of your interests (you can discuss this with me)
 - IEEE Trans. Info. Theory (years 2006 and later)
 - Use search engines (e.g. IEEEexplore, INSPEC, SCI, ...) to find a paper
 - Bring a print out of paper to me for an approval, by Wed. of 12th week.
 - Read the paper and its references
 - Summarize your understanding of the paper in PPT charts for 20 min. in class presentation
 - PPT charts should be written succinctly
 - Font size > 18
 - Number of pages < 20

Check the Following, at least

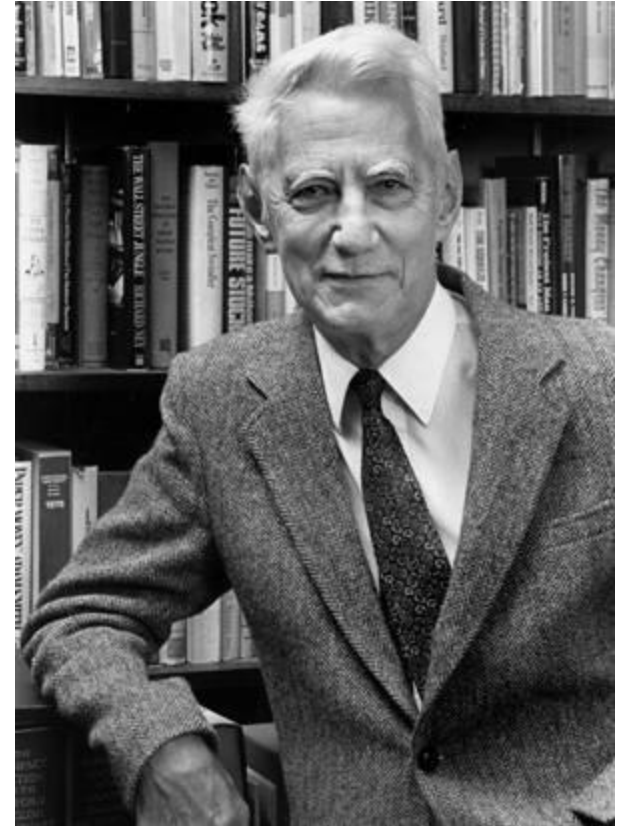
- ❖ What's the main contribution of your paper?
 - The main problem considered in the paper?
 - Give a clear problem statement
 - Who else attempted to solve the problem?
 - What approaches are taken?
 - What are the major differences?
- ❖ Discuss the technical details of the paper
 - What are the technical approaches for main results?
 - Provide insights and examples
- ❖ Discuss the results
 - Metrics for comparison?
 - What's new?
 - What could we do with the new knowledge?

Scope of this course

- ❖ Information Theory, as created by Shannon 1948
- ❖ Plus
- ❖ Modern information theory
 - MIMO channel capacity
 - Network information theory
 - Compressed Sensing

Claude E. Shannon (1916 -- 2001)

- ❖ Math/EE Bachelor from UMich (1936)
- ❖ MSEE and Math Ph.D. from MIT (1940)
- ❖ A landmark paper “Mathematical Theory of Communications” (1948)
 - Founder of Information Theory
 - Fundamental limits on communications
 - Information quantified as a logarithmic measure
- ❖ For more info on him, make a visit to <http://www.bell-labs.com/news/2001/february/26/1.html>



Textbooks

- ❖ Textbook: Thomas & Cover, *Elements of Information Theory*, 2nd Ed., Wiley, 2006
- ❖ Reference-1: Robert Gallager, *Information Theory and Reliable Communication*, John Wiley & Sons, Inc. New York, NY, USA, 1968. ISBN:0471290483
- ❖ Reference-2: Raymond Yeung, *A First Course in Information Theory*, Kluwer Academic/Plenum Publishers, 2002

❖ Now, let's begin by reviewing Probability/Random Variables

Primer on Probability/Random Variables

The 0th Module

Real World Experiments and Mathematical Abstraction

❖ Experiments

- Measurement of voltage across a resistance
- Roll a die

❖ Three entities in the real world experiments

- The set of all possible *outcomes*
- Grouping of the *outcomes* into classes, called *results*
- The *relative frequencies* of occurrences of the *results*

❖ The corresponding mathematical abstractions

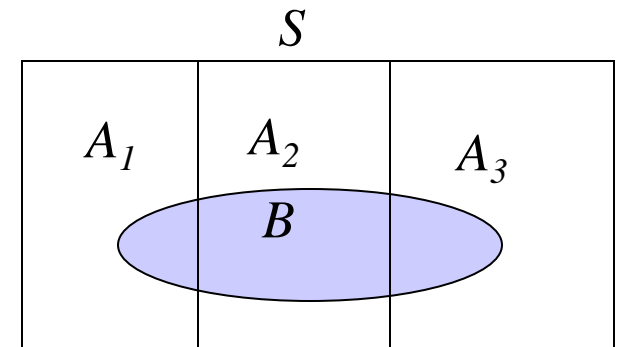
- *The sample space*
- *The set of events*
- *The probability measure* assigned on each of these *events*

Fundamental Definitions in Set Theory

- ❖ A set is a collection of *objects (elements)*.
 - $A = \{v: 0 \leq v \leq 5 \text{ volts}\}$
 - $B_1 = \{1, 2, 3, 4\}, B_2 = \{\text{head, tail}\}$
- ❖ A subset C of A is another set whose elements are also elements of A .
 - $C = \{1, 2\} \subset B_1$
 - We say C belongs to B_1
- ❖ Set operations: Union and Intersection
 - $B_1 \cup B_2 = \{1, 2, 3, 4, \text{head, tail}\}$
 - $B_1 \cap C = \{1, 2\}$ (Sometimes, a shorthand notation, $B_1 C$, is used)
- ❖ The *empty set* or *null set* $\{\emptyset\}$ (or simply \emptyset) is the set having no elements.

Fundamental Definitions in Set Theory

- ❖ Two sets A and B are *mutually exclusive* or *disjoint* if they have no common elements.
 - $A \cap B = AB = \emptyset$
- ❖ A partition U of a set S is a collection of mutually exclusive subsets A_i of S whose union equals S .
 - $S = A_1 \cup A_2 \cup A_3$ and $A_i \cap A_j = \emptyset$ for any $i, j \neq i$
- ❖ In the figure below, $U = [A_1, A_2, A_3]$, and the subset $B = (A_1 \cap B) \cup (A_2 \cap B) \cup (A_3 \cap B)$

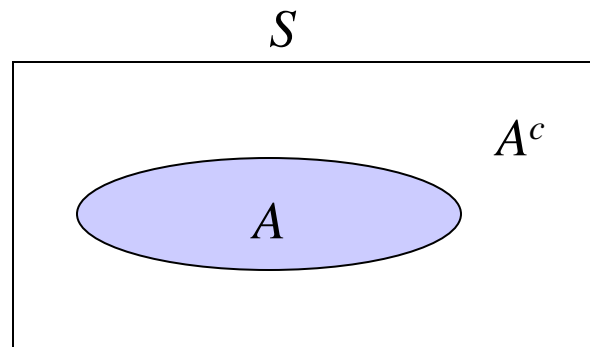


Sample Spaces and *Events*

- ❖ A sample space Ω , which is called *the certain event of a particular experiment*, is the **collection of all experimental outcomes** (objects).
 - An object in Ω is called *a sample point*; is usually denoted by ω .
- ❖ Subsets of a sample space is called *events*.
 - **Grouping of the outcomes** into the subsets
 - A set of sample points
 - $A = \{\omega: \text{some condition(s) on } \omega \text{ is provided here}\}$, the event A is the set of all ω satisfying the condition(s) on ω .
 - An event consisting of a single element is called an *elementary event*.

Complement of an Event

- ❖ We define a complement of an event A as the set of all outcomes of S which are not included in A .
- ❖ We denote $A^c = S \setminus A$.



Examples of *Sample Spaces* and *Events* (Results)

- ❖ Die experiment: $\Omega = \{1, 2, 3, 4, 5, 6\}$
 - $A = \{\omega: \text{odd}\} = \{1, 3, 5\}$
 - $B = \{\omega: \text{even}\} = \{2, 4, 6\}$
- ❖ The closed interval of the real line:
 $\Omega = [0, 1] = \{\omega: 0 \leq \omega \leq 1\}$
 - $A = \{\omega: 0.2 \leq \omega \leq 0.7\}$
- ❖ All time functions $f(t)$, $-\infty < t < \infty$
 - An event may be a set of all time functions whose energy is less than 1.
- ❖ A finite sample space of N elements \rightarrow There are 2^N possible subsets.

Trial

- ❖ A single performance of an experiment is called a *trial*.
- ❖ In each trial we observe a single outcome $a_i \in S$.
- ❖ We say an event A occurs during this trial when A contains a_i .
- ❖ From a single trial, multiple events can occur.
- ❖ Roll a die: $\Omega = \{1, 2, 3, 4, 5, 6\}$.
 - Now, suppose after a trial, an outcome “1” is observed.
 - Then, the events $\{1\}$, $\{1, 3, 5\}$, $\{1, 3\}$, and all the rest $2^5 - 3$ events that contain “1” as an element, it can be said, have occurred.

On the Occurrence of Events In a Trial

- ❖ We say an event $A = \{a_1, a_2, a_3\}$ has occurred in a trial, if any one element of the set, namely, a_1 , a_2 , or a_3 , was the outcome of the trial.
- ❖ The event Ω occurs in every trial.

Probability Measure

- ❖ *An assignment* of a real number from the interval $[0, 1]$ to the *events* defined on Ω .
 - Ex) Fair die: All faces occur equally likely with probability $1/6$.
 - Ex-2) Unfair die: face-1 event occurs with probability $1/3$, the rest 5 faces with $2/15$.
 - Ex-3) You can create and use your own rule which suits your needs the most (your betting rule in Gambling for example).
- ❖ Probability measure $P(A)$ is assigned to a *field* E of subsets (events) of the sample space Ω .
 $P: E \rightarrow [0, 1]$

Relative Frequency vs. Probability Measure

- ❖ The assignment of probability measure to an event A , $P(A)$, may be done in terms of relative frequency of occurrences in N independent trials

$$P(A) = \lim_{n \rightarrow \infty} n_A/N$$

where n_A is the number of occurrence of event A in N trials

- ❖ Ex-1) a coin is tossed 100 times.
 - The event of head occurred 51 times.
 - Then, $P(A) = 51/100$
- ❖ Ex-2) An experienced gambler watches the cards played, and updates his table of probability measures assigned only on the events of his interests and makes bets accordingly

Axiomatic Definition of Probability

- ❖ The assignment of probability to events should follow the three fundamental rules (Kolmogoroff's axioms)
- ❖ 1. $0 \leq P(A) \leq 1$ (The frequency of an event)
- ❖ 2. $P(\Omega) = 1$ (In every trial there is an outcome)
- ❖ 3. If $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$
 - Die: frequency(1 or 2) = frequency(1) + frequency(2),
 $\{1\} \cap \{2\} = \emptyset$
- ❖ In the theory of probability, all conclusions are direct or indirect consequences of these three axioms.
- ❖ These conclusions allow us to predict -- by calculation – the probability of occurrence of observable(or wanting-to-observe) events in real world experiments.
- ❖ Reference: Web-site: <http://www.kolmogorov.com/Kolmogorov.html>

Examples

- ❖ A coin toss experiment: $S = \{h, t\}$
- ❖ Events are the four subsets of S , $\{\emptyset\}$, $\{t\}$, $\{h\}$, $\{h, t\}$.
- ❖ It forms a sigma field.
- ❖ A sigma field is a collection of sets which is closed under the *union* and the *complement* operations.
- ❖ A complement of $\{t\}$ is $\{h\}$ in this example.
- ❖ We will use superscript c to denote complement, i.e., $\{t\}^c = \{h\}$ and $S^c = \{\emptyset\}$.
- ❖ We may assign $P\{t\} = p$ and $P\{h\} = q$, i.e., $p + q = 1$.

Coin Toss Three Times Experiment (1)

- ❖ $S = \{hhh, hht, hth, htt, ttt, tth, tht, thh\}$.
- ❖ Assume a fair coin; then head/tail occurs with equal prob.
- ❖ First, consider the naive case such that all 2^3 possible events are of interest, and then the probability assignment is trivial.
- ❖ Ex) The probability of an event $\{hht, hhh\}$ is
 - $P\{hht, hhh\} = P\{hht\} + P\{hhh\} = 2/8$.

Coin Toss Three Times Experiment (2)

- ❖ Now, consider a non trivial case:
- ❖ Suppose we are interested in the occurrence of an event $A = \{hth, tht\}$ only.
- ❖ Then, we assign probability to only those events in the sigma field formed by A , i.e., $\{A, A^c, \{\emptyset\}, S\}$.
- ❖ Thus, assign $P(A) = 1/4$ (The coin is a fair coin).
- ❖ We note that the probability measure satisfies all the conditions of the Kolmogoroff's axioms.

Conditional Probability

- ❖ Given any two events A and B , the conditional probability $P(A|B)$ of an event A is defined as

$$P(A | B) := P(AB)/P(B)$$

whenever $P(B) \neq 0$.

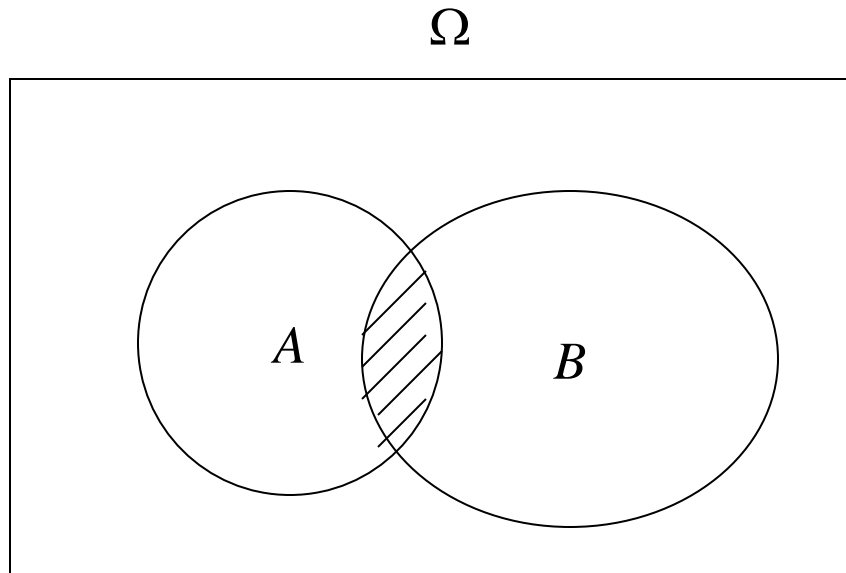
- ❖ $P(A | A) = 1$

- ❖ In the Coin-Toss Three Times experiment, let $A = \{hhh\}$ and $B = \{\text{a head in the first toss}\} = \{hhh, hht, hth, htt\}$

$$P(A | B) = (1/8)/(1/2) = 1/4 .$$

Probability of Joint Event

- ❖ Notation: $P(A, B) = P(AB) = P(A \cap B)$
- ❖ We refer $P(A, B)$ as the probability of a “joint event A and B .”



Probability of Joint Event

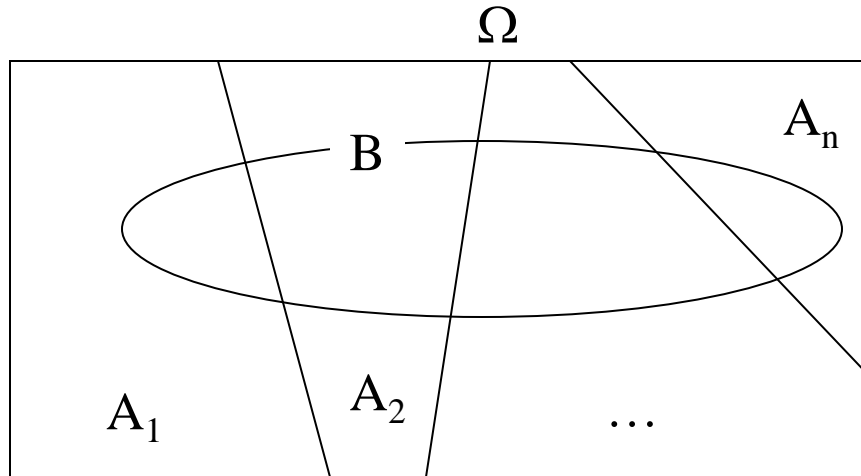
$$\begin{aligned} \blacklozenge P(A, B) &= P(A | B) P(B) \\ &= P(B | A) P(A) \end{aligned}$$

- \blacklozenge A box contains three white balls, w_1 , w_2 , and w_3 and two red balls r_1 and r_2 . We remove two balls in succession. What is the probability that the first removed is white and the second is red?

Independence

- ❖ If $P(A|B) = P(A)$ or $P(B|A) = P(B)$, the two events A and B are, said to be, (statistically) independent with each other.
- ❖ Coin Toss Twice:
 - $\Omega = \{hh, ht, th, tt\}$
 - Suppose we use numbers a and b in $[0, 1]$ with $a + b = 1$ in the following manner:
 - $P\{hh\} = a^2$, $P\{ht\} = P\{th\} = ab$, $P\{tt\} = b^2$
 - Note that the assignment satisfies the axioms: $a^2 + 2ab + b^2 = (a+b)^2 = 1$
 - Now, define two events $A = \{\text{head at the first toss}\}$ and $B = \{\text{head at the second toss}\}$
 - Note $P(A) = aa + ab = a$ and $P(B) = ba + aa = a$
 - $P(A, B) = P\{hh\} = a^2 = P(A) P(B)$
 - Then, we note A and B are mutually independent.

Theorem of Total Probability (Very Important)



- ❖ If $U=[A_1, A_2, \dots, A_n]$ is a partition of Ω and B is an arbitrary event, then

$$\begin{aligned} P(B) &= P(B, A_1) + P(B, A_2) + P(B, A_3) + P(B, A_4) \\ &= P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3) + P(B|A_4)P(A_4) \end{aligned}$$

Bayes' Theorem [Very Important]

- ❖ From the results of the conditional probability and the total probability theorem, we could easily get the following,

$$\begin{aligned} P(A_i|B) &= \frac{P(A_i, B)}{P(B)} \\ &= \frac{P(B|A_i) P(A_i)}{\sum_{k=1}^n P(B|A_k) P(A_k)} \end{aligned}$$

Examples of Bayes' Theorem

- ❖ Box-1 contains a white balls and b black balls. Box-2 contains c white balls and d black balls. One ball is drawn from Box-1 and inserted into Box-2. Then, a ball is drawn from Box-2.
- ❖ What is the probability that a ball drawn from Box-2 is white?

- ❖ What is the probability that the first draw from Box-1 was black, given that a white ball was obtained at the second draw from Box-2

Permutation/Combination

- ❖ Consider a set of N distinct objects
- ❖ *Permutation*: The total number of distinctive arrangements (each in an ordered sequence) of N distinct objects is

$$N!$$

- ❖ The total number of distinctive arrangements when taking K objects out of N distinct objects is

$$N(N-1)(N-2) \dots (N-K+1) = N!/(N-K)!$$

- ❖ *Combination*: The total number of ways to select K objects out of N distinct objects is

$$\binom{N}{K} = \frac{N!}{(N-K)!K!}$$

Bernoulli Trials

- ❖ Observe the occurrence of an event A in each trial
- ❖ The event A occurs with $P(A) = p$ and $P(A^c) = 1 - p = q$
- ❖ Find the probability of a compound event that there are k occurrences of event A in N trials
- ❖ None in $N \dots (1-p)^N$
- ❖ One in $N \dots N p(1-p)^{N-1}$
- ❖ Two in $N \dots \binom{N}{2} p^2(1-p)^{N-2}$
- ...
- ❖ In general,

$$\begin{aligned} P\{A \text{ occurs } k \text{ times in } N \text{ trials}\} \\ = \binom{N}{k} p^k (1-p)^{N-k} \end{aligned}$$

Random Variable and Processes

❖ A signal is a function of time

ex) $y(t) = \sin(2\pi f_c t)$, this is a deterministic signal

❖ A random signal: the value of the signal at a fixed time t is a random variable

ex) $y(t) = \sin(2\pi f_c t + \theta)$, $0 \leq t \leq T$

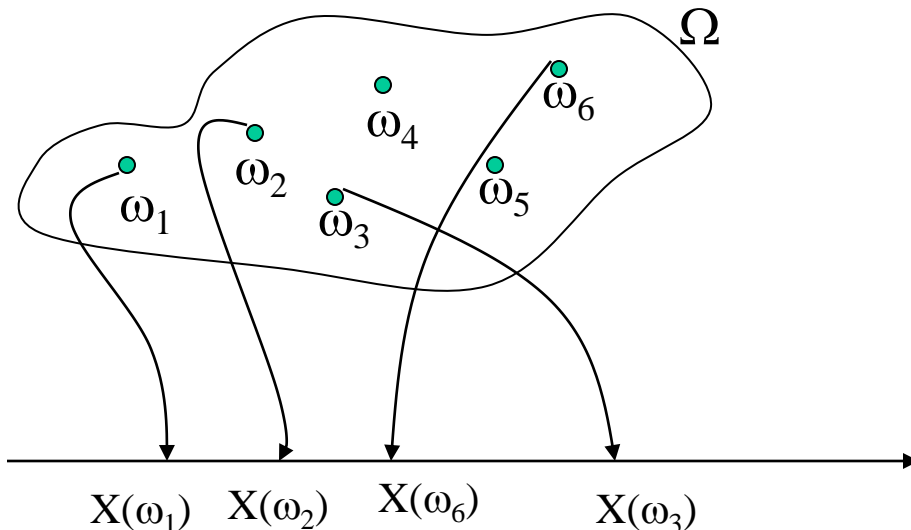
where θ is +180 degree with probability 1/2 or -180 degree with prob. 1/2

❖ A random process $y(t)$ is a collection of different random variables at each time t

– Stochastic processes

Random Variable

- ❖ A function $X: \Omega \rightarrow \mathbb{R}$ (Domain is Ω , range is \mathbb{R})
 - Given any ω , the function specifies a finite real number $X(\omega)$



A random variable is a function whose domain is Ω , the range of this function is usually a real line (Real-valued random variable). Also, it has a probability distribution $\Pr\{X \leq x\}$ associated with it.

Motivation for RV

- ❖ It may be easier to deal with numbers, instead of abstract objects.

Events Described by Random Variables

- ❖ We know now that we assign probability to the field of subsets of Ω .
- ❖ Note that with the use of a random variable, the subsets of *range* space are associated with the subsets of Ω .
- ❖ Thus, events defined on the outcomes of experiments can be described by the subsets of the range space of the function.

Examples of Random Variables

❖ Roll a die experiment

- 6 outcomes, $\Omega = \{f_1, f_2, f_3, \dots, f_6\}$
- We may define a random variable X_1 which has the following rule
 $X_1(f_1) = 10, X_1(f_2) = 20, X_1(f_3) = 30, X_1(f_4) = 40, X_1(f_5) = 50,$ and
 $X_1(f_6) = 60$
- We may also define a random variable X_2 which uses the following rule
 $X_2(f_1) = -1, X_2(f_2) = -2, X_2(f_3) = -3, X_2(f_4) = +3, X_2(f_5) = +2, \dots,$ and
 $X_2(f_6) = +1$
- It's up to the designer to choose a map for convenience

Examples of Random Variables (2)

- ❖ According to the r.v.s X_1 and X_2 , we can say the following:
- ❖ A subset $\{\omega: X_1(\omega) = 10, 30, 50\}$ is equivalent to the event $\{f_1, f_3, f_5\} = \{\omega: \text{odd}\}$.
- ❖ Similarly, a subset $\{\omega: X_2(\omega) = -1, -2\}$ is equivalent to the event $\{f_1, f_2\}$.
- ❖ Thus, we can talk about assigning a probability measure on the events described by random variables, in exactly the same way we do with the events of Ω .

Distribution Function

- ❖ Suppose the probability measure defined on the die experiment was

$$P(f_k) = 1/6 \text{ for all } k=1, 2, \dots, 6$$

- ❖ Then, correspondingly we could have the probability measure defined on the random variables X_1 and X_2

- ❖ For X_1 , we have

$$P(X_1 = 10) = 1/6, P(X_1=20) = 1/6, \dots$$

- ❖ For X_2 , we have

$$P(X_1 = -1) = 1/6, P(X_2=-2) = 1/6, P(X_2=-3) = 1/6\dots$$

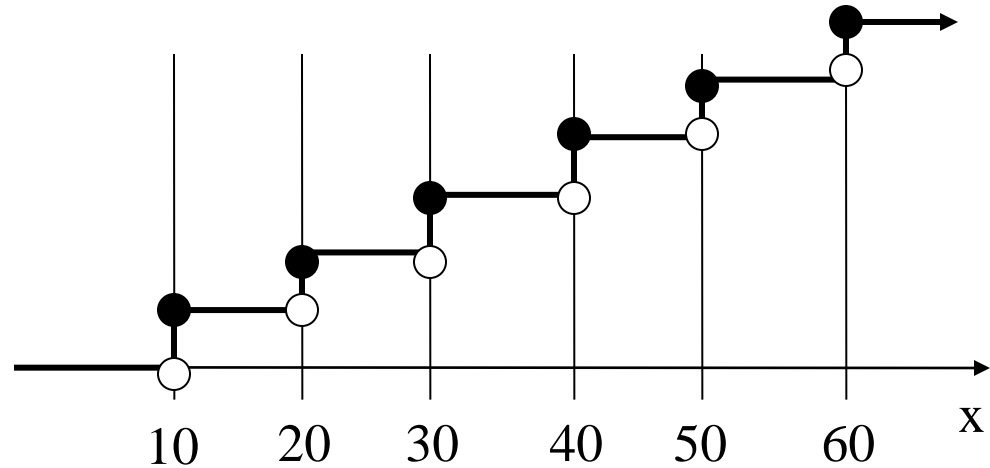
- ❖ Probability assignment is easy with finite and countable sample space.

Distribution Function (2)

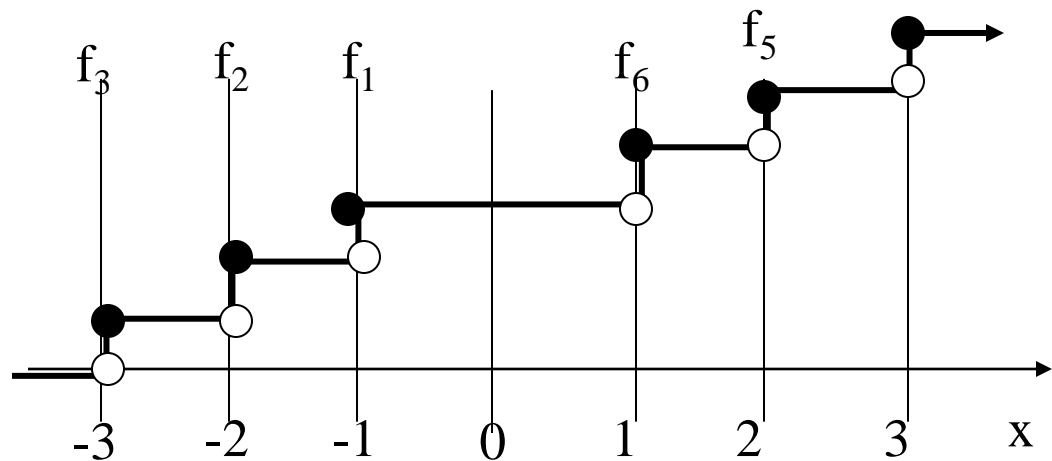
- ❖ We use a cumulative distribution function to deal with an infinite uncountable sample space.
 - For example, $S = [0, 1]$.
- ❖ The probability is assigned on the intervals of interest.
 - A collection of intervals, say events, is of interest.
 - A sigma field can be formed for the collection of intervals.
 - Distribution function $F_X(x)$ of a random variable X is defined as
$$F_X(x) := P(\omega: X(\omega) \leq x)$$
 - It is called the cumulative distribution function (CDF) of X .
- ❖ Examples) Find the distribution functions for random variable X_1 and X_2 that were defined in the roll-a-die experiment.

Distribution Function (3)

- ❖ $F_{X_1}(x) = P(X_1 \leq x)$
- ❖ Note that the function is right continuous



- ❖ $F_{X_2}(x) = P(X_2 \leq x)$



Properties of Distribution Function $F(x)$

❖ Non-decreasing function of x : For $x_2 > x_1$, $F(x_2) \geq F(x_1)$

❖ Continuous from the right.

$$\lim_{\varepsilon \downarrow 0} F(x+\varepsilon) = F(x),$$

❖ $F(-\infty) = P(X \leq -\infty) = 0$

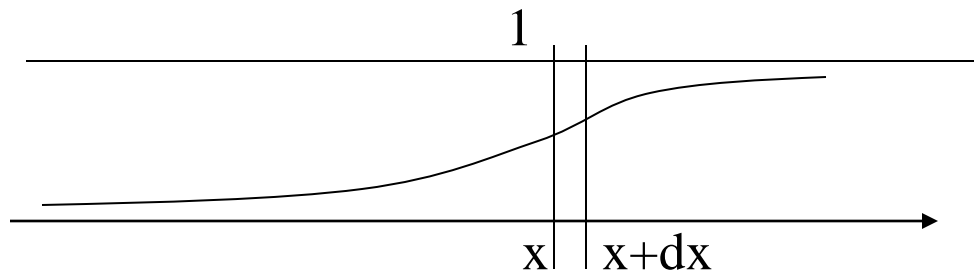
❖ $F(+\infty) = P(X \leq \infty) = 1$

❖ $0 \leq F(x) \leq 1$

Probability Density Function $f(x)$

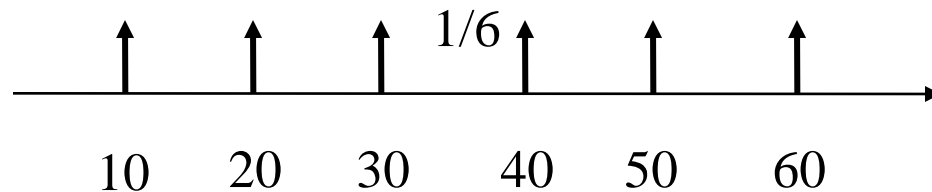
❖ $f(a) := dF(x)/dx \big|_{x=a}$

❖ $P\{x < X \leq X+dx\} = P\{X \leq x+dx\} - P\{X \leq x\} = f(x) dx$



❖ Example of pdf of X_1 :

$$f_1(x) = (1/6) \sum_{k=1}^6 \delta(x - 10k)$$



Ensemble Averages (Expected Value)

❖ Ensemble Average $E(X)$

❖ 1st moment: $m_1 = E\{X\} := \int_{-\infty}^{\infty} x f(x) dx$

❖ Note that this operator is a linear operator

❖ 2nd moment: $m_2 = E\{X^2\} = \int_{-\infty}^{\infty} x^2 f(x) dx$

❖ $\text{Var}(X) = E\{(X - E\{X\})^2\} = E\{(X - m_1)^2\}$
 $= E(X^2) - E\{X\}m_1 - m_1 E\{X\} + E(m_1^2) = m_2 - m_1^2$

Ensemble Average of Product XY

- ❖ X and Y are two random variables with PDF $f_x(x)$ and $f_y(y)$
- ❖ Then, $f_{XY}(x, y)$ is the joint density function
- ❖ $E\{XY\} = \int \int x y f_{xy}(x, y) dx dy$
 - This is called the **Correlation** of the two random variables X and Y
 - Note, what happens when X and Y are independent
 - When $E\{XY\} = E\{X\}E\{Y\}$, X and Y are said to be **mutually uncorrelated**
 - Note, if you have two indep. r.v.s, then they are uncorrelated, but not vice versa
- ❖ $E\{(X-E(X))(Y-E(Y))\}$ is called the **Covariance**
 - Note what happens when two are uncorrelated

Binomial Distribution

❖ Binomial PDF: prob. of obtaining k “1”s in N Bernoulli trials

$$P(k) = \binom{N}{k} p^k (1 - p)^{N-k}$$

By letting $x = k$, where $k = 0, 1, 2, \dots, N$

$$f(x) = \sum_{k=0}^N P(k) \delta(x - k)$$

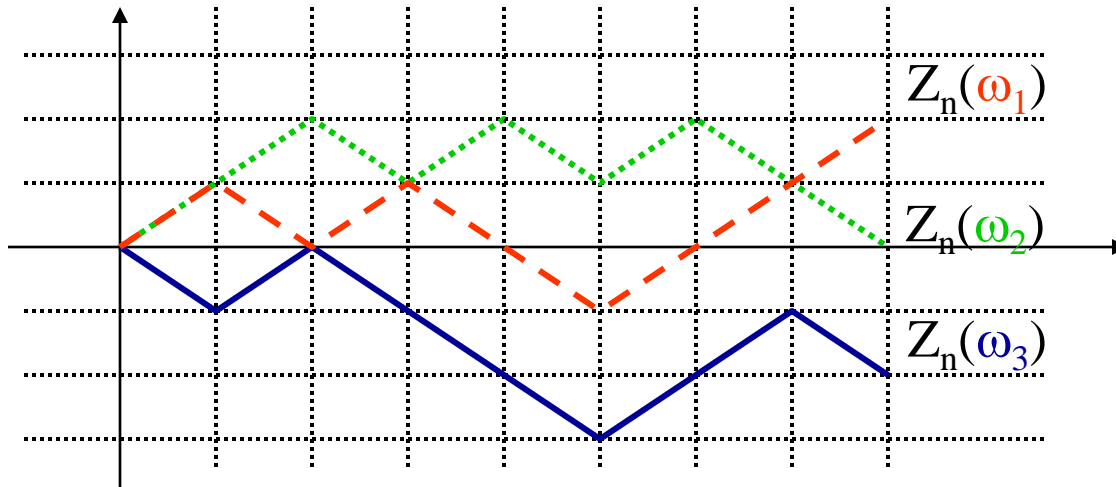
– Binomial expansion: $(p + q)^N = \sum_{k=0}^N P(k)$

$$= \sum_{k=0}^N \binom{N}{k} p^k (1 - p)^{N-k}$$

Random Processes (Stochastic Processes)

- ❖ A random process can be described as a *collection of random variables* parameterized by time index t .
- ❖ Continuous random process $\{x_t, t \in [0, \infty]\}$
 - For a fixed t , x_t is a random variable.
- ❖ Discrete-time random process $\{x_k\}$, such that $x_1, x_2, \dots, x_k, \dots$
 - Again, each x_k is a random variable.
- ❖ Ex) Flipping a coin repeatedly $x_k = 1$ with prob. p
or -1 with prob. $1-p$
- ❖ Ex2) $Z_n := \sum_{k=1}^n X_k$

Random Processes (Stochastic Processes)



Each is called a *sample path*.

Ensemble:
Collection of every possible sample paths

- ❖ Suppose we observe a path being taken by the random process Z_n in 8 steps.
- ❖ There are 2^8 possible paths. This collection is called *ensemble*.
- ❖ In an observation, Z_n takes a particular path. It is called a *sample path* taken by the random process in an experiment.
- ❖ We may interpret it as an outcome of a random experiment: choosing one object out of 2^8 objects.
- ❖ We use $Z_n(\omega)$ to denote a particular sample path.

Stationary Processes (Strict Sense)

- ❖ A random process $x(t)$ is said to be stationary to the order N , if for any t_1, t_2, \dots, t_N ,
$$f_x(x(t_1), x(t_2), \dots, x(t_N)) = f_x(x(t_1+t_0), x(t_2+t_0), \dots, x(t_N+t_0))$$
where t_0 is any arbitrary real constant.
- ❖ That is, the joint distribution function is shift-invariant in time.
- ❖ If this holds for any N , then we say the process is *strictly stationary*.

Ergodic Random Process (Important)

- ❖ If time average = ensemble average, then ergodic.
- ❖ A random process is said to be *ergodic* if the time average of any sample path is equal to the ensemble average (expectation).
 - $E(x(t)) = \lim_{T \rightarrow \infty} (1/T) \int_T x(t) dt$ (ergodic in mean)
 - $E(x^2(t)) = \lim_{T \rightarrow \infty} (1/T) \int_T x^2(t) dt$ (ergodic in 2nd moment)
- ❖ An ergodic process must be a stationary process (but not vice versa).
 - If a process is non-stationary, then the ensemble average of the process changes over time.
 - Not all stationary processes are ergodic.
 - Select a coin from a box containing two coins with different weight in a box and throw them repeatedly.

Example of Ergodic Processes

- ❖ Show that $x(t) = \cos(2\pi f_0 t + \theta)$ is ergodic in mean and 2nd moment, where θ is uniformly distributed over $[0, 2\pi]$.
- ❖
$$\begin{aligned} E(x(t)) &= (1/2\pi) \int_0^{2\pi} \cos(2\pi f_0 t + \theta) d\theta \\ &= (1/2\pi) \sin(2\pi f_0 t + \theta) \Big|_0^{2\pi} \\ &= 0 \end{aligned}$$
- ❖
$$\begin{aligned} E(x^2(t)) &= (1/2\pi) \int_0^{2\pi} \cos^2(2\pi f_0 t + \theta) d\theta \\ &= (1/2\pi) (1/2) \int_0^{2\pi} 1 + \cos(2\pi 2f_0 t + 2\theta) d\theta \\ &= 1/2 \end{aligned}$$
- ❖ $T_0 = 1/f_0$
- ❖ $\langle x(t) \rangle = (1/T_0) \int_0^{T_0} \cos(2\pi f_0 t + \theta) dt = 0$
- ❖ $\langle x^2(t) \rangle = (1/T_0) \int_0^{T_0} \cos^2(2\pi f_0 t + \theta) dt$
 $= (1/2T_0) \int_0^{T_0} 1 + \cos(2\pi 2f_0 t + 2\theta) dt = 1/2$

HW#0

- ❖ Complete the following problems and submit by the next lecture.
- ❖ Will be checked, but not graded.

Mutually Exclusive vs. Independence

- ❖ The events A and B are mutually exclusive. Can they be independent?

Probability/Random Variable/Distribution

- ❖ A coin with $\Pr\{\text{tail}\} = p$ is tossed n times.
 - (a). Find the probability of the event that shows k heads in n trials.
 - (b). What is the conditional probability that the first toss is head given that there are 2 heads in n tosses?
 - (c). Let X be the random variable denoting the number of heads. Specify the domain and the range of this random variable.
 - (d). Sketch the cumulative distribution function of X for $n = 6$. Assume $p = 0.1$.

Probability

- ❖ Consider a box shown above. It has 10 pockets. Two balls are thrown into the box in sequence. A ball can be placed in any pocket with equal probability. No pocket can hold two balls. No balls can be placed outside the box.
- (a) What is the probability that both balls are placed into the same column?
 - (b) What is the probability that both balls are placed into the same row?
 - (c) What is the probability that the two balls are separated into both different row and different column?
 - (d) Is there any other case? Justify your answer.

Joint distribution/conditional probability

- ❖ Two i.i.d. (indep. identically distr.) binary random variables, X_1 and $X_2 \in \{1, -1\}$ with p and $(1-p)$. What's the conditional probability $\Pr(X_1=1|X_2=1)$?
- ❖ Now consider a series of binary random variables, X_1, X_2, X_3, \dots . X_1 produces equally likely outcomes, the second and the rest are i.i.d. random variables producing the outcome 1 with probability p and outcome -1 with probability $(1-p)$ where p is a number between zero and 1. The number p is determined at the first experiment. p is $1/2$ if $X_1 = 1$ or $1/4$ if $X_1 = -1$.
 - What is $\Pr\{X_4 = 1\}$?
 - What about $\Pr\{X_1 + X_2 = 2\}$?

Joint distribution/conditional probability

❖ In this problem, θ , U , V , e_1 and e_2 are all binary $\{0, 1\}$ random variables. Let's use notation $P_\theta = \Pr(\theta=1)$, and thus $\Pr(\theta = 0) = 1 - P_\theta$. The same goes for the other random variables. For example, $P_{e_1} = \Pr(e_1 = 1)$, and $P_{e_2} = \Pr(e_2 = 1)$.

❖ Suppose U and V are binary random variable, i.e.

$$U = \theta + e_1 \text{ modulo } 2$$

$$V = \theta + e_2 \text{ modulo } 2$$

where $P_\theta = p$, $P_{e_1} = p_1$ and $P_{e_2} = p_2$, and θ , e_1 and e_2 are mutually independent.

1. For $P_\theta = 0.6$, $P_{e_1} = 0.1$ and $P_{e_2} = 0.2$, find the joint distribution $\Pr(U = x, V = y)$.
2. Repeat 1 with $P_\theta = 0.9$, $P_{e_1} = 0.01$ and $P_{e_2} = 0.02$.

Urn Problem

- ❖ A box contains m white balls and n black balls. Balls are drawn at random one at a time without replacement. Find the probability of encountering a white ball by the k -th draw.

Total Probability/Bayes' Theorem

- ❖ Suppose there is a test for a prostate cancer which is known to be 95% accurate. A person took the test and the result came out positive. Suppose that the person comes from a population of a million, where 20,000 people suffer from that disease. What can we conclude about the probability that the person under test has that particular cancer.

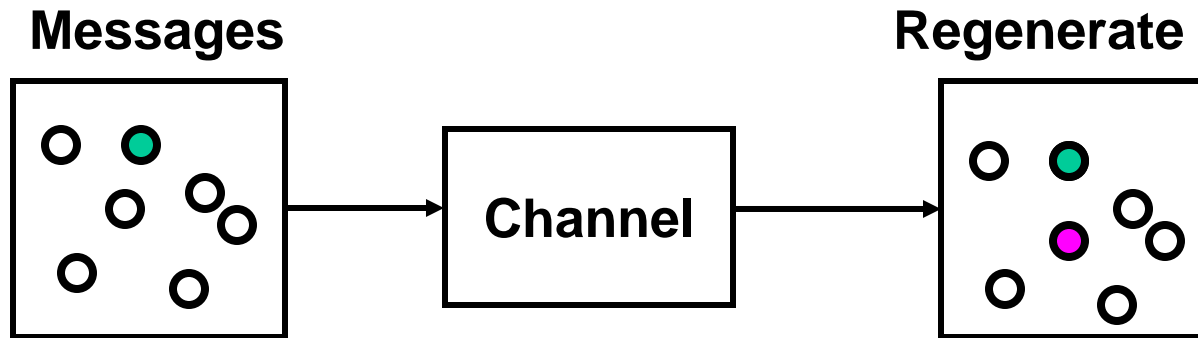
Information Theory

The 1st Module

Claude E. Shannon (1916-2001)

- ❖ Math/EE Bachelor from UMich (1936)
- ❖ MSEE and Math Ph.D. from MIT (1940)
- ❖ A landmark paper “Mathematical Theory of Communications” (1948)
 - Founder of Information Theory
 - Fundamental limits on communications
 - Information quantified as a logarithmic measure
- ❖ For more info on him, make a visit to
<http://www.bell-labs.com/news/2001/february/26/1.html>

Novel Perspective on Communications



- ❖ Communications: Transfer of information from a source to a receiver
- ❖ Messages (information) can have semantic meaning; but they are irrelevant for the design of a comm. system.
- ❖ What's important then?
 - A message is selected from a set of all possible messages and transmitted, and regenerated at the receiver.
 - The size of the message set has something to do with the amount of information.
- ❖ *The capacity of the channel is the maximum size of message set that can be transferred over the channel and can be regenerated almost error-free at the receiver*

The Size M of Message Set

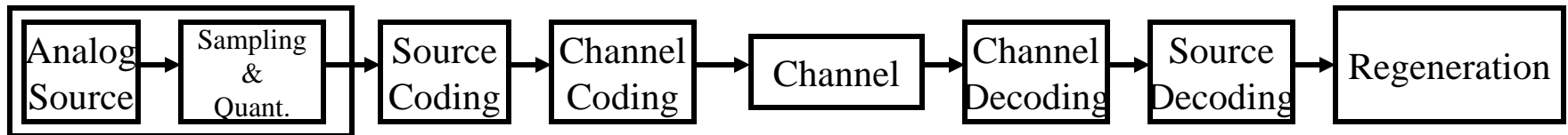
- ❖ Is the Amount of information
- ❖ M or any monotonic function of M can be used as a measure of information.
- ❖ His choice was the logarithmic function. Why?
 - If $M_1 > M_2$, $\log(M_1) > \log(M_2)$
 - When base 2, $\log_2(M)$ is the number of memory cells.
 - We call the resulting unit “bits.”
 - A four-bit register can represent a message set of size 2^4 , and a three-bit register 2^3 .
 - The amount of information is $\log_2(2^4) = 4$ bits (and 3 bits).
 - This choice was made out of convenience; but considered appropriate (See the axiomatic definition of entropy in Cover & Thomas 1st Ed., Prob2.4)

Fundamental Limits on Communications Systems

- ❖ The Sampling and Modulation Theorem (Nyquist and Hartley 1928)
- ❖ Source and Channel Coding Theorem (Shannon)

- ❖ Can we define a quantity which measures the amount of information produced by a digital or an analog source?
- ❖ Rate Distortion and Source Coding Theorem:
 - “ n -bit quantization”: Distortion will increase if we reduce n .
 - *Source code* takes away redundancy in the source and reduces the number of bits required.

- ❖ How about the size of message set that can be transferred over a noisy channel almost error-free?
- ❖ Channel Capacity and Channel Coding Theorem:
 - *Channel code* adds redundancy in order to gain protection against random error occurring in the channel



Uncertainty and Entropy

- ❖ Suppose a set of n possible outcomes, each having the probability of occurrence as p_1, p_2, \dots, p_n .
- ❖ After a random experiment, we have an outcome.
- ❖ Then, we can say about the occurrence of an event.
- ❖ **Entropy is a measure of uncertainty** (randomness) on the occurrence of an event.
- ❖ We use logarithmic measures (non-negative)
 - $\log(1/p_i) \geq 0$,
- ❖ If $p_i < p_j$, then $\log(1/p_i) > \log(1/p_j)$.
 - Less probable event means larger uncertainty.
 - More probable event means smaller uncertainty.
 - The sure event has zero uncertainty.

Definition of Entropy

- ❖ Entropy is the average *measure* of uncertainty of a distribution, p_1, p_2, \dots, p_n .

$$H(p_1, p_2, \dots, p_n) := \sum_{j=1}^n p_j \log(1/p_j)$$

Some Properties of Entropy

- ❖ Uncertainty = Amount of Information = The number of bits needed in representation
- ❖ More uncertain event carries more information.
- ❖ The sure event carries zero amount of information
 - A binary source generates “1” with probability 1. Then, the source produces zero amount of information, i.e., $\log(1/1) = 0$.
 - A binary source generates “1” and “0” with equal probability. Each event carries the same amount of information. Then, this source generates 1 bit of information.

Entropy of a RV

❖ Let X be a random variable with alphabet $A = \{x_1, x_2, \dots, x_n\}$ and its probability mass function $p(x) = \Pr\{X=x_i \in A\}$

❖ We define entropy for r.v. X

$$H(X) := \sum_{x \in \mathcal{X}} p(x) \log(1/p(x))$$

– Note that in fact this measure has nothing to do with the random variable X , but has everything to do with the distribution.

– The range of X does not play any role in the calculation of $H(X)$.

❖ When the base of the logarithm is 2, the unit is “bits.”

❖ When the base is e , the unit is “nats.”

$H(X)$ is the Average Uncertainty (Information) of X

- ❖ Let's take some examples
- ❖ Ex1) When X is binary
- ❖ Ex2) When X is quaternary

Entropy gives the largest lower bound on the number of bits required to represent the set of events

- ❖ Ex3) Average Information Content in English
- ❖ Assume all 26 letters occur equally likely from a source
 - $H = \log_2(26) = 4.7$ bits/character

Entropy gives the largest lower bound on the number of bits required to represent the set of events

❖ Assume some distribution other than uniform









- a, e, o, t with prob = 0.1
- h, i, n, r, s with prob = 0.07
- c, d, f, l, m, p, u, y with prob. = 0.02
- b, g, j, k, q, v, w, x, z with prob. = 0.01
- **H = 4.17** bits/character

❖ Thus, if there was a source generating letters according to this distribution (ignoring spaces, commas, etc), then the source's information rate is 4.17 bits per character.

Entropy and Information

- ❖ Entropy is the minimum attainable average length of any binary description system.
 - I'll explain this with the next example.
- ❖ Ex4) Suppose a race of 8 horses. The race was held in LA yesterday. We are here in Gwangju. There is a reporter in LA. The reporter can only make a binary answer—Yes or No—to our question. Now, knowing that the winning prob. of each horse is $(1/2, 1/4, 1/8, 1/16, 1/64, 1/64, 1/64, 1/64)$ respectively; which horse would you ask first to be the winning horse? The objective is to determine the winning horse as quickly as possible.
 - Note that the entropy is $H = 2$ bits.

Entropy and Information

p_i				Length
1/2	0		0	1
1/4	1		10	2
1/8	2		110	3
1/16	3		1110	4
1/64	4		111100	6
1/64	5		111101	6
1/64	6		111110	6
1/64	7		111111	6

- ❖ The map from the horse index to the binary sequence is a code.
- ❖ This coding strategy achieves the entropy bound.
- ❖ The average length = $1(1/2) + 2(1/4) + 3(1/8) + 4(1/16) + 6(1/64)*4 = 2$
(which is the same as $H = 2$)
- ❖ What happens if the horse index, $0, 1, \dots, 7$, was used for the coding? How many bits would be needed then?

Joint Entropy and Conditional Entropy

- ❖ Joint Entropy: The joint entropy $H(X, Y)$ of a pair of discrete random variable (X, Y) with a joint distribution $p(x, y)$ is defined as

$$\begin{aligned} H(X, Y) &:= - \sum_x \sum_y p(x, y) \log p(x, y) \\ &= - E\{\log p(X, Y)\} \end{aligned}$$

- ❖ Conditional Entropy:

$$\begin{aligned} H(Y | X) &:= - \sum_x \sum_y p(x, y) \log p(y | x) \\ &= - E\{\log p(Y|X)\} \\ &= - \sum_x p(x) H(Y | X = x) \\ &= - \sum_x p(x) \sum_y p(y|x) \log p(y|x) \end{aligned}$$

Chain Rule: $H(X, Y) = H(X) + H(Y|X)$

$$\begin{aligned} \blacklozenge H(X, Y) &:= - \sum_x \sum_y p(x, y) \log p(x, y) \\ &= - \sum_x \sum_y p(x, y) \log[p(x) p(y|x)] \\ &= - \sum_x \sum_y p(x, y) [\log p(x) + \log p(y|x)] \\ &= - \sum_x p(x) \log p(x) - \sum_x \sum_y p(x, y) \log p(y|x) \\ &= H(X) + H(Y|X) \\ &\quad \text{or similarly} \\ &= H(Y) + H(X|Y) \end{aligned}$$

Example

- ❖ $H(X) = 3/8 * \log_2(8/3) + 5/8 * \log_2(8/5) = 0.9544$
- ❖ $H(Y) = 6/8 * \log_2(8/6) + 2/8 * \log_2(8/2) = 0.8113$
- ❖ $H(Y|X) = \sum_x p(x) H(Y|X=x)$
 $= 3/8 * H(Y|X=0) + 5/8 * H(Y|X=1)$
 $= 3/8 * H(2/3, 1/3) + 5/8 * H(4/5, 1/5)$
 $= 3/8 * 0.9183 + 5/8 * 0.7219$
 $= 0.7955$
- ❖ $H(X, Y) = H(X) + H(Y|X) = 1.75$
- ❖ $H(X, Y) = - E\{\log p(X, Y)\}$
 $= 2/8 * \log_2(4) + (4/8) * \log_2(2) + 2 * 1/8 * \log_2(8)$
 $= 1/4 * 2 + 1/2 + 2 * 3/8 = 1 + 3/4 = 1.75$

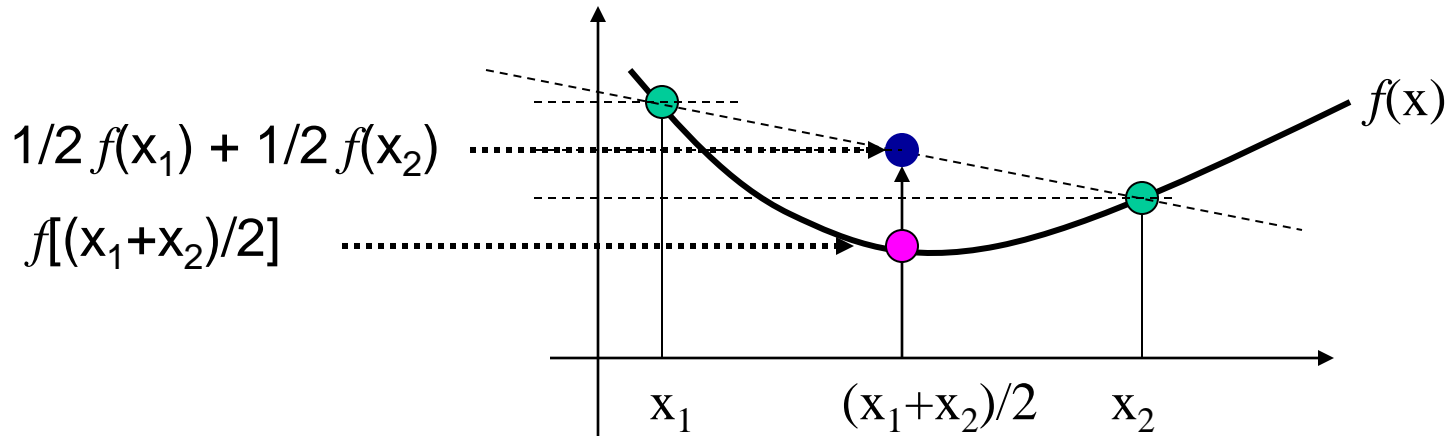
		X	
		0	1
Y			
	0	2/8	4/8
1	1/8	1/8	

The units are [bit].

Max. entropy when uniform

- ❖ $H(X) \leq \log|\mathcal{X}|$, where $|\mathcal{X}|$ is the size of alphabet, with equality iff X is uniform over \mathcal{X} .
 - Non-uniform gives maximum entropy under a certain input criteria
 - cf) Gaussian distribution gives max. entropy under average energy constraint.
 - I owe you the proof of this statement, especially the only if part.

Jensen's Inequality



❖ For any $f(x)$ convex U, it is easy to see

$$1/2 f(x_1) + 1/2 f(x_2) \geq f[(x_1+x_2)/2]$$

❖ This holds true for any distribution $p_1 + p_2 = 1$ such that

$$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2)$$

❖ For r.v. X and function f convex U,

$$E\{f(X)\} \geq f(E\{X\})$$

– For strictly convex U $f(x)$, equality *iff* X is a constant

❖ What if a function is concave \cap ?

What about in terms of the second derivative?

$\log(x)$ concave?

over what interval?

Relative Entropy is Non-Negative!

- ❖ $D(p \parallel q)$ = Kullback Leibler Distance between two distributions $p(z)$ and $q(z)$ or Relative Entropy
$$:= \sum_z p(z) \log(p(z)/q(z))$$
- ❖ Suppose $p(z)$ and $q(z)$ are *strict positive* distributions (no zero probability masses). Let S_p and S_q denote their alphabets respectively.
- ❖ - $D(p \parallel q) = \sum_{z \in S_p} p(z) \log[q(z)/p(z)]$
$$\leq \log\{\sum_{z \in S_p} p(z) [q(z)/p(z)]\}$$

(log is strict concave \cap ;thus equality only if $p(z)/q(z)$ constant)

$$= \log\{\sum_{z \in S_p} q(z)\}$$
$$\leq \log\{\sum_{z \in S_q} q(z)\} = \log(1) = 0$$
- ❖ Thus, $D(p \parallel q) \geq 0$ with equality *iff* $p(z) = q(z)$.
 - Is the equality iff part easy to prove?

See page 87 for another way!

Example on Relative Entropy

- ❖ Let $\mathcal{X} = \{0, 1\}$ and two distr.'s $p(x)$ and $q(x)$
- ❖ $p(x=0) = 1-r$, $p(x=1) = r$
- ❖ $q(x=0) = 1-s$, $q(x=1) = s$
- ❖ $D(p \parallel q) = (1-r) \log[(1-r)/(1-s)] + r \log[r/s]$
- ❖ $D(q \parallel p) = (1-s) \log[(1-s)/(1-r)] + s \log[s/r]$
- ❖ Thus, $D(p \parallel q) \neq D(q \parallel p)$ in general
 - *Relative Entropy is not symmetric in general*
- ❖ Ex) when $r = s$, then $D(p \parallel q) = D(q \parallel p) = 0$
- ❖ Ex) when $r = 1/2$, $s = 1/4$, $D(p \parallel q) = 0.2075$, $D(q \parallel p) = 0.1887$

Relative Entropy is Non-Negative!

(Other Approach)

- ❖ Suppose $p(z)$ and $q(z)$ are *strict positive* distributions (no zero probability masses). Let S_p and S_q denote their alphabets respectively.
- ❖ If the sum $\sum_{z \in S_p} p(z) \log(p(z)/q(z)) = 0$, then $p(z) = q(z)$ for all $z \in S_p$.
- ❖ Proof:

$$\begin{aligned} \sum_z p(z) \log(p(z)/q(z)) &\geq \sum_z p(z) (1 - q(z)/p(z)) && \text{(Why?)} \\ &= \sum_{z \in S_p} p(z) - \sum_{z \in S_p} q(z) \\ &\geq (1 - 1) = 0 && \text{(Why?)} \end{aligned}$$

The first equality holds only if $p(z)/q(z) = 1$.
The second equality holds only if $S_p = S_q$.

Entropy is maximum, when uniform distributed

❖ Proof: Let $u(x)$ be uniform on \mathcal{X}

$$\begin{aligned} H(p) &= \sum_x p(x) \log(1/p(x)) \\ &= \sum_x p(x) \{ \log(1/p(x)) + \log(u(x)) - \log(u(x)) \} \\ &= - \sum_x p(x) \log(u(x)) + \sum_x p(x) \{ \log[u(x)/p(x)] \} \\ &= \log|\mathcal{X}| - D(p \parallel u) \end{aligned}$$

Mutual Information is Non-Negative!

$$\begin{aligned} \blacklozenge \quad I(X; Y) &:= \sum_x \sum_y p(x, y) \log[p(x, y)/p(x)p(y)] \\ &= D(p(x, y) \parallel p(x)p(y)) \end{aligned}$$

----- Distance between the joint and the product distribution.

----- Thus, Mutual Information is non-negative.

$$= E_{(x,y)} \{ \log[p(X, Y)/p(X)p(Y)] \} \geq 0$$

$$I(X, Y) = H(X) - H(X | Y)$$

$$\begin{aligned}
 \text{❖ } I(X; Y) &= \sum_{x \in X} \sum_{y \in Y} p(x, y) \log[p(x, y)/p(x)p(y)] \\
 &= \sum_{x \in X} \sum_{y \in Y} p(x, y) \log[\cancel{p(y)} p(x | y) / p(x) \cancel{p(y)}] \\
 &= \sum_{x \in X} \sum_{y \in Y} p(x, y) \{ \log[p(x | y)] - \log[p(x)] \} \\
 &= H(X) - H(X|Y)
 \end{aligned}$$

❖ Reduction in uncertainty of X due to the knowledge of Y

❖ Also, $I(X; Y) = H(Y) - H(Y|X)$

❖ How much can I tell about X knowing Y?

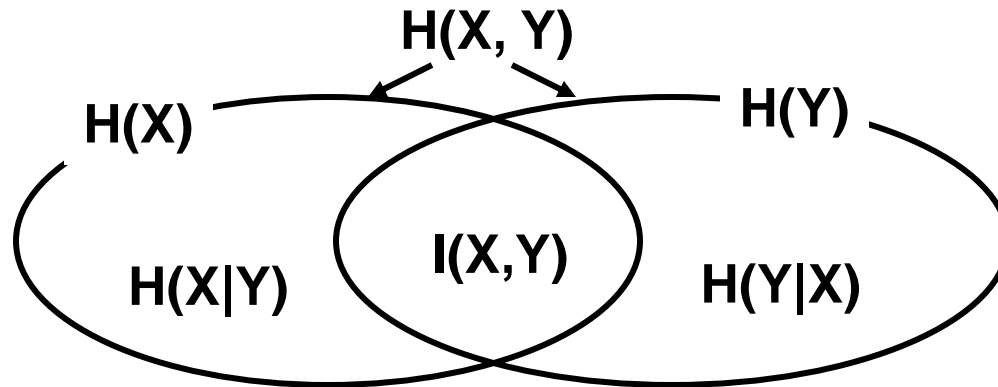
❖ How much can I tell about Y knowing X?

❖ $I(X; Y) = I(Y; X)$

Mutual Information?

- ❖ The measure of amount of information about X we can have knowing Y (vise versa).
 - Cf) Measure of correlation between X and Y , see P2.11.
- ❖ Ex) Suppose $Y = X$, then $H(X|Y) = 0$ (no uncertainty). \rightarrow Self-mutual information is entropy.
 - Thus, knowing Y means knowing X exactly (the full information $H(X) = H(Y)$ is obtained)
- ❖ Ex) Suppose Y and X independent, then $H(X|Y) = H(X)$, then $I(X;Y) = H(X) - H(X) = 0$.
 - Knowing Y cannot tell anything about X .
 - Can you show that if $I(X; Y) = 0$, then X and Y independent?

Relationships



❖ $I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$

❖ Thus, $I(X; Y) = H(X) + H(Y) - H(X, Y)$

--- use $H(X, Y) = H(X) + H(Y|X)$

Conditioning reduces entropy

- ❖ $H(X|Y) \leq H(X)$, with equality *iff* X and Y independent
 - $I(X; Y) = H(X) - H(X|Y) \geq 0$
- ❖ cf) $I(X; Y) = 0$ *iff* X and Y independent.

Chain Rules

❖ Let X_1, X_2, \dots, X_n drawn from $p(x_1, x_2, \dots, x_n)$. Then,

$$H(X_1, X_2) = H(X_1) + H(X_2|X_1)$$

$$H(X_1, X_2, X_3) = H(X_1) + H(X_2, X_3|X_1)$$

$$= H(X_1) + H(X_2|X_1) + H(X_3 |X_1, X_2)$$

...

$$H(X_1, X_2, \dots, X_n) = \sum_{i=1}^n H(X_i |X_{i-1}, \dots, X_1 \}$$

 **Watch out for the notation**

Results in previous page lead to

❖ $H(X_1, X_2, \dots, X_n) \leq \sum_{i=1}^n H(X_i)$
with equality iff X_i are independent

Conditional Mutual Information

$$\begin{aligned} \blacklozenge \quad I(X; Y | Z) &= H(X | Z) - H(X | Y, Z) \\ &= E\{\log[p(X, Y | Z)/p(X | Z)p(Y | Z)]\} \end{aligned}$$

\blacklozenge Can we say this?

– $I(X; Y | Z) = 0$ IFF X and Y indep. given Z .

Chain Rule for Information

$$\begin{aligned} \blacklozenge & I(X_1, X_2, X_3; Y) \\ &= E\{\log[p(X_1, X_2, X_3, Y)/\mathbf{p(X_1, X_2, X_3)}p(Y)]\} \\ &= H(X_1, X_2, X_3) - H(X_1, X_2, X_3|Y) \\ &= H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) \\ &\quad - H(X_1|Y) - H(X_2|X_1, Y) - H(X_3|X_1, X_2, Y) \\ &= I(X_1; Y) + I(X_2; Y|X_1) + I(X_3; Y|X_1, X_2) \end{aligned}$$

In general, we have

$$I(X_1, \dots, X_n; Y) = \sum_{i=1}^n I(X_i; Y | X_1, \dots, X_{i-1})$$

Concavity of log: *Log Sum Inequality*

- ❖ For non-negative a_1, a_2, \dots, a_n and b_1, b_2, \dots, b_n

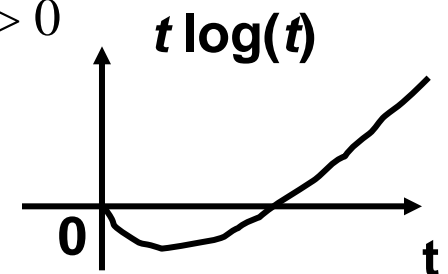
$$\sum_{i=1}^n a_i \log(a_i/b_i) \geq (\sum_{i=1}^n a_i) \log[\sum a_i/\sum b_i]$$

with equality iff a_i/b_i constant.

Note, sum of numbers \geq a single number.

- ❖ Proof:

- $f(t) = t \log t, t > 0$, is strictly convex ($f''(t) = 1/t > 0$ for $t > 0$)
- Use the Jensen's Inequality: avg. of maps \geq map of avg.
- $\sum_{i=1}^n \alpha_i f(t_i) \geq f(\sum \alpha_i t_i)$ for $\alpha_i \geq 0$ and $\sum_i \alpha_i = 1, t_i > 0$
- Substitute $\alpha_i = b_i/\sum_i b_i$, and $t_i = a_i/b_i$
- Equality iff a_i/b_i constant



Use the Log Sum Inequality to show $D(p \parallel q) \geq 0$

$$\begin{aligned} \blacklozenge D(p \parallel q) &= \sum p(x) \log[p(x)/q(x)] \\ &\geq \sum p(x) \log[\sum p(x)/\sum q(x)] \\ &= 1 \log(1/1) = 0 \end{aligned}$$

$D(p \parallel q)$ is convex in the pair (p, q)

- ❖ Mixing distributions decreases the relative entropy
- ❖ Consider two pairs (p_1, q_1) and (p_2, q_2) of distributions
- ❖ Which one is bigger?
 - Avg. of relative entropies, $0.5(D(p_1 \parallel q_1) + D(p_2 \parallel q_2))$ – (1)
 - Relative entropy of avg. distribution: $D(0.5(p_1 + p_2) \parallel 0.5(q_1 + q_2))$ – (2)
- ❖ (1)’: $p_1(x) \log(p_1(x)/q_1(x)) + p_2(x) \log[p_2(x)/q_2(x)]$
- ❖ (2)’: $(p_1(x) + p_2(x)) \log[(p_1(x) + p_2(x))/(q_1(x) + q_2(x))]$
- ❖ (1)’ \geq (2)’ – the Log Sum Inequality
- ❖ Summing over all x , we have (1) \geq (2)

Concavity of Entropy

❖ Recall the proof that entropy is maximum when the distribution is uniform.

❖ Let $u(x)$ be uniform on \mathcal{X}

$$\begin{aligned} H(p) &= \sum_x p(x) \log(1/p(x)) \\ &= \sum_x p(x) \{ \log(1/p(x)) + \log(u(x)) - \log(u(x)) \} \\ &= -\log(u(x)) + \sum_x p(x) \{ \log[u(x)/p(x)] \} \\ &= \log|\mathcal{X}| - D(p \parallel u) \end{aligned}$$

❖ Not only is entropy maximum for uniform distribution but also a concave function of $p(x)$.

Concavity of Entropy (other approach)

- ❖ $H(p)$ is a *concave* function of a distribution $p(x)$
- ❖ This means if you mix distributions, the entropy increases.
- ❖ Let $X_1 \sim p_1(x)$ and $X_2 \sim p_2(x)$
- ❖ Let $Z = X_\theta$ where $\theta = 1$ with prob. λ and 2 with $1-\lambda$
- ❖ Thus, the distr. of Z is $\lambda p_1(x) + (1 - \lambda) p_2(x)$
- ❖ We know $H(Z) \geq H(Z | \theta)$
 - conditioning reduces entropy

❖ Thus, we have

$$H[\lambda p_1(x) + (1 - \lambda) p_2(x)] \geq \lambda H[p_1(x)] + (1-\lambda) H[p_2(x)].$$

- This shows $f(E) \geq E(f)$. Thus, entropy is a concave function of distribution.

Concavity of $I(X; Y)$ over $p(x)$ given $p(y|x)$

- ❖ $I(X; Y) = H(Y) - H(Y|X)$
- ❖ $H(Y)$ is a concave function of $p(y)$.
 - Note $p(y) = \sum p(x) p(y|x)$ is a linear function of $p(x)$.
 - Thus, $H(Y)$ is a concave function of $p(x)$.
- ❖ $H(Y|X) = \sum p(x) H(Y|X = x)$, is a linear function of $p(x)$.
- ❖ Thus, $I(X; Y)$ is a concave function of $p(x)$ given $p(y|x)$.

Sequence of results so far

- ❖ Relative entropy is non negative. Proved!
- ❖ Relative entropy is zero IFF the two distributions are identical. Proved!
- ❖ Entropy $H(X)$ is maximum with $X \sim$ uniform distribution.
- ❖ Mutual information is a relative entropy.
- ❖ Mutual information is thus non negative.
- ❖ MI $I(X; Y) = 0$ IFF X and Y independent.
- ❖ Conditioning reduces entropy.
- ❖ Entropy is a concave function of distribution.
- ❖ MI $I(X; Y)$ is a concave function of $p(x)$ given $p(y|x)$.

HW#1

- ❖ Cover & Thomas: Ch2: 1, 2, 5, 8, 12, 14, 18,
- ❖ Showing the convexity of $f(x) = e^x$ is easy. Use the Calculus: Take the derivatives twice and show that it's positive everywhere. Now, prove the convexity of $f(x)$ using the general convexity proving technique learned in this lecture.

- ❖ (Challenge; Optional) Consider arbitrary random variables X_1, X_2 , and

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} + \begin{pmatrix} N_1 \\ N_2 \end{pmatrix}$$

where the matrix elements $[a_{ij}]$ are arbitrary non zero constants and N_1 and N_2 are independent random variables. Let's denote $\mathbf{X} := \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$.

Prove or disprove $I(\mathbf{X}; Y_1, Y_2) \leq I(\mathbf{X}; Y_1) + I(\mathbf{X}; Y_2)$.

HW#1

- ❖ $I(X_1, X_2; Y)$ and $I(\mathbf{X}; Y)$. Are they different?
- ❖ Recall the HW#0 problem on the joint distribution of U and V .
 - (a) For the first case where $p_1 = 0.1$ and $p_2 = 0.2$, find the following measures: $H(U)$, $H(V)$, $H(U|e_1)$, $H(V|e_2)$, $H(U|V)$, $H(V|U)$, $H(U, V)$, $I(U; V)$, $I(U; \theta)$, $I(V; \theta)$.
 - (b) Repeat for $p_1=0.01$ and $p_2 = 0.02$.
 - (c) Note there is a notable change in $I(U; V)$ between (a) and (b). Describe this change and make qualitative statements explaining the change. What would happen to $I(U; V)$ when p_1 and p_2 approach zero? What would happen if they both approach $1/2$.

Information Theory

2nd Module

Agenda

- ❖ Markov Chain and Entropy
- ❖ Sufficient Statistics
- ❖ Fano's Inequality
- ❖ Different Types of Convergences
- ❖ Asymptotic Equipartition Property
- ❖ High Probable Set vs. Typical Set
- ❖ Homeworks

Markov Chain

- ❖ Consider random variables X , Y , and Z .
- ❖ A chain of random variables $X \rightarrow Y \rightarrow Z$ is called Markov chain if

$$p(z | x, y) = p(z | y) .$$

- ❖ Note it implies $p(x, z | y) = p(x | y) p(z | x, y) = p(x | y) p(z | y)$.
 - The first equality is due to conditional probability.
 - The second is due to Markov chain.
 - Thus, a MC $X \rightarrow Y \rightarrow Z$ implies, conditional independence between X and Z knowing Y .
- ❖ ***Conditioning on current, future and past are independent.***

Data Processing Inequality

❖ If $X \rightarrow Y \rightarrow Z$, then $I(X; Y) \geq I(X; Z)$.

❖ Proof:

$$I(X; Y, Z) = I(X; Y) + I(X; Z | Y)$$

or
$$= I(X; Z) + I(X; Y | Z)$$

– We know $I(X; Z | Y) = 0$ and $I(X; Y | Z) \geq 0$. (why?)

– Thus, $I(X; Y) \geq I(X; Z)$

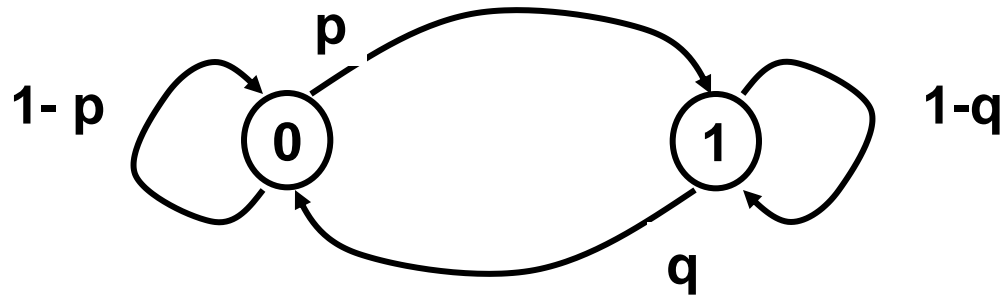
– Equality *iff* $I(X; Y | Z) = 0$, i.e., $X \rightarrow Z \rightarrow Y$ is a Markov chain.

❖ Let's use $Z := g(Y)$, a function of Y .

❖ The function implies an arbitrary data processing on Y .

❖ The inequality implies then any data processing will not help us understand X any better.

Markov Chain



- ❖ Consider a Markov chain, X_0, X_1, \dots, X_n
 - Transition matrix $\mathbf{P} = [1-p \ q ; p \ 1-q]$.
 - Initial distr. $\pi = [\alpha ; 1 - \alpha]$.
 - Stationary distr. $s_0 = q/(p+q)$, $s_1 = p/(p+q)$, $\mathbf{s} = [s_0; s_1]$.
 - $[\Pr\{X_1=0\}; \Pr\{X_1=1\}] = \mathbf{P} \pi$
 - $\Pr\{X_1=0\} = \Pr\{X_1=0|X_0=0\}\Pr\{X_0=0\} + \Pr\{X_1=0|X_0=1\}\Pr\{X_0=1\}$
 - $\Pr\{X_1=1\} = \Pr\{X_1=1|X_0=0\}\Pr\{X_0=0\} + \Pr\{X_1=1|X_0=1\}\Pr\{X_0=1\}$

Markov Chain and Entropy

- ❖ Distr. at any n is $\mathbf{t}_n := [\Pr\{X_n=0\}; \Pr\{X_n=1\}] = \mathbf{P}^n\boldsymbol{\pi}$
- ❖ The stationary distr. is $\mathbf{s} = \lim_{n \rightarrow \infty} \mathbf{t}_n$
 - Or, simply solve $\mathbf{s} = \mathbf{P}\mathbf{s}$.
- ❖ Ex) $p = 0.1$, $q=0.3$,
 $\mathbf{P} = [0.9 \ 0.3; 0.1 \ 0.7]$,
 $\mathbf{P}^\infty = [0.75 \ 0.75; 0.25 \ 0.25]$,
 $\mathbf{s} = [0.75; 0.25]$
- ❖ Consider the following cases
 - $\boldsymbol{\pi} \sim$ uniform, $\mathbf{s} \sim$ non-uniform: $H(\mathbf{t}_n)$ is decreasing toward $H(\mathbf{s})$
 - $\boldsymbol{\pi} \sim$ non-uniform, $\mathbf{s} \sim$ uniform: $H(\mathbf{t}_n)$ is increasing toward $H(\mathbf{s})$

The Second Law of Thermodynamics

- ❖ *Entropy of an isolated system is non-decreasing.*
- ❖ This comes from the notion that the micro states in a thermodynamic system reach equally likely states in equilibrium (uniform stationary distr.)
 - If started off with non-uniform initial distr., then, entropy increases.
 - If started off with uniform initial distr., then, entropy stays the same.

Sufficient Statistics

- ❖ Suppose an index set $\{\theta: 1, 2, \dots, n\}$ and a family of pmf's parameterized by θ , $\{f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_n(\mathbf{x})\}$.
- ❖ Let
 - \mathbf{X} be a sample from a distribution in this family and
 - $T(\mathbf{X})$ be a function of the sample (a statistic) for inference of θ .
- ❖ MC: $\theta \rightarrow \mathbf{X} \rightarrow T(\mathbf{X})$
- ❖ Thus, in general $I(\theta; \mathbf{X}) \geq I(\theta; T(\mathbf{X}))$.
- ❖ When the equality is achieved, we call $T(\mathbf{X})$ a sufficient statistic for inference on θ .
 - Basically, it implies that $T(\mathbf{X})$ contains all the information for θ .
 - No loss of information for θ .

Example on Sufficient Statistics

- ❖ Consider a sequence of coin tosses, X_1, X_2, \dots, X_n , iid with $X_i \in \{0,1\}$, with an unknown parameter $\theta = \Pr\{X_i = 1\}$.
- ❖ Given n , the number of 1's in n -trials is a *sufficient statistic* for θ .
 - $T(X_1, \dots, X_n) = \sum_{i=1}^n X_i$
 - $\Pr\{X_1=1, X_2=1, \dots, X_n=0, \text{ i.e. } k \text{ 1's}\} = \theta^k (1-\theta)^{n-k}$, for any $k \in \{0, 1, \dots, n\}$.
- ❖ Also $\hat{\theta} = \frac{T}{n}$ is the sufficient statistic for θ .
- ❖ Thus, we note that $\Pr\{X_1=x_1, X_2=x_2, \dots, X_n=x_n \mid T = k\}$
$$= \begin{cases} 1/(n \text{ choose } k) & \text{if } \sum_{i=1}^n x_i = k \\ 0 & \text{o.w.} \end{cases}$$
- ❖ θ is independent of the sequence $\{X_i\}$ given T . Thus, $\theta \rightarrow \{X_i, i=1, \dots, n\} \rightarrow T$ forms a MC. Thus, T is a sufficient statistic for θ .

Sufficient Statistics (2nd Ex)

❖ Other examples of sufficient statistics...

Fano's Inequality

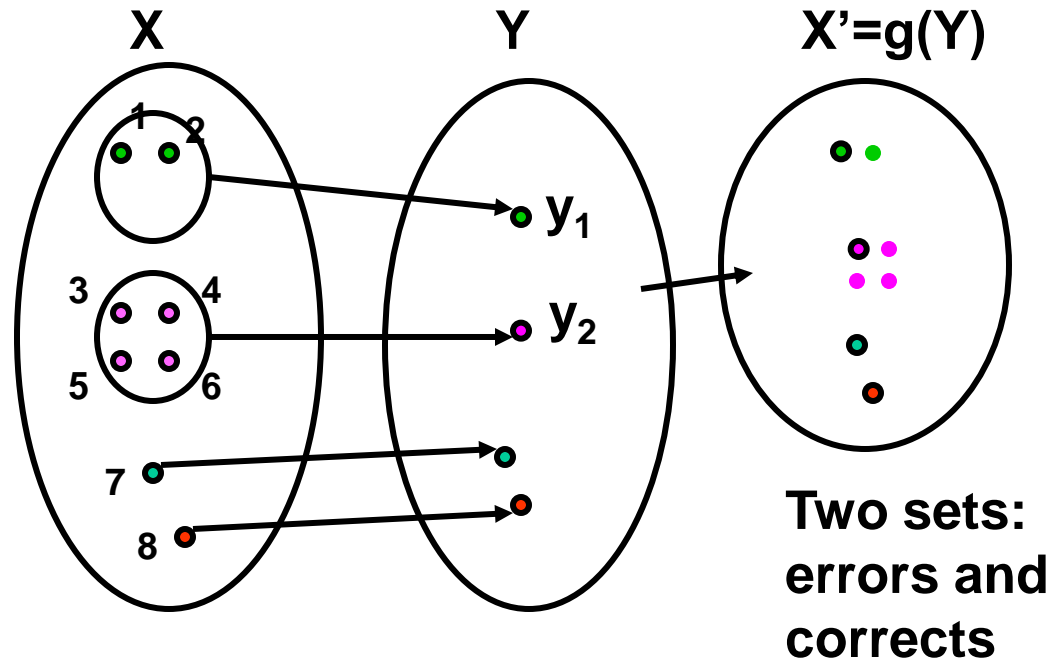
- ❖ Consider the problem of “send X , observe Y , and make a guess $g(Y)$ on X .”
- ❖ Note that $X \rightarrow Y \rightarrow X' = g(Y)$ forms a MC.
- ❖ FI relates the $P_e := \Pr\{X' := g(Y) \neq X\}$ with $H(X|Y)$.
- ❖ We know $H(X|Y) \geq 0$ with “=” iff X is func. of Y :

$$\Pr\{X'(Y) \neq X\} = 0 \text{ iff } H(X|Y) = 0$$

- ❖ Thus, we expect “*small* P_e for *small* $H(X|Y)$.”

Fano's Inequality

- ❖ A thought experiment
- ❖ y_1 observed: two possibilities on X
 - P_e is $1/2$
- ❖ y_2 observed: 4 possibilities on X
 - P_e is $3/4$
- ❖ We can divide the set $\{X = x\}$ into two disjoint sets
 - $\{X' = X\} = \{1, 3, 7, 8\}$
 - $\{X' \neq X\} = \{2, 4, 5, 6\}$



Fano's Inequality (2)

❖ $H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X|Y)$

❖ Or a weaker version is

$$1 + P_e \log|\mathcal{X}| \geq H(X|Y) \text{ or}$$

$$P_e \geq (H(X|Y) - 1)/\log|\mathcal{X}|$$

❖ Proof:

$$\text{Consider } E := \begin{cases} 1 & \text{if } X' \neq X \\ 0 & \text{o.w.} \end{cases}$$

$$\begin{aligned} \text{Chain rule gives } H(E, X | Y) &= H(X | Y) + H(E | X, Y) \\ &= H(E | Y) + H(X | Y, E) \end{aligned}$$

Fano's Inequality (3)

$$\begin{aligned}
 H(X | Y) + H(E | X, Y) &= H(E | Y) + H(X|Y, E) \\
 &\leq H(E) = H(P_e) \leq 1.0
 \end{aligned}$$

The last term can be bounded as

$$\begin{aligned}
 H(X|Y, E) &= \Pr\{E=1\} H(X|Y, E=1) + \Pr\{E=0\} H(X|Y, E=0) \\
 &= P_e \sum_y p(y) H(X|Y=y, E=1)
 \end{aligned}$$

---- But, we know $H(X|Y=y, E=1) \leq \log(|\mathcal{X}| - 1)$

for any y (There is at least one ω $X'(\omega) = X(\omega)$)

$$\leq P_e \log(|\mathcal{X}| - 1)$$

Therefore,

$$H(X|Y) \leq H(P_e) + P_e \log(|\mathcal{X}| - 1) \leq 1 + P_e \log(|\mathcal{X}| - 1) \quad \text{Q.E.D.}$$

Types of Convergences

❖ *In distribution*: $X_n \Rightarrow X$ in distribution if

$$F_n(x) = \Pr\{X_n \leq x\} \rightarrow F(x) = \Pr\{X \leq x\} \text{ as } n \rightarrow \infty$$

– *Ex*) Let X_1, X_2, \dots iid fair binary $\{-1, +1\}$ rvs. Let $S_n = (1/\sqrt{n}) \sum_{i=1}^n X_i$. Then, $F_n(y) := \Pr(S_n \leq y) \rightarrow \mathcal{N}(0, 1)$ (C.L.T.).

❖ *In probability*: $X_n \Rightarrow X$ in probability as $n \rightarrow \infty$ if $\forall \varepsilon > 0$

$$\Pr\{\omega: |X_n(\omega) - X(\omega)| > \varepsilon\} \rightarrow 0 \text{ as } n \rightarrow \infty$$

❖ *In almost sure, almost everywhere sense, or with prob. 1*:

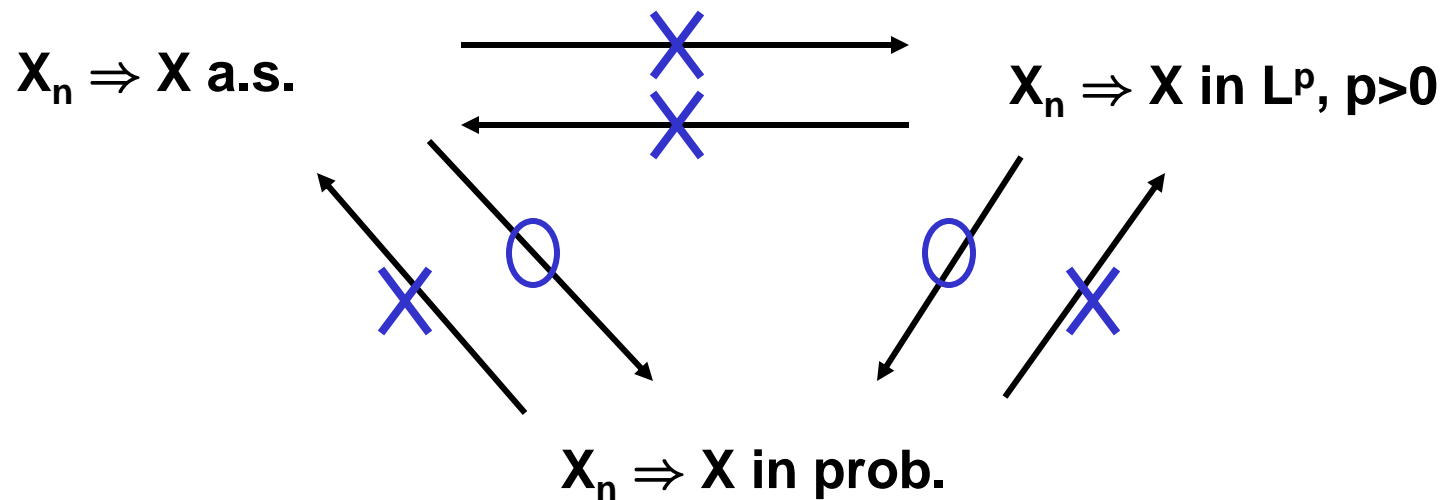
$X_n \Rightarrow X$ a.s. as $n \rightarrow \infty$, if

-- $\Pr\{\omega: \lim X_n(\omega) = X(\omega)\} = 1$, or

-- For $\forall \varepsilon$, $\Pr\{\omega: |X_n(\omega) - X(\omega)| > \varepsilon, \text{ i.o.}\} = 0$, as $n \rightarrow \infty$

❖ *In L^2* : $X_n \Rightarrow X$ in L^2 , if $E\{|X_n - X|^2\} \rightarrow 0$, as $n \rightarrow \infty$

Relationship Between Different Types



Richard Durrett, Probability: Theory and Examples, 1991, Wadsworth

“ $X_n \Rightarrow X$ a.s.” \Rightarrow “ $X_n \Rightarrow X$ in prob.”

❖ $X_n \Rightarrow X$ a.s. implies that for $\forall \varepsilon > 0$

$$\lim_{k \rightarrow \infty} \mathbf{P}\{\bigcup_{n \geq k} [|\mathbf{X}_n - \mathbf{X}| > \varepsilon]\} = 0$$

❖ Since $\{|\mathbf{X}_k - \mathbf{X}| > \varepsilon\} \subseteq \bigcup_{n \geq k} \{|\mathbf{X}_n - \mathbf{X}| > \varepsilon\}$,

$$\Pr\{|\mathbf{X}_k - \mathbf{X}| > \varepsilon\} \leq \Pr(\bigcup_{n \geq k} \{|\mathbf{X}_n - \mathbf{X}| > \varepsilon\})$$

❖ Taking the limit on both sides,

$$\lim_{k \rightarrow \infty} \Pr\{|\mathbf{X}_k - \mathbf{X}| > \varepsilon\} \leq \lim_{k \rightarrow \infty} \Pr(\bigcup_{n \geq k} \{|\mathbf{X}_n - \mathbf{X}| > \varepsilon\}) = 0$$

Q.E.D.

$X_n \Rightarrow X$ in prob. $\not\Rightarrow X_n \Rightarrow X$ a.s.
 (Converse is not true)

❖ Consider a series of r.v.'s $X_n := 1_{A_n}$ where A_n are defined as

$$A_1 = [0, 1];$$

$$A_2 = [0, 1/2), A_3 = [1/2, 1];$$

$$A_4 = [0, 1/4), A_5 = [1/4, 1/2), A_6 = [1/2, 3/4), A_7 = [3/4, 1];$$

...

❖ Let $\Pr\{X_n = 1\} = \text{length}(A_n)$ (Lebesgue)

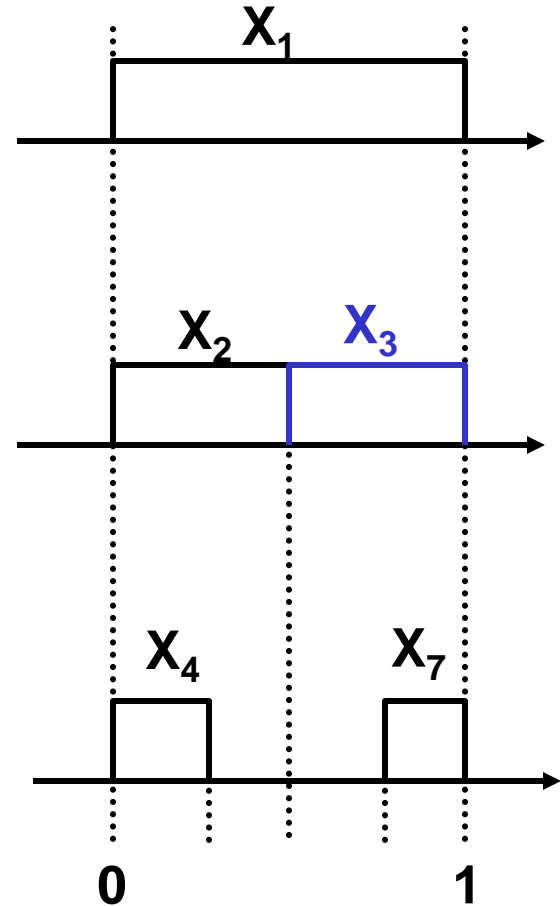
❖ Now, let $X = 0$. Then,

❖ For $\forall \varepsilon > 0$, $\Pr(|X_n - X| > \varepsilon) \rightarrow 0$ as $n \rightarrow \infty$

❖ But, $\{\omega: \lim X_n(\omega) = X(\omega)\} = \emptyset$

Thus, $\Pr\{\omega: \lim X_n(\omega) = X(\omega)\} = 0$.

Q.E.D.



Example for both “in prob.” and “a.s.”

- ❖ Consider a series of r.v. $X_n = 1_{A_n}$ where $A_1 = [0, 1]$; $A_n = [0, 1/n]$, with the Lebesgue measure as the prob.
- ❖ Let $X = 0$.
- ❖ With this example, we note that $X_n \Rightarrow X$ in both “in prob” and “a.s.” senses

Laws of Large Numbers

❖ *Weak Law* of Large Numbers: Let X_1, X_2, \dots be i.i.d. with $E|X_1| < \infty$ and $E\{X_1\} = \mu$, and as $n \rightarrow \infty$,

$$S_n/n \Rightarrow \mu \text{ in probability}$$

where $S_n = X_1 + X_2 + \dots + X_n$.

❖ *Strong Law* of Large Numbers: $S_n/n \Rightarrow \mu$ **a.s.** as $n \rightarrow \infty$.

– That is, it is in fact a.s.

❖ L^2 Weak Law: Let X_1, X_2, \dots, X_n be uncorrelated r.v.'s with $E\{X_i\} = \mu$ and $\text{var}(X_i) \leq C < \infty$. Then, as $n \rightarrow \infty$

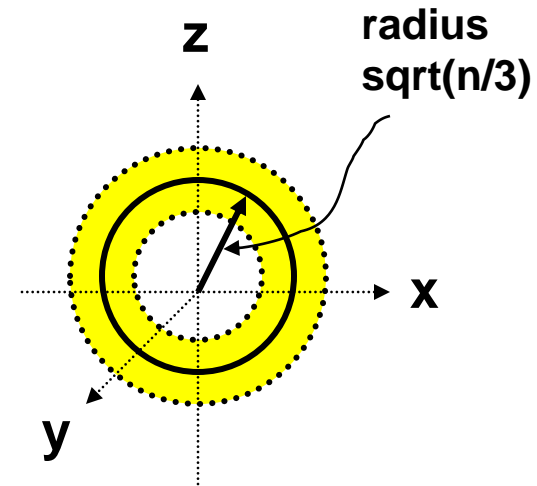
$$S_n/n \Rightarrow \mu \text{ in } L^2$$

Surface Hardening

- ❖ A high-dimensional cube $[-1, 1]^n$ is almost the boundary of a ball.
- ❖ Let X_1, X_2, \dots be independent uniformly distributed on $[-1, 1]$.
 - Then, $EX_i^2 = 1/3$.
- ❖ Then, the WLLN implies

$$(X_1^2 + \dots + X_n^2)/n \rightarrow 1/3 \text{ in probability as } n \rightarrow \infty$$
- ❖ Consider an n -dimensional random vector $\mathbf{X} := (X_1, \dots, X_n)$, and its length $\|\mathbf{X}\| = \text{sqrt}(X_1^2 + \dots + X_n^2)$
- ❖ Thus, for $\forall \epsilon > 0$, you can always find a large enough n , such that $\Pr\{|\|\mathbf{X}\|^2/n - 1/3| > \epsilon\} \approx 0$
- ❖ $\Pr\{\mathbf{X} \in \mathbb{R}^n: 1/3 - \epsilon < \|\mathbf{X}\|^2/n < 1/3 + \epsilon\} \approx 1$

$$\Pr\{\mathbf{X} \in \mathbb{R}^n : \sqrt{n(1/3 - \epsilon)} < \|\mathbf{X}\| < \sqrt{n(1/3 + \epsilon)}\} \approx 1$$



$$\begin{aligned} \text{Length}^2 &= \text{norm}^2 \\ &= \sum x_i^2 \end{aligned}$$

Asymptotic Equi-partition Property

❖ Let X_1, X_2, \dots , i.i.d. with $p(x)$.

❖ The **sample entropy**

$$- H_n' = - (1/n) \log p(X_1=x_1, \dots, X_n=x_1) = - (1/n) \sum_i \log p(X_i=x_i)$$

Converges in prob. to

the true entropy $H(X) = - \sum_i p(x_i) \log p(X_1=x_i)$.

❖ As $n \rightarrow \infty$, Ω can be divided into two mutually exclusive sets: The **typical set** and the non-typical set.

– The sequences in the typical set have the sample entropy $\approx H(X)$

– Those in the non-typical set have the sample entropy $\neq H(X)$

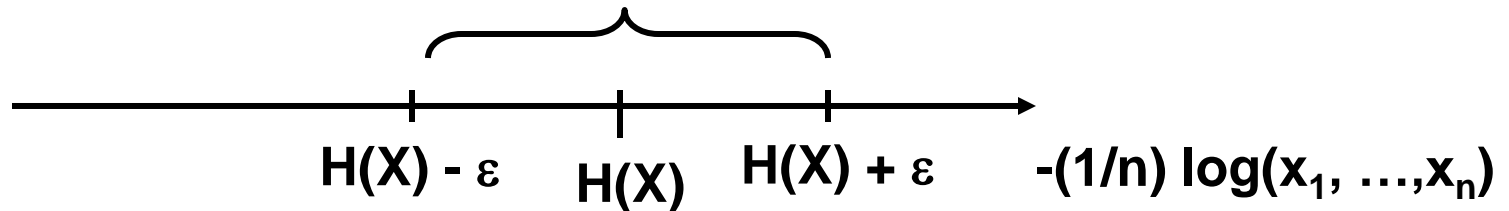
❖ From WLLN, $\Pr\{\text{Typical set}\} \approx 1.0$ as $n \rightarrow \infty$

Asymptotic Equi-partition Property (2)

- ❖ AEP: If X_1, X_2, \dots iid with $p(x)$, then
$$H_n' := - (1/n) \log p(X_1, X_2, \dots, X_n) = - (1/n) \sum_i \log p(X_i)$$
$$\Rightarrow - E(\log p(X_1)) = H(X) \text{ in prob.}$$

(due to WLLN)
- ❖ This means, for $\forall \varepsilon > 0$
$$\Pr\{(x_1, \dots, x_n): | H_n' - H(X) | > \varepsilon \} \rightarrow 0 \text{ as } n \rightarrow \infty$$
 - Prob. of the *atypical* set goes to zero
 - Prob. of the *typical* set goes to 1
- ❖ We can divide the entire set Ω , the set of all possible sequences of length n , into two mutually exclusive sets
 - Typical set $A_\varepsilon^{(n)} := \{(x_1, \dots, x_n): | H_n' - H(X) | \leq \varepsilon \}$
 - Atypical set $\Omega - A_\varepsilon^{(n)}$

A sequence in the Typical Set $A_\varepsilon^{(n)}$



- ❖ For any sequence $(x_1, \dots, x_n) \in A_\varepsilon^{(n)} := \{(x_1, \dots, x_n) : |-(1/n) \log p(x_1, \dots, x_n) - H(X)| \leq \varepsilon\}$, the prob. of the sequence must have the following property

$$|-(1/n) \log p(x_1, \dots, x_n) - H(X)| \leq \varepsilon$$

$$H(X) - \varepsilon \leq -(1/n) \log p(x_1, \dots, x_n) \leq H(X) + \varepsilon$$

$$2^{-n(H(X)+\varepsilon)} \leq p(x_1, \dots, x_n) \leq 2^{-n(H(X) - \varepsilon)}$$

- ❖ Since we can choose a very small ε , the prob. of a sequence can be made very close to $2^{-nH(X)}$, as $n \rightarrow \infty$.

$\Pr\{A_\varepsilon^{(n)}\} > 1 - \varepsilon$, for n sufficiently large

- ❖ For any $\varepsilon > 0$ and $\delta > 0$, there exists an n_0 such that $n > n_0$,
 $\Pr\{ | -(1/n) \log[p(x_1, \dots, x_n)] - H(X) | \leq \varepsilon \} > 1 - \delta$.
- ❖ Choose $\delta = \varepsilon$.

The Size of the Typical Set $A_\varepsilon^{(n)}$

❖ The size of the typical set satisfies

1. $|A_\varepsilon^{(n)}| \leq 2^{n(H(X) + \varepsilon)}$

2. $(1-\varepsilon) 2^{n(H(X) - \varepsilon)} \leq |A_\varepsilon^{(n)}|$

❖ Proof of 1:
$$\begin{aligned} 1 &= \sum_{\mathbf{x} \in \mathcal{X}^n} p(\mathbf{x}) \\ &\geq \sum_{\mathbf{x} \in A_\varepsilon} p(\mathbf{x}) \\ &\geq \sum_{\mathbf{x} \in A_\varepsilon} 2^{-n(H(X) + \varepsilon)} \\ &= |A_\varepsilon^{(n)}| 2^{-n(H(X) + \varepsilon)} \quad \text{Q.E.D.} \end{aligned}$$

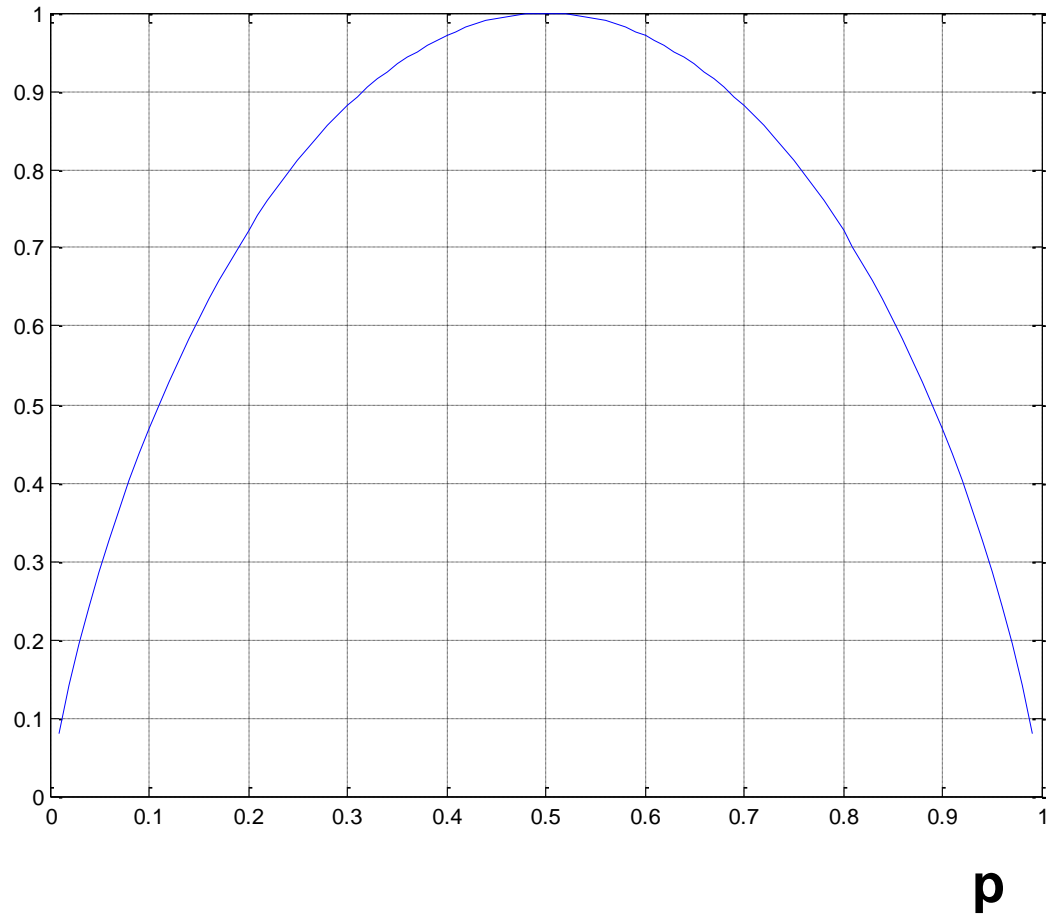
❖ Proof of 2: $1 - \varepsilon \leq \Pr\{A_\varepsilon^{(n)}\} \leq |A_\varepsilon^{(n)}| 2^{-n(H(X) - \varepsilon)}$

Example

- ❖ $X_1 \sim$ binary r.v. taking 1 or 0, with prob. p and $(1-p)$
- ❖ Let X_1, X_2, \dots, X_n i.i.d.
- ❖ Ex. with $n=6, p=2/3$
 - The most typical sequences have 4 ones ($np = 6 \cdot 2/3 = 4$).
 - The prob. of any sequence with 4 ones is $p^4 (1-p)^2$. There are $\binom{6}{4}$ number of such sequences.
 - There are total of 2^6 possible sequences.
- ❖ We can divide the complete set into the typical and the non-typical sets.
- ❖ In a trial, the sequences in the non-typical set occur rarely while those in the typical set occur very often.

$$H(p)$$

$H(p)$



Example (2)

❖ Consider $p = 0.5$

- Then, we note $H(X) = 1$; the size of typical set is 2^6 ; each and every sequences happens equally likely with prob. $1/2^6$

Example (3) at $n=6$

- ❖ Consider $p = 0.061$, $H(0.061) = 0.33$; the size of typical set is $2^{6*(0.33+1/6)} = 7.88$; compared to $2^6 = 64$
- ❖ A sequence in the typical set is expected to have $np = 0.061*6 = 0.37$ number of 1's
- ❖ Exact calculation:
 - a seq. with no 1: $(1-p)^6 = 0.6855$
(The most probable sequence and also most typical)
 - seq.'s with a single 1: $C_1^6(1-p)^5 p = (6) 0.0445 = 0.2672$
 - These two kinds of sequences (7 seq's) account for 95% occurrences.

Example (4): n=10

❖ Consider $n = 10$ with $p = 0.14$. Then, $H(0.14) = 0.58$; $nH = 5.8$; the size of typical set is $2^{10 \cdot (0.58 + 1/10)} \approx 111$;
 prob. = $(1/111) = 0.009$; $2^{10} = 1024$

❖ Exact calculation:

- a seq. with no 1:

$$(1-p)^{10} = 0.22$$

- seq.'s with a single 1:

$$C_{1}^{10} (1-p)^9 p = 0.036 \times 10 = 0.36$$

- seq.'s with two 1's: $C_{2}^{10} (1-p)^8 p^2 = (45) 0.0059 = 0.27$

85%

- seq.'s with three 1's: $C_{3}^{10} (1-p)^7 p^3 = (120) 9.5e-4 \times 120 = 0.11$

96%

- size of the 96% occurrence set is $1 + 10 + 45 + 120 = 176$

Most typical set

Most probable

Example (5): $n = 100$

❖ Consider $n = 100$ with $p = 0.02$. Then, $H(0.02) = 0.1414$; $nH \approx 14$; the size of typical set is $2^{14} \approx 18054$; $\text{prob.} = 1/(18K) = 5.538e-5$; $2^{100} = (1024)^{10}$

❖ Exact calculation:

- a seq. with no 1: $(1-p)^{100} = 0.1326$
- seq.'s with a single 1: $(1-p)^{99} p = 0.0027$, (x 100) = 0.27
- seq.'s with two 1's: $(1-p)^{98} p^2 = 5.25e-5$, (x 4950) = 0.2734
- seq.'s with three 1's: $(1-p)^{97} p^3 = 1.12e-6$, (x 161700) = 0.1823
- seq.'s with four 1's: $(1-p)^{96} p^4 = 2.3e-8$, (x 3.9M) = 0.09
- size of the 95% occurrence set is about 4 Million

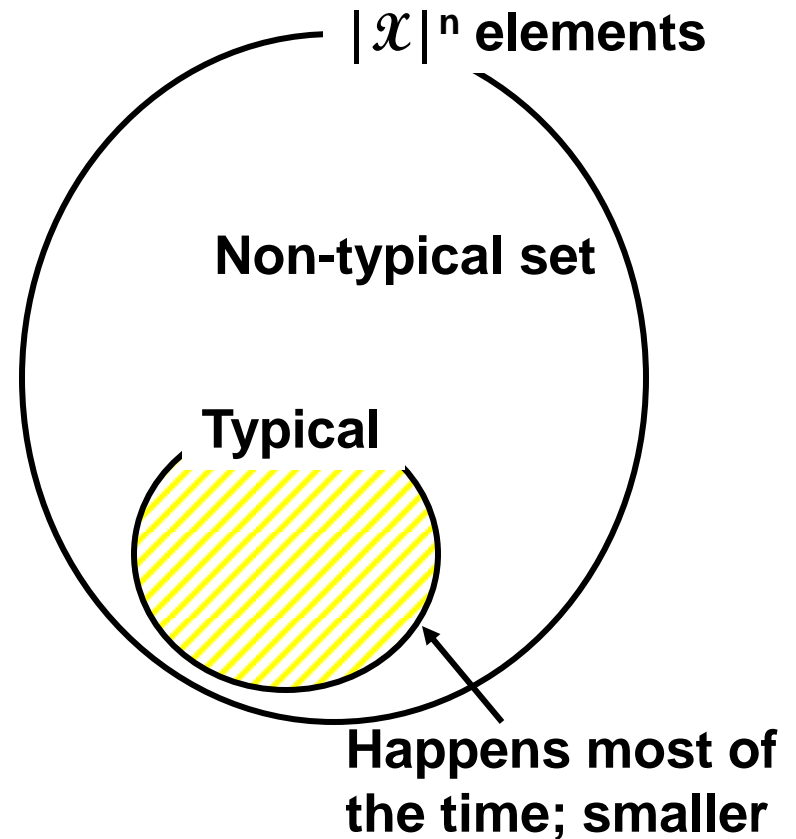
Most probable

95%

Most typical set

Consequences of AEP: Data Compression

- ❖ The size of the typical set is $2^{n(H(X) + \epsilon)}$
- ❖ Data Compression Scheme:
- ❖ Seq.'s in typical set: In general, we need $(nH(X) + \epsilon) + 1$ bits to represent them
 - Let's use 0 as prefix to denote membership to the typical set
 - $n(H(X) + \epsilon) + 2$ bits in total
- ❖ Seq.'s in atypical set:
 - $n \log_2 |\mathcal{X}| + 1$ bits (Use prefix 1)



High Probability Sets and the Typical Sets

- ❖ Typical set is a small set that accounts for the most of the probability.
- ❖ But, is there a set smaller than the typical set, that accounts for the most of the probability?
- ❖ Theorem 3.3.1 states that the size of the typical set is the same as the size of the high probability set, to the first order in the exponent
 - The proof is easy, and outlined in prob. 3.11

High Probability Sets and the Typical Sets

❖ High Probability Set $B_\delta^{(n)} \subset \mathcal{X}^n$ is defined as a set

$$\Pr\{B_\delta^{(n)}\} \geq 1 - \delta, \quad \text{for } 1/2 > \delta > 0.$$

❖ The theorem indicates that the size of this set is

$$\lim_{n \rightarrow \infty} (1/n) \log (|B_\delta^{(n)}|/|A_\delta^{(n)}|) = 0$$

❖ At a finite n , $(1/n) \log (|B_\delta^{(n)}|/|A_\delta^{(n)}|) = \varepsilon > 0$

$$|B_\delta^{(n)}| = |A_\delta^{(n)}| 2^{n\varepsilon}$$

– Both sizes grow exponentially fast

– But the exponent of the growth is linear, nH

❖ Using Example (5), we note that the most probable set must include the all 0 sequence by definition; but the typical set may not include it (the most typical set include all the sequences with two ones).

Homework #2, #3

❖ HW#2

- P2.6 (Conditional vs. unconditional mutual information)
- P2.23 (Conditional MI)
- P2.26 (Relative entropy is non negative)
- P2.29 (Inequalities)
- P2.34 (Entropy of initial condition)
- P2.40 (Discrete Entropies)
- P2.43 (MI of heads and tails)
- P2.48 (Sequence length)

❖ HW#3

- P2.21 (Markov inequality)
- P2.30 (Maximum entropy)
- P2.32, P2.33 (Fano's inequality)
- P3.1 (Markov and Chebyshev inequalities)
- P3.2 (AEP and MI)
- P3.4 (AEP)
- P3.10 (Random box size)
- P3.13 (Calculation of typical set) Note the table on pg. 69 might have some errors. Generate your own and do the problem.

❖ Challenge Problems, not required:

1. Let X_1, X_2, \dots, X_n be uncorrelated r.v.'s with $E\{X_i\} = \mu$ and $\text{var}(X_i) \leq C < \infty$. Prove $(X_1 + \dots + X_n)/n \Rightarrow \mu$ in L^2 as well as in probability. (Hint: prove L^2 first and use Chebyshev's inequality)
2. Find a tight lower bound on the probability of making errors for BPSK signaling over the AWGN channel. (Hint: Fano's inequality)
 1. Obtain the bound expression
 2. Numerically evaluate your expression
 3. Draw the bound on P_e as the function of E_b/N_o . (Draw only from -1 to 5 dB SNR in the interval of 0.5 dB)

Information Theory

3rd Module

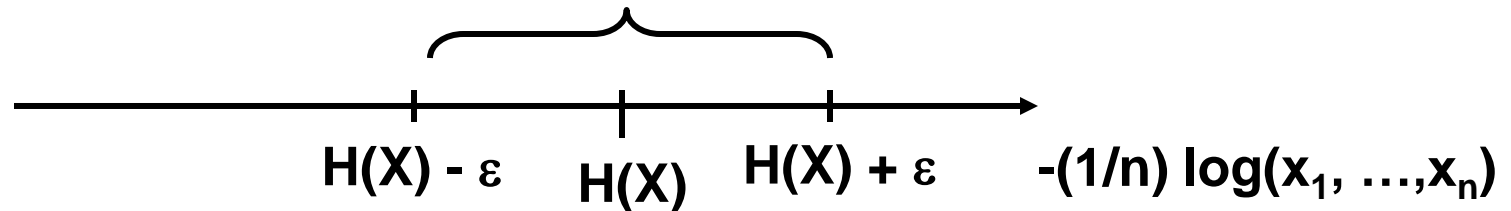
Agenda

- ❖ Entropy Rates of a Stochastic Process (Chapter 4)
- ❖ Compression (Chapter 5)

Tentative Schedule

<i>Weekly Course Schedule</i>		
<i>Calendar</i>	<i>Description</i>	<i>*Remarks</i>
1st week, 9/2	<i>Introduction to Information Theory, Entropy</i>	
2nd week, 9/7, 9	Entropy, Relative Entropy and Mutual Information	
3rd week, 9/14, 16	Entropy, Relative Entropy and Mutual Information	
4th week, 9/2,23	Asymptotic Equipartition Property	
5th week, 9/28, 30	Asymptotic Equipartition Property/Entropy Rates of a Stochastic Process	
6th week, 10/5, 7	Entropy rates of Markov Chain	
7th week, 10/12, 14	Data compression	
8th week, 10/19, 21	Data compression/Channel capacity	Midterm 10/21
9th week, 10/26, 28	Channel capacity theorems/forward/reverse	
10th week, 11/2, 4	Differential entropy	
11th week, 11/9, 11	Gaussian channel capacity	Selection of papers due
12th week, 11/16, 18	MIMO channel capacity theorem	보충?
13th week, 11/23, 25	Multiple access channel capacity theorem	
14th week, 11/30, 12/2	Slepian Wolf	
15th week, 12/6, 9	Student Presentation	
16th week, 12/16	Final Exam (12/16 일)	

A sequence in the Typical Set $A_\varepsilon^{(n)}$



- ❖ For any sequence $(x_1, \dots, x_n) \in A_\varepsilon^{(n)} := \{(x_1, \dots, x_n) : |-(1/n) \log_2[p(x_1, \dots, x_n)] - H(X)| \leq \varepsilon\}$, the prob. of the sequence must have the following property
- ❖ $|-(1/n) \log_2[p(x_1, \dots, x_n)] - H(X)| \leq \varepsilon$
- ❖ $H(X) - \varepsilon \leq -(1/n) \log_2[p(x_1, \dots, x_n)] \leq H(X) + \varepsilon$
- ❖ $2^{-n(H(X)+\varepsilon)} \leq p(x_1, \dots, x_n) \leq 2^{-n(H(X) - \varepsilon)}$
- ❖ Since we can choose a very small ε , the prob. of a sequence can be made very close to $2^{-nH(X)}$, as $n \rightarrow \infty$.

Basically, AEP tells us that

- ❖ *For length n i.i.d. sequence of r.v.'s*
- ❖ nH bits are good enough for describing the typical sequence.
 - Size of the typical message set is $2^{nH(X)}$.
 - Each sequence in the message set occurs with $2^{-nH(X)}$.
- ❖ In an experiment, usually a sequence in the typical set happens.
- ❖ Shannon's Theorem 3 is AEP (see [page 13](#)).
- ❖ But what happens if the r.v.'s are dependent?
 - Motivation to consider the *entropy rate*

Shannon's Paper

- ❖ Shannon uses Markov chain to describe English.
 - [Shannon's 1948 paper](#)
 - Zero-order, first-order, second-order letters
 - First-order word, second-order word
 - Let's take a look at his paper.

- ❖ “Can we define a quantity which will measure, in some sense, how much information is “produced” by such a process, or better, at what rate information is produced?”

Some Definitions for MC

❖ Stationary stochastic process:

$$\Pr(\mathbf{X}_1 = \mathbf{x}_1, \dots, \mathbf{X}_n = \mathbf{x}_n) = \Pr(\mathbf{X}_{t+1} = \mathbf{x}_1, \dots, \mathbf{X}_{t+n} = \mathbf{x}_n) \text{ for all } t.$$

❖ MC

$$- \Pr(\mathbf{X}_{n+1} = a \mid \mathbf{X}_n = b, \dots, \mathbf{X}_1 = \mathbf{x}_1) = \Pr(\mathbf{X}_{n+1} = a \mid \mathbf{X}_n = b)$$

❖ MC is *time-invariant* (almost always we assume this) if

$$P(\mathbf{X}_{n+1+t} = a \mid \mathbf{X}_{n+t} = b) = P(\mathbf{X}_{n+1} = a \mid \mathbf{X}_n = b)$$

- Transition matrix \mathbf{P} stays the same.

- *Stationary* distribution: $\mathbf{s} = \mathbf{P}\mathbf{s}$

❖ If the initial distribution is \mathbf{s} , then the MC is stationary.

Per-Symbol Entropy (*Entropy Rate*)

- ❖ Consider a sequence of r.v.s X_1, X_2, \dots, X_n
- ❖ How does the **entropy** of a sequence of n r.v.'s grow with n ?
- ❖ Let's define the *per symbol entropy*
 $H(\mathcal{X}) := \lim_{n \rightarrow \infty} (1/n) H(X_1, \dots, X_n)$ when it exists
- ❖ Examples:
 - ❖ When X_1, X_2, \dots are iid, the rate attains the *maximum* $H(X_1)$.
 $H(\mathcal{X}) = \lim_{n \rightarrow \infty} (1/n) H(X_1, \dots, X_n) = nH(X_1)/n = H(X_1)$.
 - ❖ When X_1, X_2, \dots are indep. but different distr.
Then, $H(X_1, \dots, X_n) = \sum_i H(X_i)$.

Conditional Entropy Rate

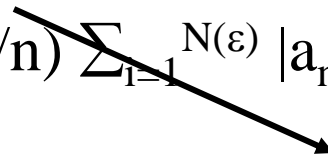
- ❖ $H'(\mathcal{X}) := \lim_{n \rightarrow \infty} H(X_n | X_{n-1}, \dots, X_1)$
- ❖ *For a stationary process, $H'(\mathcal{X}) = H(\mathcal{X})$, both limits exist and equal.*
- ❖ $H(X_{n+1} | X_n, \dots, X_1) \leq H(X_{n+1} | X_n, \dots, X_2)$
 - why?
 - $= H(X_n | X_{n-1}, \dots, X_1)$
 - why?
- ❖ We know that
 - It is a non increasing series of non-negative numbers.
 - It is bounded from below.
- ❖ Then, the limit exists. (convergence from above)

Cesaro Mean (from Analysis):

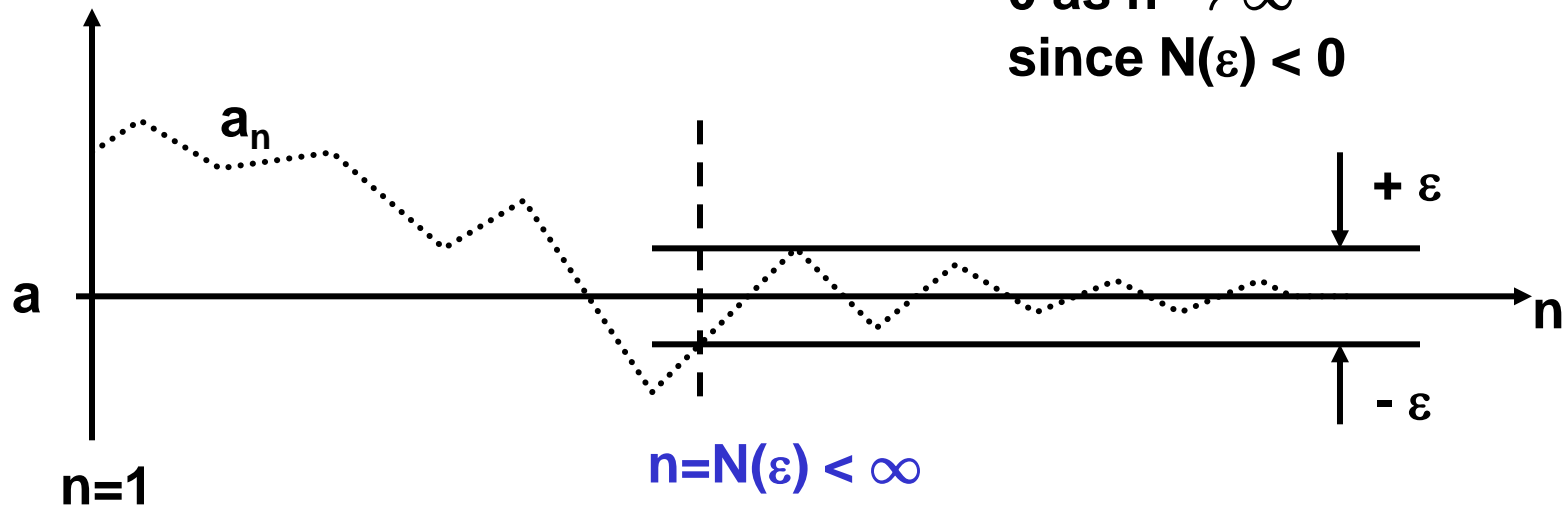
❖ Cesaro Mean (from Analysis):

If $a_n \rightarrow a$ and $b_n = (1/n) \sum_{i=1}^n a_i$, then $b_n \rightarrow a$

$$|b_n - a| \leq (1/n) \sum_{i=1}^n |a_i - a| \leq (1/n) \sum_{i=1}^{N(\varepsilon)} |a_i - a| + \varepsilon$$



**0 as $n \rightarrow \infty$
since $N(\varepsilon) < \infty$**



$H(X) = H'(X)$ for stationary process

❖ From the chain rule:

$$(1/n) H(X_1, \dots, X_n) = (1/n) \sum_{i=1}^n H(X_i | X_{i-1}, \dots, X_1)$$

❖ By applying the Cesaro Mean, we know

$$H(\mathcal{X}) = \lim (1/n) H(X_1, \dots, X_n) = \lim H(X_n | X_{n-1}, \dots, X_1) = H'(\mathcal{X})$$

❖ Implications: For a stationary process,

- There are about $2^{nH(\mathcal{X})}$ typical sequences of length n .
- The prob. of typical set is close to 1.
- $nH(\mathcal{X})$ bits are usually needed to represent the length n typical sequences.

Entropy Rate of a stationary MC

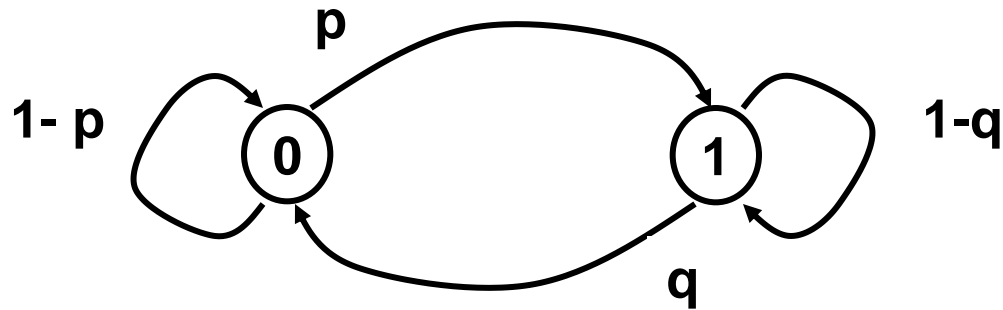
$$\begin{aligned} \blacklozenge \quad H(\mathcal{X}) &= H'(\mathcal{X}) = \lim_{n \rightarrow \infty} H(X_n | X_{n-1}, \dots, X_1) \\ &= \lim_{n \rightarrow \infty} H(X_n | X_{n-1}) \\ &= H(X_2 | X_1) \end{aligned}$$

\blacklozenge Let vector \mathbf{s} denote the stationary distribution and \mathbf{P} the transition matrix of a stationary MC.

$$\mathbf{s} = \mathbf{P}\mathbf{s}$$

$$H(\mathcal{X}) = H(X_2 | X_1) = \sum_i s_i (- \sum_j P_{ij} \log P_{ij})$$

Entropy rate of two state MC



- ❖ $\mathbf{s} = [s_0 \ s_1]'$; $\mathbf{P} = [1-p \ p; 1-q \ q]$;
- ❖ $H(\mathcal{X}) = H(X_2|X_1) = \sum_i s_i (-\sum_j P_{ij} \log P_{ij})$
- ❖ Then, the entropy rate $H(\mathcal{X}) = (q \cdot H(p) + p \cdot H(q)) / (p+q)$

Shannon's Examples

- ❖ See section 7 of his paper.
- ❖ He was interested in finding the entropy rate of an information source (English).
 - How much redundancy is there in the source?
 - Redundancy in English ≈ 0.5 .

 - This is my example.
 - I _a_t _o _o h_m_ _nd p_ly _it_ m_ k_d_.
 - is it possible to make out the meaning?
 - Deleted about 13 characters ($13/40 = 33\%$)

Answer to my example

- I want to go home and play with my kids. (40 char's and spaces)

❖ Chapter 5: Data Compression

Information generated from a source can be compressed without distortion.

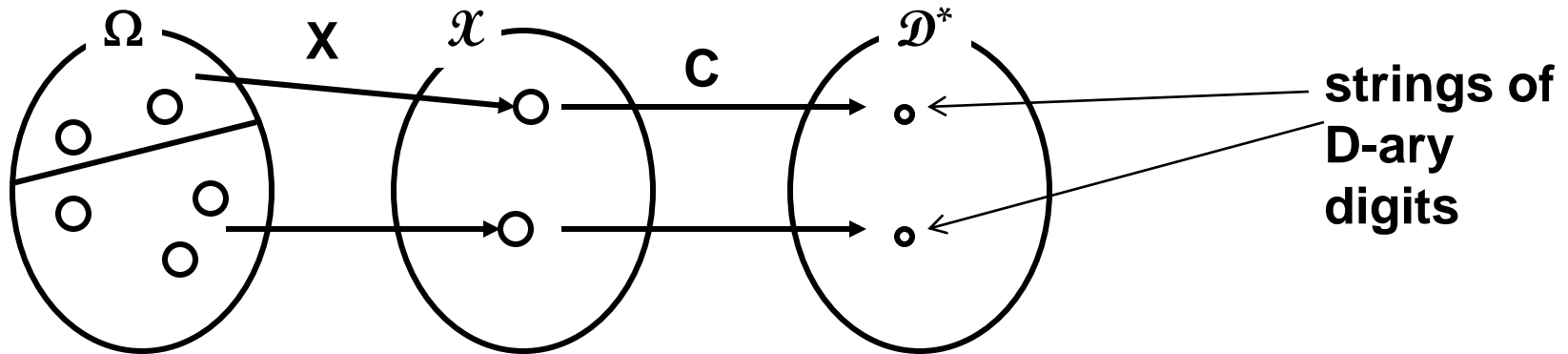
In this chapter, we are interested in distortion-less *Source Compression*.

Data Compression (Distortionless)

- ❖ Intuition tells us that we would want to use
 - Short description for frequent outcomes
 - Long description for less frequent outcomes
- ❖ A code constructed this way will have a small minimum description length.
- ❖ A source code does this efficiently.
- ❖ Information sources include data sources such as type writer, signals and telegraph.
- ❖ Shortest description length of a random source
 - A variable/a process \sim Entropy/Entropy rate
- ❖ Expected description length \geq Entropy

A code is a map

- ❖ A source code C for a random variable X is a *mapping* from $\mathcal{X} = \{X(\omega): x_1, x_2, \dots\}$ to \mathcal{D}^* , the set of codewords
 - A codeword $C(x)$ is a finite length string of D -ary digits assigned to x .



- ❖ Let $l(x)$ be the length of $C(x)$.
- ❖ The expected length $L(C) = E\{l(X)\} = \sum_{x_i} l(x_i) p(x_i)$.

A string of a D -ary digit

- ❖ D -ary alphabet is $\mathcal{D} = \{0, 1, 2, \dots, D-1\}$.
- ❖ Ex) 012201 is a ternary alphabet string.

Non-singular code

❖ A code is said to be *non-singular* if every element of \mathcal{X} maps into a different string in \mathcal{D}^* , i.e.,

$$x_i \neq x_j \quad \Rightarrow \quad C(x_i) \neq C(x_j)$$

- This statement does not allow many-to-one mapping.
- Mapping—by the definition—is either only one-to-one or many-to-one (no one-to-many).
- Thus, a map (a code) can include either one-to-one or many-to-one assignments.
- Hence, a *non-singular* code is a code that does not allow any many-to-one assignment in a map.
- Non-singularity thus implies the one-to-one mapping which is sufficient for unambiguous decoding.

Some Definitions on Functions

- ❖ A function $f: A \rightarrow B$ is a relation between A and B satisfying the following conditions:
 - For *each* $a \in A$, there *exists* $b \in B$ such that $(a, b) \in f$, AND
 - If (a, b) and (a, c) are in f , then $b = c$.

- ❖ A function $f: A \rightarrow B$ is said to be
 - One-to-one if given $b \in B$, there is at most one $a \in A$.
 - Onto if for each $b \in B$, there is at least one $a \in A$, i.e. $b=f(a)$
 - One-to-one correspondence if a function is both one-to-one and onto.

Extension C^ of C is a map*

❖ An *extension* C^* of a code C is a mapping from finite length strings of \mathcal{X} to finite length strings of \mathcal{D} , defined by

$$C(x_1x_2\dots x_n) = C(x_1)C(x_2) \dots C(x_n)$$

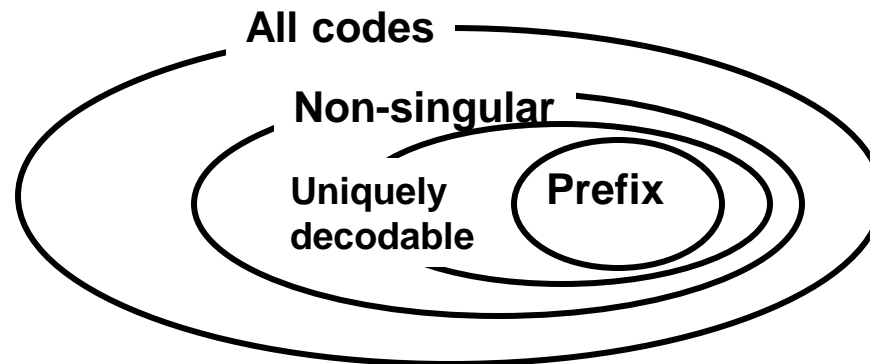
- It is *concatenation* of codewords.
- Ex) If $C(x_1) = 00$ and $C(x_2) = 11$, then $C(x_1x_2) = 0011$
- Why? Efficiency! It can get rid of a space symbol which would have been needed for separating any pair of different length codewords.
- Ex) $C(x_1)=11$, $C(x_2)=10$, $C(x_3) = 110$, $C(x_4) = 01$
 - $110110 \sim 110, 110$ or $11,01,10$ (not decodable)
 - Space symbols useful but wasteful!
 - Not efficient!

Uniquely Decodable Code

- ❖ A code C is called *uniquely decodable* if its k -th *extension* is one-to-one mapping from \mathcal{X}^k to \mathcal{D}^* for all $k \geq 0$.
 - (see P5.21, not P5.18)
 - Uniquely decodable = non singular when extended

Prefix Code

- ❖ A code C is called a *prefix(-free)* code or an *instantaneous* code if no codeword is a prefix of any other codeword.
- ❖ Ex) $C = \{0, 10, 110, 1110\}$
 - $0101100110 \rightarrow 0, 10, 110, 0, 110$ uniquely decodable
 - Instantaneous, since the end of a codeword is immediately recognizable
 - Self-punctuating



Examples

- ❖ $\mathcal{X} = \{1, 2, 3, 4\}$
- ❖ Consider the following codes (the elements are ordered)
 - Non-singular, uniquely decodable, instantaneous?
- ❖ Map#1: $\{0, 0, 1, 1\}$
- ❖ Map#2: $\{0, 010, 01, 10\}$
- ❖ Map#3: $\{10, 00, 11, 110\}$
- ❖ Map#4: $\{0, 10, 110, 111\}$

Kraft Inequality

If A is a prefix code, then the lengths of A 's codeword satisfies Kraft inequality.

- ❖ For any *instantaneous* code over D -ary alphabet, *the collection of codeword lengths, l_1, l_2, \dots, l_m , must satisfy the inequality*

$$\sum_{i=1}^m D^{-l_i} \leq 1$$

where m is the number of codewords.

If the lengths of a code A satisfies Kraft inequality, A is a prefix code.

True?

- ❖ (Converse) Given *a col. of codeword lengths* that satisfy this inequality, there *exists* an instantaneous code with these word lengths.

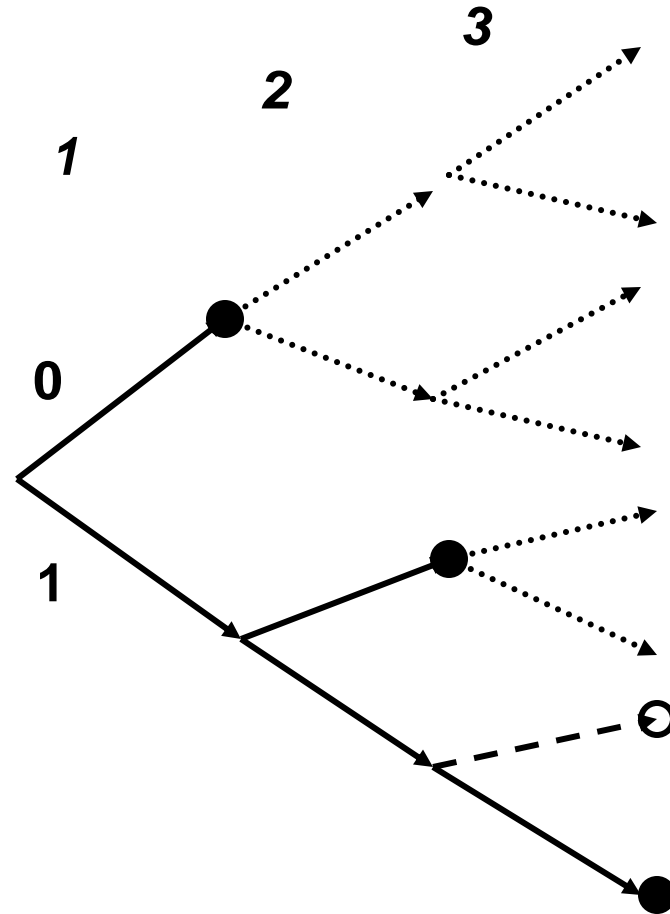
- Sufficiency test for existence of a prefix code.

If the lengths satisfies Kraft inequality, there exists a prefix code with these word lengths.

- ❖ Fundamental constraints on the lengths of a prefix code.

Kraft Inequality (Proof)

- ❖ Prefix code ~ each codeword has no child.
- ❖ l_1, \dots, l_m
- ❖ $l_{\max} = \max\{l_1, \dots, l_m\}$
- ❖ $D^{l_{\max}} \geq \sum_i D^{l_{\max} - l_i}$
 $\Rightarrow 1 \geq \sum_i D^{-l_i} . \text{(QED)}$
- ❖ When do you have equality?
- ❖ Ex) $2^3 \geq 4 + 2 + 1$ or $4+2+1+1$



The previous proof is not rigorous!

- ❖ The proof was based on induction ($D=2, 3, \dots$).
 - Consider the leaves at the maximum depth of the tree.
 - The total number of leaves is greater than or equal to the sum of the leaves displaced by codewords.
 - All descents of a codeword are displaced
 - Any leaf occupied by a codeword is displaced as well.
 - Sum of codewords + descendants of codewords cannot be greater than the total number of leaves at the maximum depth.
- ❖ Any way to improve the proof?

Kraft Inequality for UD code

- ❖ For any *uniquely decodable* code over D -ary alphabet, *the collection of codeword lengths, l_1, l_2, \dots, l_m* , must satisfy the inequality

$$\sum_{i=1}^m D^{-l_i} \leq 1$$

where m is the number of codewords.

- ❖ (Converse) Given *a col. of codeword lengths* that satisfy this inequality, there *exists* a uniquely decodable code with these word lengths.

- Sufficiency test for existence of a uniquely decodable code.

- ❖ Fundamental constraints on the lengths of a UD code.

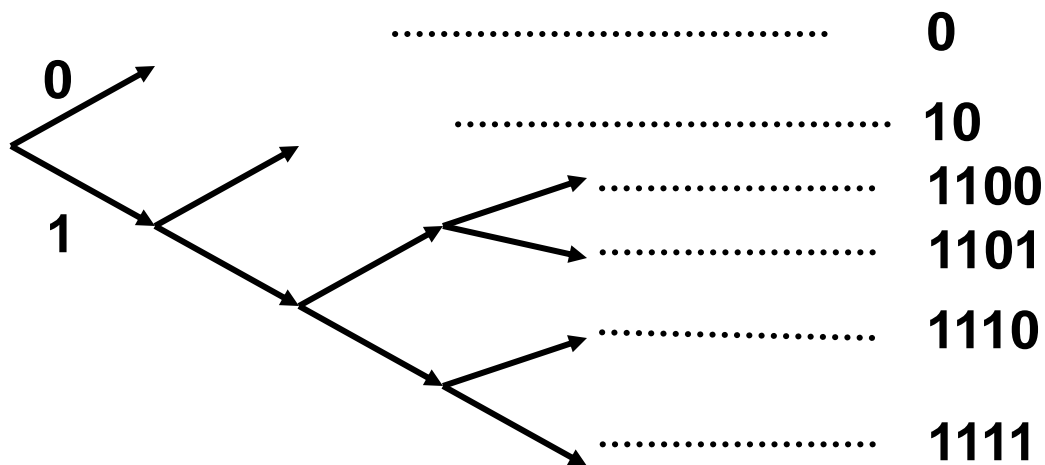
Examples

- ❖ Is there an instantaneous prefix code with a col. of lengths $\{1, 2, 2, 3\}$?
 - NO, since $2^{-1} + 2^{-2} + 2^{-2} + 2^{-3} > 1$
- ❖ How about $\{1, 2, 3, 3\}$?

Example

❖ {1, 2, 4, 4, 4, 4}

$$2^{-1} + 2^{-2} + 4 * 2^{-4} = 1/2 + 1/4 + 4 * 1/16 = 1$$



Optimal Codes

- ❖ Find a prefix code that has the minimum $L=E(l)$.

There are many prefix codes satisfying the Kraft inequality.

$E(l)$ means there is a distribution.

- ❖ First, find the lengths $\{l_1, \dots, l_m\}$ which satisfy the Kraft inequality and have the minimum L .

Given the distribution, p_i 's, map them to lengths l_i 's.

Find the lengths l_i 's satisfying KI, map them to p_i 's.

Looks like we are doing the first.

Given p_i 's find l_i 's.

- ❖ Second, construct the prefix code using the tree.

Let's use the Calculus, just to have an idea

❖ Use the Lagrange multiplier method to solve

– Minimize $L = \sum p_i l_i$

– Subject to $\sum D^{-l_i} \leq 1$

Find l_i 's
given p_i 's.

– Let's relax the condition and let l_i be any positive number.

– Let $l_i = l(C(x_i))$, $i=1, \dots, |\mathcal{X}|$ and $p_i = \Pr\{X = x_i\}$.

– Note that $|\mathcal{X}| = m$

Lagrange multiplier minimization

❖ Let $J = \sum p_i l_i + \lambda (\sum D^{-l_i})$

❖ Differentiate wrt each l_i and obtain

$$\partial J / \partial l_i = p_i - \lambda D^{-l_i} \log D,$$

❖ Setting it equal to 0, we have

$$D^{-l_i} = p_i / (\lambda \log D)$$

❖ Substitute this in the constraint $\sum D^{-l_i} = 1$ and obtain $\lambda = 1 / \log D$

❖ Thus, $p_i = D^{-l_i}$ or $l_i^* = -\log_D p_i$

❖ Then, $L = E\{l(X)\} = \sum p_i l_i^* = -\sum p_i \log_D p_i = H_D(X)$

$$L \geq H_D(X)$$

- ❖ Since *codeword length* must be integer, we must do the ceiling operation, i.e.

$$l_i = \lceil l_i^* \rceil$$

- ❖ Thus, $L \geq H_D(X)$, equality *iff* l_i^* integers or $p_i = D^{-l_i}$
- ❖ Ex) $\{p_i\} = \{1/2, 1/4, 1/16, 1/16, 1/16, 1/16\}$
 - $\{C(x_i)\} = \{0, 10, 1100, 1101, 1110, 1111\}$
 - Lengths 1, 2, 4, 4, 4, 4
 - $E(L) = (1/2) * 1 + (1/4) * 2 + (1/4) * 4 = 2$
 - $H = (1/2) * \log_2(2) + (1/4) * \log_2(4) + 4 * (1/16) * \log_2(16) = 2$
 - $E(L) = H$ since $p_i = 2^{-l_i}$, $l_i = 1, 2, 4, 4, 4, 4$

D-adic distribution

- ❖ A distribution where each of the probabilities is equal to D^{-n} .
- ❖ The lower bound on expected length, $L \geq H_D(X)$ is achieved iff the distribution is *D-adic* (i.e. $p_i = D^{-li}$).

Information Theoretic Proof: $L \geq H_D(X)$

$$\begin{aligned} \spadesuit L - H_D(X) &= \sum_x p(x) l(x) - \sum_x p(x) \log_D(1/p(x)) \\ &= \sum_x p(x) \{ \log_D[p(x)] - \log_D[D^{-l(x)}] \} \\ &\quad \text{--- let } r(x) = D^{-l(x)}/r_o \\ &\quad \text{--- with } r_o = \sum D^{-l(x)} \leq 1 \text{ (Kraft Inequality)} \\ &= \sum_x p(x) \{ \log_D[p(x)] - \log_D[r(x)r_o] \} \\ &= \sum_x p(x) \{ \log_D[p(x)] - \log_D[r(x)] - \log_D[r_o] \} \\ &= \sum_x p(x) \log_D[p(x)/r(x)] + \log_D[1/r_o] \\ &\geq 0 \text{ (why?)} \end{aligned}$$

A possible procedure to find an optimal code

- ❖ Find the D -adic distribution that is closest (in the relative entropy sense) to the distribution of X
 - $p_i = D^{-l_i}$, $i = 1, \dots, |\mathcal{X}|$
 - So now you have $\{l_i\}$, the col. of codeword lengths.
- ❖ Construct a D -adic tree according to the col.
- ❖ Assign codewords on the leaves of the tree.

- ❖ The first step of searching for the closest D -adic distribution is not trivial.
- ❖ We may use a sub-optimal procedure.

$$H_D(\mathbf{X}) + 1/n > L_n \geq H_D(\mathbf{X})$$

- ❖ Previous pages say we have an overhead of maximum one bit per symbol.
- ❖ Now, consider a series of r.v.s, X_1, X_2, \dots ,
- ❖ We want to get rid of the one overhead bit per symbol by spreading out over many symbols.
- ❖ For a simple example, consider a group of iid $X_1, \dots, X_n \sim p(x)$. Then, the distribution is $P_n(x) = \prod_{i=1}^n p(x)$.
- ❖ Then, we have

$$H_D(X_1, \dots, X_n) \leq E\{l(X_1, \dots, X_n)\} < H_D(X_1, \dots, X_n) + 1$$
- ❖ Dividing by n , we have the expected length per symbol L_n

$$H_D(\mathbf{X}) \leq L_n < H_D(\mathbf{X}) + 1/n$$

X_1, X_2, \dots, X_n is a stationary stochastic process

- ❖ We know $H(X_1, \dots, X_n)/n \rightarrow H(\mathcal{X})$ (Cesaro Mean)
 - $nH(\mathcal{X})$ bits \sim sufficient for description of typical seq. of length n
- ❖ Then, the minimum expected codeword length per symbol converges to the entropy rate of the process

$$L_n^* \rightarrow H(\mathcal{X}) \text{ as } n \rightarrow \infty$$

Messages learned

- ❖ Thus, one can always construct a near optimal prefix code.
 - Use the Shannon code, $l_i = \lceil \log_D(1/p_i) \rceil$
 - Use a sequence, rather than a symbol (vector processing, rather than a symbol processing)

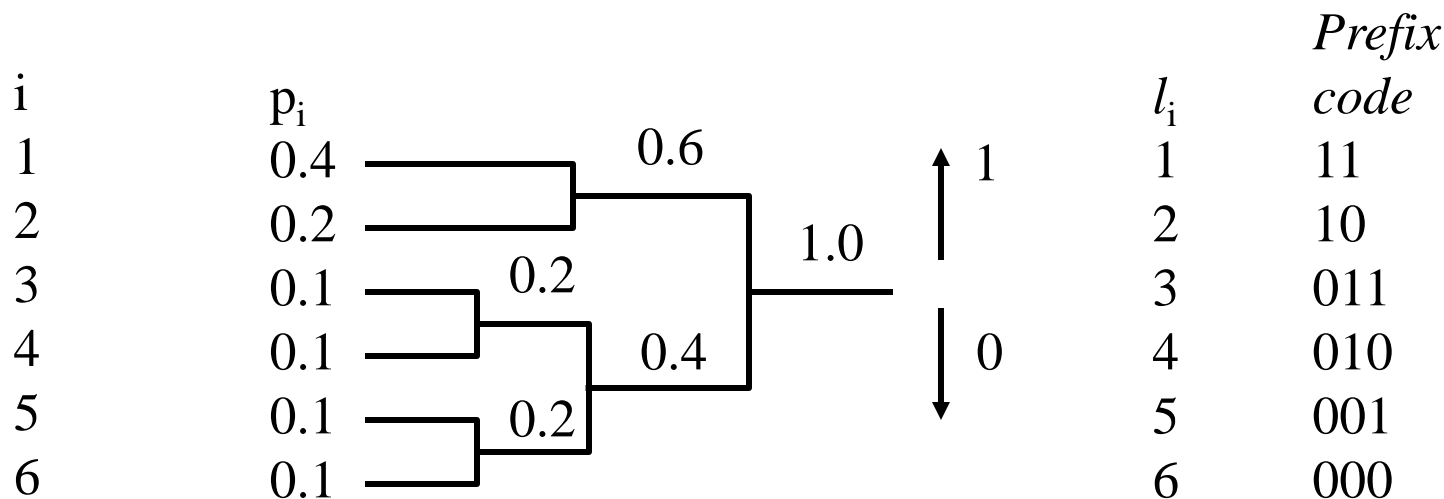
Two Important Coding Methods

❖ Huffman code

❖ Lempel-Ziv code

Huffman Code (Huffman Tree)

- ❖ Huffman code is an optimal *prefix* code, with the shortest *expected* length for a given distribution, which can be constructed by the Huffman algorithm.
- ❖ It minimizes $L = \sum_i p(x_i) l_i$.
- ❖ Compare it with $H = \sum_i p(x_i) \log(1/p(x_i))$.



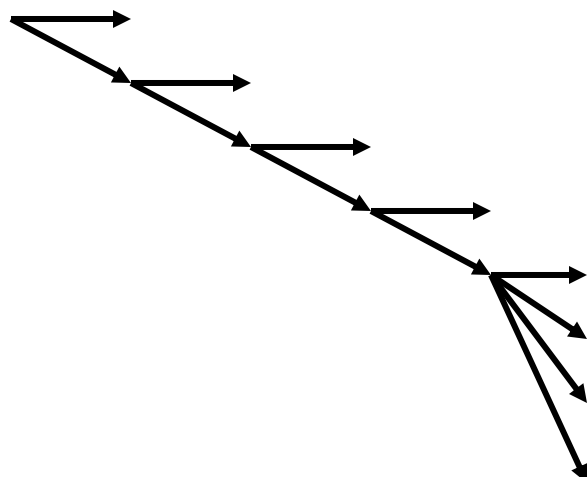
Huffman Code

❖ $L = 0.4*2 + 0.2*2 + 4*0.1*3 = 2.4$

❖ $H = 2.319$ (Lower bound)

- It should be noted here that the expected length turned out quite close to the lower bound.

Recall the 8-Horse Race Problem

p_i				Length
1/2	0		0	1
1/4	1		10	2
1/8	2		110	3
1/16	3		1110	4
1/64	4		111100	6
1/64	5		111101	6
1/64	6		111110	6
1/64	7		111111	6

- ❖ The code we constructed is a Huffman code.
- ❖ The distribution is 2-adic.
- ❖ $L = 2 = H$

Huffman Encoding Algorithm

- ❖ Construct a tree in the following routine:
 - First, have the probabilities (and the index) listed in the decreasing order.
 - Second, go over the list and locate two indices with lowest probabilities. Group these low probable items, and add the probabilities and label it for the group. Now note that the size of the list is reduced by one.
 - Repeat the second step exhaustively.
- ❖ Once a tree is completed, we can assign “1” for up and “0” for down from the top of the tree.

Optimality of Huffman Codes

- ❖ Given a distribution, there could be many Huffman codes which all lead to the same expected length.
- ❖ Let's call the Huffman tree procedure Huffman coding.
- ❖ Huffman coding is optimal in the sense that if C^* is a Huffman code, and C' is a code from other coding procedure, then $L(C^*) \leq L(C')$
 - See Ch5.8 for proof by induction

Lempel-Ziv Coding (Ch13)

- ❖ Construction of Huffman code requires knowledge of priors.
- ❖ Huffman code does not make use of correlation between words and phrases (based on the assumption that text is generated from a random variable, rather than a random process)
- ❖ Lempel-Ziv code figures out the correlation structure of the source in a sequence, and further compress
 - LZ code is a universal code (does not need to know the distribution or correlation of the source).
 - LZ code is dictionary based.
- ❖ It's adaptive and simple.
- ❖ **Operation:** parse the data stream into segments that are the shortest subsequences not appeared previously.

Lempel-Ziv Encoding

❖ Let's take a simple example for illustration of the algorithm

❖ Consider a data stream

aaababbbbaaabaabbbababb

❖ Starting from the left of the stream, parse the data stream into segments not appeared previously

a, aa, b, ab, bb, aaa, ba, aaaa, bba, bab, b

Index-- 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11

Encoding-- a, 1a, b, 1b, 3b, 2a, 3a, 6a, 5a, 7b, 3

❖ (Appeared, new) = (Index, new) = (4 bits, 7 bits)

❖ The total number of bits per segment needed in this example is 4 bits for the index and 7 bit ascii code for the new part.

– 11 segments \times (4 + 7) bits = 121 bits

– 24 characters \times 7 bits = 168 bits

– (168-121)/168 is the compression ratio in this example

Lempel-Ziv Coding (MATLAB)

- ❖ We will compress the following text with an LZ algorithm that I programmed in MATLAB, `Lempel_ziv.m`
- ❖ [Jordan_text.txt](#)
- ❖ [H1N1_virus.txt](#)

Huffman Codes vs. Shannon Codes

- ❖ X is binary, X=1 with prob. 2^{-10} , X=0 with prob. $1 - 2^{-20}$
- ❖ Shannon Code: $l(x=1) = \lceil \log_2(2^{10}) \rceil = 10$, $l(x=0) = 1$
- ❖ Huffman Code: $l(x=1) = 1$ and $l(x=0) = 1$
- ❖ A codeword for infrequent symbol in Shannon Code is much longer than in an optimal code.
- ❖ But, this does not mean that a codeword in an optimal code is always shorter than those in the Shannon code.
- ❖ Ex) $X \sim (1/3, 1/3, 1/4, 1/12)$
 - two kinds of Huffman trees with lengths (2, 2, 2, 2) or (1, 2, 3, 3)
 - Shannon lengths (2, 2, 2, 4)

HW

❖ P4.4, 4.7, 4.11, 4.24

❖ P5.3, 5.6, 5.7, 5.16, 5.18

Kraft Inequality for Uniquely Decodable Codes

- ❖ Recall that the uniquely decodable codes is a super-set of the prefix code
 - Prefix code is uniquely decodable
 - A uniquely decodable code may not be a prefix code
 - Let A be the set of codeword lengths for uniquely decodable code
 - Let B be the set of those for prefix code
 - $|A|$
- ❖ (McMillan): $\{l_1, \dots, l_m\}$ with $\sum D^{-l_i} \leq 1$
 - \Leftrightarrow
 - Uniquely decodable code with lengths $\{l_1, \dots, l_m\}$

McMillan Inequality: $\sum D^{-li} \leq 1$

- ❖ “ \Rightarrow ” part is easy since prefix code is a uniquely decodable code
- ❖ Let’s prove “ \Leftarrow ” part: set of codeword-lengths of a uniquely decodable code must satisfy the Kraft inequality.
- ❖ Let $l(x)$ be the codeword length of the symbol $x \in \mathcal{X}$.
- ❖ Let $l(x_1, \dots, x_k)$ be the codeword lengths of the k -th extension code

$$l(\mathbf{x}) := l(x_1, \dots, x_k) = \sum_{i=1}^k l(x_i) . \quad \text{---(1)}$$

- ❖ We need to show that $\sum_{\mathbf{x} \in \mathcal{X}} D^{-l(\mathbf{x})} \leq 1$.

Back to Proof of McMillan

$$\begin{aligned}
 \spadesuit \left(\sum_{x \in \mathcal{X}} D^{-l(x)} \right)^k &= \sum_{x_1 \in \mathcal{X}} \cdots \sum_{x_k \in \mathcal{X}} D^{-l(x_1)} \cdots D^{-l(x_k)} \\
 &= \sum_{\mathbf{x} \in \mathcal{X}^k} D^{-l(\mathbf{x})} \quad \text{--- (1)}
 \end{aligned}$$

By gathering terms wrt wordlengths

$$\begin{aligned}
 &= \sum_{m'=1}^{k l_{\max}} a(m') D^{-m'} \\
 &\leq \sum D^{m'} D^{-m'} \\
 &= k l_{\max}
 \end{aligned}$$

$$\spadesuit \sum_{x \in \mathcal{X}^k} D^{-l(x)} \leq (k l_{\max})^{1/k} \rightarrow 1 \text{ as } k \rightarrow \infty$$

Optimality of Huffman Codes (Off)

- ❖ Definition of Optimality here: $\sum p_i l_i$ is minimum
- ❖ Huffman algorithm gives an optimal code
- ❖ Lemma 5.8.1: An *optimal* instantaneous code satisfies the following properties:
 - Let $p_1 \geq p_2 \geq \dots \geq p_m$
 - $p_i > p_j \Rightarrow l_i \leq l_j$
 - The two longest codewords have the same length.
 - The two longest codewords differ only in the last bit and correspond to the two least likely symbols.

Proof (1) (Off)

- ❖ If $p_j > p_k$, then $l_j \leq l_k$
- ❖ Consider an optimal code C_m
- ❖ Consider C'_m , with the codewords j and k of C_m swapped
- ❖
$$\begin{aligned} L(C'_m) - L(C_m) &= \sum p_i l'_i - \sum p_i l_i \\ &= p_j l_k - p_j l_j + p_k l_j - p_k l_k \\ &= p_j (l_k - l_j) + p_k (l_j - l_k) \\ &= (p_j - p_k) (l_k - l_j) \\ &\geq 0 \quad (\text{Since } C_m \text{ is optimal}) \end{aligned}$$
- ❖ Thus, if $(p_j - p_k) > 0$, $l_k - l_j \geq 0$

Proof (3) (Off)

- ❖ The two longest codewords differ only in the last bit and correspond to the two least likely symbols

Summary of Lemma (Off)

- ❖ If $p_1 \geq p_2 \geq \dots \geq p_m$, then there exists an optimal code with $l_1 \leq l_2 \leq \dots \leq l_{m-1} = l_m$, and the codewords $C(x_{m-1}) = C(x_m)$ differ only in the last bit.

- ❖ We now want to prove that a code satisfying the lemma is optimal
- ❖ Use the idea of “merging”– just as in Huffman algorithm.

“Merged” code C_{m-1} (Off)

- ❖ Take the common prefix of the longest words and assign prob.

$$p_{m-1} + p_m$$

- ❖
$$L(C_m) = \sum_{i=1}^m p_i l_i$$

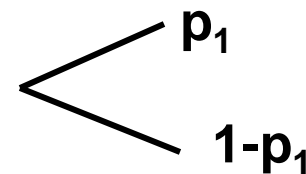
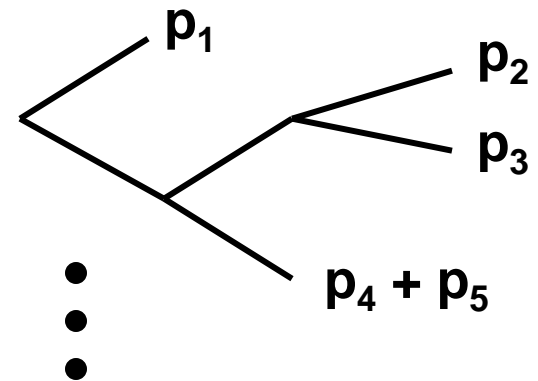
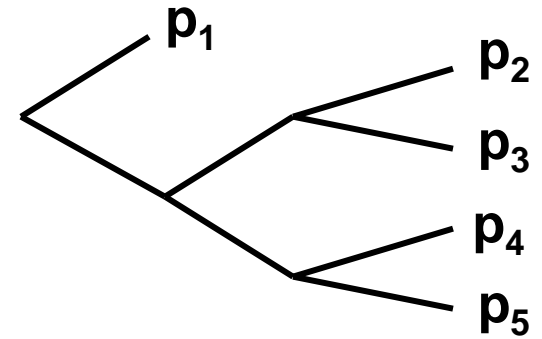
$$= \sum_{i=1}^{m-2} p_i l_i + p_{m-1} l_{m-1} + p_m l_m$$

$$= \sum_{i=1}^{m-2} p_i l_i + (p_{m-1} + p_m)(l_m - 1)$$

$$+ p_{m-1} + p_m$$

$$= L(C_{m-1}) + p_{m-1} + p_m$$

- ❖ Thus, minimization of $L(C_m)$ is equivalent to minimization of $L(C_{m-1})$, and so on



“Merged” code C_{m-1} (Off)

- ❖ When we have only two symbols left, the solution is simple, assign 0 for one and 1 for the other.
- ❖ $L(C_2)$ is minimized.
- ❖ Then, $L(C_3)$ is minimized, ... and so on.
- ❖ Thus, $L(C_m)$ satisfying the lemma is minimum.
- ❖ Thus, C_m which satisfies the lemma is an optimal code.

Huffman Coding is Optimal (Off)

- ❖ If C^* is Huffman and C is any other code, then $L(C^*) \leq L(C)$

Information Theory

4th Module

Agenda

- ❖ $L = E(L)$ in this lecture note
- ❖ Entropy bound on data compression

$$L := \sum p_i l_i \geq H_D(X)$$

- ❖ Shannon Code: $l_i = \lceil \log_D(1/p_i) \rceil$

$$L < H_D(X) + 1$$

- ❖ McMillan inequality:

$$\text{Uniquely decodable code} \Leftrightarrow \sum D^{-li} \leq 1$$

- ❖ Huffman code:

$$L^* = \min_{\sum D^{-li} \leq 1} \sum p_i l_i$$

- ❖ Chapter 8: Channel Capacity

D-adic distribution

- ❖ A distribution where each of the probabilities is equal to D^{-n} .
- ❖ Lower bound on expected length, $L \geq H_D(X)$ is achieved iff the distribution is *D-adic* (i.e. $p_i = D^{-li}$).

Procedure of finding an optimal code

- ❖ Find the D -adic distribution that is closest (in the relative entropy sense) to the distribution of X
 - $p_i = D^{-l_i}$, $i = 1, \dots, |\mathcal{X}|$
 - So now you have $\{l_i\}$, the set of codeword lengths.
- ❖ Construct a D -adic tree according to the set.
- ❖ Assign codewords on the leaves of the tree.

- ❖ The first step of searching for the closest D -adic distribution is not trivial.
- ❖ We may use a sub-optimal procedure.

$$H(X) \leq L < H(X) + 1 \text{ (Off)}$$

- ❖ Let's study a coding scheme. The expected length of the code is to be within one additional bit of the lower bound.
- ❖ We showed $L - H_D(X) = D(p \parallel r) - \log(\sum D^{-l_i}) \geq 0$.
- ❖ The choice of word length $l_i = \log_D(1/p_i)$ yields $L = H$, which is the case for a D-adic distribution.

- ❖ But what if we do the ceiling operation (the Shannon code), i.e.?

$$l_i = \lceil \log_D(1/p_i) \rceil \geq \log_D(1/p_i),$$

- ❖ These lengths surely satisfy the Kraft inequality since

$$\sum D^{-l_i} \leq \sum D^{-\log_D\{1/p_i\}} = \sum p_i = 1.$$

- ❖ By the definition of ceiling operation, we have

$$\log_D(1/p_i) + 1 > l_i \geq \log_D(1/p_i) \Rightarrow H_D(X) + 1 > L \geq H_D(X)$$

$$H_D(X) + 1/n > L_n \geq H_D(X) \text{ (Off)}$$

- ❖ Previous page says we have an overhead of maximum one bit per symbol.
- ❖ Now, consider a series of r.v.s, X_1, X_2, \dots ,
- ❖ We want to get rid of the one overhead bit per symbol by spreading out over many symbols.
- ❖ For a simple example, consider a group of iid $X_1, \dots, X_n \sim p(x)$. Then, the distribution is $P_n(x) = \prod_{i=1}^n p(x)$.
- ❖ Then, we have
$$H_D(X_1, \dots, X_n) \leq E\{l(X_1, \dots, X_n)\} < H_D(X_1, \dots, X_n) + 1$$
- ❖ Dividing by n , we have the expected length per symbol L_n
$$H_D(X) \leq L_n < H_D(X) + 1/n$$

X_1, X_2, \dots, X_n is a stationary stochastic process

- ❖ We know $H(X_1, \dots, X_n)/n \rightarrow H(\mathcal{X})$ (Cesaro Mean)
 - $nH(X)$ bits \sim sufficient for description of typical seq. of length n
- ❖ Then, the minimum expected codeword length per symbol converges to the entropy rate of the process

$$L_n^* \rightarrow H(\mathcal{X}) \text{ as } n \rightarrow \infty$$

Kraft Inequality for Uniquely Decodable Codes

- ❖ Recall that the uniquely decodable codes is a super-set of the prefix code
 - Prefix code is uniquely decodable
 - A uniquely decodable code may not be a prefix code
 - Let A be the set of codeword lengths for uniquely decodable code
 - Let B be the set of those for prefix code
 - $|A|$
- ❖ (McMillan): $\{l_1, \dots, l_m\}$ with $\sum D^{-l_i} \leq 1$
 - \Leftrightarrow
 - Uniquely decodable code with lengths $\{l_1, \dots, l_m\}$

Uniquely Decodable Code (Off)

- ❖ Extension of a uniquely decodable code is non-singular.
- ❖ Consider C^k , the k -th extension of a uniquely decodable code C .
 - It's a concatenation of k repetitions of the code C .

McMillan Inequality: $\sum D^{-li} \leq 1$

- ❖ “ \Rightarrow ” part is easy since prefix code is a uniquely decodable code
- ❖ Let’s prove “ \Leftarrow ” part: set of codeword-lengths of a uniquely decodable code must satisfy the Kraft inequality.
- ❖ Let $l(x)$ be the codeword length of the symbol $x \in \mathcal{X}$.
- ❖ Let $l(x_1, \dots, x_k)$ be the codeword lengths of the k -th extension code

$$l(\mathbf{x}) := l(x_1, \dots, x_k) = \sum_{i=1}^k l(x_i) . \quad \text{---(1)}$$

- ❖ We need to show that $\sum_{\mathbf{x} \in \mathcal{X}} D^{-l(\mathbf{x})} \leq 1$.

Some Pre-thoughts (Off)

- ❖ Let $l_{\max} = \max \{l_1, \dots, l_{|X|}\}$
- ❖ Can form a set of distinct lengths, i.e., $m = 1, \dots, l_{\max}$.
- ❖ Let $a(m) = \#$ of codewords of length m .
- ❖ Ex) let $X = \{x_1, x_2, x_3, x_4\}$ and the corresponding lengths are 1, 2, 3, and 3 respectively.
 - 1 word with length 1, $a(1) = 1, m = 1$
 - 1 word with length 2, $a(2) = 1, m = 2$
 - 2 words with length 3, $a(3) = 2, m = 3$
 - $l_{\max} = 3$

Some Pre-thoughts (2) (Off)

❖ Concatenation of 2 codes:

- the maximum length of the code is $2 \times l_{\max} = 6$
- $m' = 1, 2, \dots, 6$;

❖ find the number of codewords $a(m')$ with length m'

	1	2	3	3
1	2	3	4	4
2	3	4	5	5
3	4	5	6	6
3	4	5	6	6

❖ $a(1) = 0, a(2) = 1, a(3) = 2, a(4) = 4, a(5) = 4, a(6) = 4$

❖ Note $a(m') \leq D^{m'}$

- There are at most $D^{m'}$ distinct sequences of D-ary alphabet

Back to Proof of McMillan

$$\begin{aligned} \spadesuit \left(\sum_{x \in \mathcal{X}} D^{-l(x)} \right)^k &= \sum_{x_1 \in \mathcal{X}} \cdots \sum_{x_k \in \mathcal{X}} D^{-l(x_1)} \cdots D^{-l(x_k)} \\ &= \sum_{\mathbf{x} \in \mathcal{X}^k} D^{-l(\mathbf{x})} \quad \text{--- (1)} \end{aligned}$$

By gathering terms wrt wordlengths

$$\begin{aligned} &= \sum_{m'=1}^{k l_{\max}} a(m') D^{-m'} \\ &\leq \sum D^{m'} D^{-m'} \\ &= k l_{\max} \end{aligned}$$

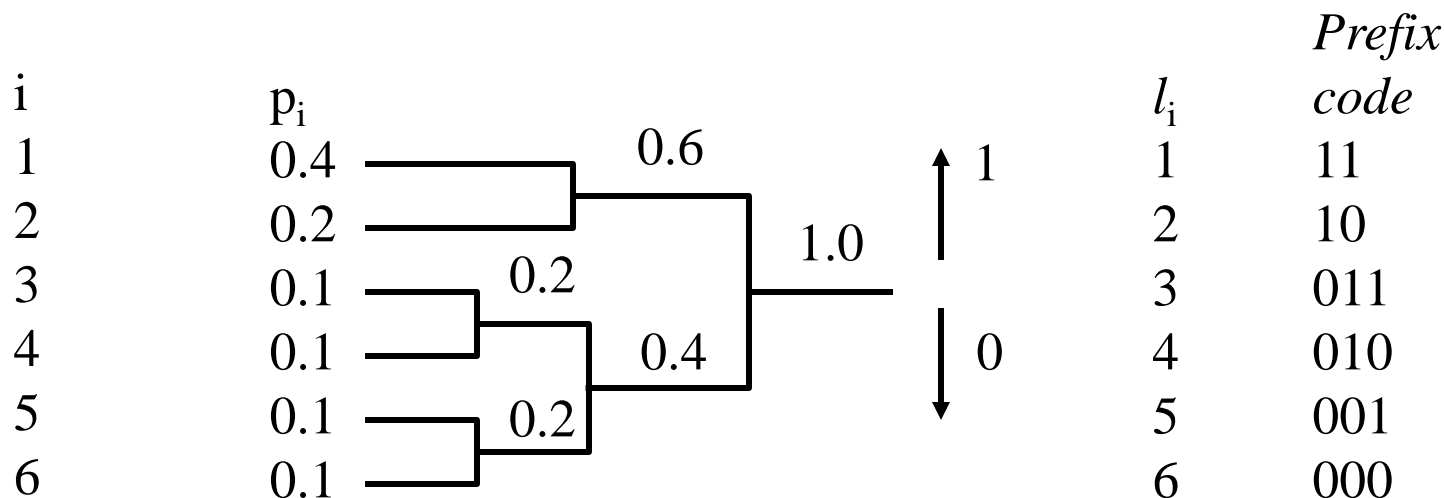
$$\spadesuit \sum_{x \in \mathcal{X}^k} D^{-l(x)} \leq (k l_{\max})^{1/k} \rightarrow 1 \text{ as } k \rightarrow \infty$$

McMillan ($\sum_{i=1}^{\infty} D^{-l_i} \leq 1$?) (Off)

- ❖ What if $l_{\max} = \infty$ (?), i.e. $|X|$ is infinite
- ❖ Property of uniquely decodable code
 - A subset is also uniquely decodable

Huffman Code (Huffman Tree)

- ❖ Huffman code is an optimal prefix code, with the smallest *expected* length for a given distribution, which can be constructed by Huffman algorithm.
- ❖ It minimizes $L = \sum_i p(x_i) l_i$.
- ❖ Compare it with $H = \sum_i p(x_i) \log(1/p(x_i))$.



Huffman Code

❖ $L = 0.4*2 + 0.2*2 + 4*0.1*3 = 2.4$









❖ $H = 2.319$ (Lower bound)

- It should be noted here that the expected length turned out quite close to the lower bound.

Huffman Encoding Algorithm

- ❖ Construct a tree in the following routine
 - First, have the probabilities (and the index) listed in the decreasing order.
 - Second, go over the list and locate two indices with lowest probabilities. Group these low probable items, and add the probabilities and label it for the group. Now note that the size of the list is reduced by one.
 - Repeat the second step exhaustively.
- ❖ Once a tree is completed, we can assign “1” for up and “0” for down from the top of the tree.

Recall the 8-Horse Race Problem

p_i				Length
1/2	0		0	1
1/4	1		10	2
1/8	2		110	3
1/16	3		1110	4
1/64	4		111100	6
1/64	5		111101	6
1/64	6		111110	6
1/64	7		111111	6

- ❖ The code we constructed is a Huffman code.
- ❖ The distribution is 2-adic.
- ❖ $L = 2 = H$

Optimality of Huffman Codes

- ❖ Given a distribution, there could be many Huffman codes which all lead to the same expected length.
- ❖ Let's call the Huffman tree procedure as Huffman coding.
- ❖ Huffman coding is optimal in the sense that if C^* is a Huffman code, and C' is a code from other coding procedure, then $L(C^*) \leq L(C')$
 - See Ch5.8 for proof by induction

Lempel-Ziv Coding (Ch12.10)

- ❖ Construction of Huffman code requires knowledge of priors.
- ❖ Huffman code does not make use of correlation between words and phrases (based on the assumption that text is generated from a random variable, rather than a random process)
- ❖ Lempel-Ziv code figures out the correlation structure of the source in a sequence, and further compress
 - LZ code is a universal code (does not need to know the distribution or correlation of the source).
 - LZ code is dictionary based.
- ❖ It's adaptive and simple.
- ❖ Operation: parse the data stream into segments that are the shortest subsequences not appeared previously.

Lempel-Ziv Encoding

❖ Let's take a simple example for illustration of the algorithm

❖ Consider a data stream

aaababbbbaaabaabbbababb

❖ Starting from the left of the stream, parse the data stream into segments not appeared previously

a, aa, b, ab, bb, aaa, ba, aaaa, bba, bab, b

Index— 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11

Encoding-- a, 1a, b, 1b, 3b, 2a, 3a, 6a, 5a, 7b, 3

❖ (Appeared, new) = (Index, new) = (4 bits, 7 bits)

❖ The total number of bits per segment needed in this example is 4 bits for the index and 7 bit ascii code for the new part.

– 11 segments \times (4 + 7) bits = 121 bits

– 24 characters \times 7 bits = 168 bits

– (168-121)/168 is the compression ratio in this example

Huffman Codes vs. Shannon Codes

- ❖ X is binary, X=1 with prob. 2^{-10} , X=0 with prob. $1 - 2^{-20}$
- ❖ Shannon Code: $l(x=1) = \lceil \log_2(2^{10}) \rceil = 10$, $l(x=0) = 1$
- ❖ Huffman Code: $l(x=1) = 1$ and $l(x=0) = 1$
- ❖ A codeword for infrequent symbol in Shannon Code is much longer than in an optimal code.
- ❖ But, this does not mean that the codeword in an optimal code is always shorter than in the Shannon code.
- ❖ Ex) $X \sim (1/3, 1/3, 1/4, 1/12)$
 - two kinds of Huffman trees with lengths (2, 2, 2, 2) or (1, 2, 3, 3)
 - Shannon lengths (2, 2, 2, 4)

Optimality of Huffman Codes (Off)

- ❖ Definition of Optimality here: $\sum p_i l_i$ is minimum
- ❖ Huffman algorithm gives an optimal code
- ❖ Lemma 5.8.1: An *optimal* instantaneous code satisfies the following properties:
 - Let $p_1 \geq p_2 \geq \dots \geq p_m$
 - $p_i > p_j \Rightarrow l_i \leq l_j$
 - The two longest codewords have the same length.
 - The two longest codewords differ only in the last bit and correspond to the two least likely symbols.

Proof (1) (Off)

- ❖ If $p_j > p_k$, then $l_j \leq l_k$
- ❖ Consider an optimal code C_m
- ❖ Consider C'_m , with the codewords j and k of C_m swapped
- ❖
$$\begin{aligned} L(C'_m) - L(C_m) &= \sum p_i l'_i - \sum p_i l_i \\ &= p_j l_k - p_j l_j + p_k l_j - p_k l_k \\ &= p_j (l_k - l_j) + p_k (l_j - l_k) \\ &= (p_j - p_k) (l_k - l_j) \\ &\geq 0 \quad (\text{Since } C_m \text{ is optimal}) \end{aligned}$$
- ❖ Thus, if $(p_j - p_k) > 0$, $l_k - l_j \geq 0$

Proof (3) (Off)

- ❖ The two longest codewords differ only in the last bit and correspond to the two least likely symbols

Summary of Lemma (Off)

- ❖ If $p_1 \geq p_2 \geq \dots \geq p_m$, then there exists an optimal code with $l_1 \leq l_2 \leq \dots \leq l_{m-1} = l_m$, and the codewords $C(x_{m-1}) = C(x_m)$ differ only in the last bit.

- ❖ We now want to prove that a code satisfying the lemma is optimal
- ❖ Use the idea of “merging”– just as in Huffman algorithm.

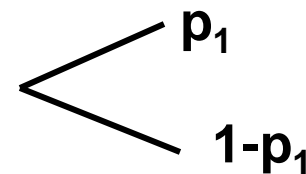
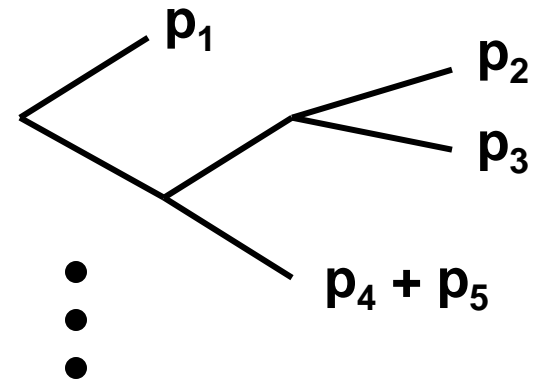
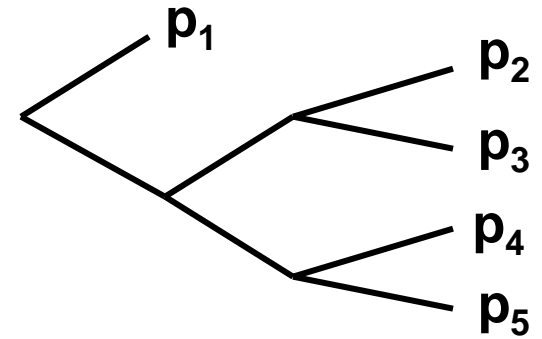
“Merged” code C_{m-1} (Off)

- ❖ Take the common prefix of the longest words and assign prob.

$$p_{m-1} + p_m$$

- ❖
$$\begin{aligned} L(C_m) &= \sum_{i=1}^m p_i l_i \\ &= \sum_{i=1}^{m-2} p_i l_i + p_{m-1} l_{m-1} + p_m l_m \\ &= \sum_{i=1}^{m-2} p_i l_i + (p_{m-1} + p_m)(l_m - 1) \\ &\quad + p_{m-1} + p_m \\ &= L(C_{m-1}) + p_{m-1} + p_m \end{aligned}$$

- ❖ Thus, minimization of $L(C_m)$ is equivalent to minimization of $L(C_{m-1})$, and so on



“Merged” code C_{m-1} (Off)

- ❖ When we have only two symbols left, the solution is simple, assign 0 for one and 1 for the other.
- ❖ $L(C_2)$ is minimized.
- ❖ Then, $L(C_3)$ is minimized, ... and so on.
- ❖ Thus, $L(C_m)$ satisfying the lemma is minimum.
- ❖ Thus, C_m which satisfies the lemma is an optimal code.

Huffman Coding is Optimal (Off)

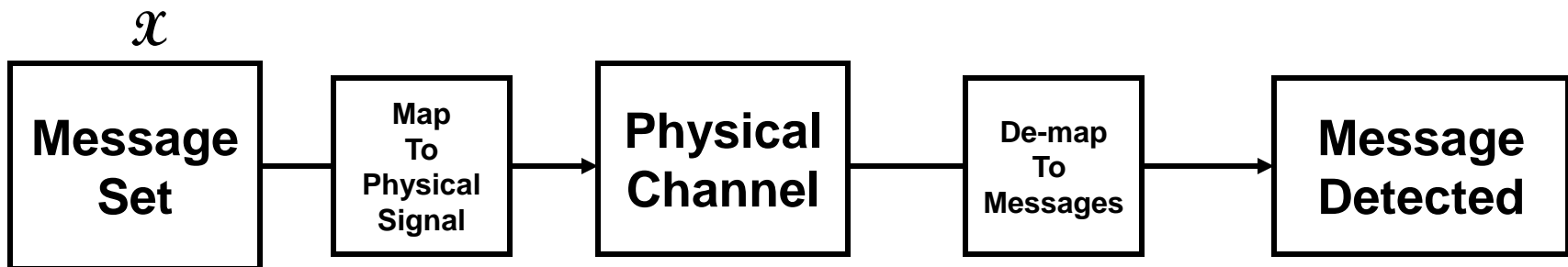
- ❖ If C^* is Huffman and C is any other code, then $L(C^*) \leq L(C)$

Skip Some Sideline Topics

- ❖ Chapter 7: Gambling and Data Compression
- ❖ Chapter 8: Kolmogoroff Complexity

Channel Capacity (Chapter 8)

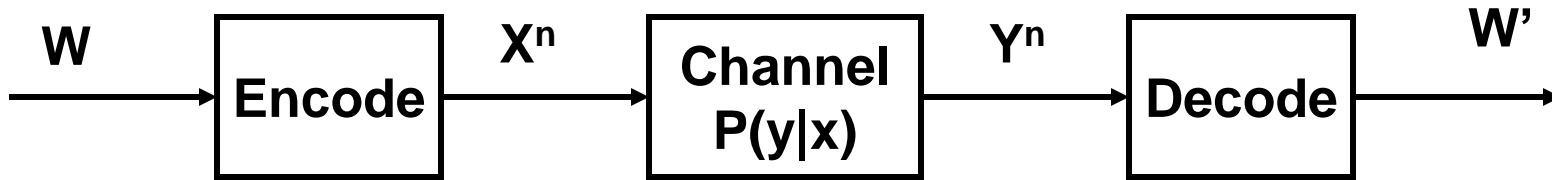
- ❖ Most successful application of Shannon Theory
- ❖ What's Channel Capacity?
- ❖ What's the size of message set that can be transmitted over the channel and be recovered almost error free?
- ❖ The size is small for channel with a lot of noise and distortion.
- ❖ The size is large for clean channel with no distortion,



Channel Capacity

- ❖ $\text{Capacity}_1 = \log(\text{Size of the Message Set})$
- ❖ $T = \text{Time required to transmit a message in the set.}$
- ❖ $\text{Capacity} = \text{Capacity}_1/T, \text{ bits/sec}$
- ❖ $\text{Capacity} = \text{Capacity in } n\text{-use of channel} / n = \textit{bits/channel use}$
- ❖ Capacity is the number of bits that can be transferred over a given channel with almost no error.

Discrete Channel



- ❖ Input alphabet X , output alphabet Y , and the probability transition matrix $p(y|x) \sim$ prob. of observing y given x transmitted .
- ❖ The channel is memoryless if $P(y^n|x^n) = \prod p(y|x)$.
- ❖ We can transform the physical channel into the discrete channel.
- ❖ Example) A cable with a certain bandwidth F_{\max} and noise power spectral density \rightarrow Sample at twice the bandwidth and obtain the discrete channel (sample IN sample OUT) .

“*Information*” Channel Capacity

❖ $C = \max_{p(x)} I(X; Y)$

- The maximum is taken over all input distr. $p(x)$.
- Compared with the “*operational*” channel capacity ~ the highest rate in bits/channel-use, at which information can be sent with an arbitrary small probability of error.
- Shannon proved that the *operational* channel capacity is equal to the *information* channel capacity.

Noiseless Binary Channel

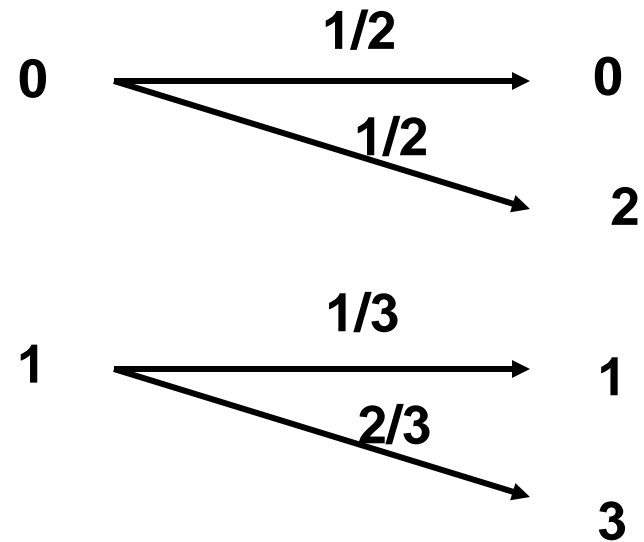
❖ $C = \max_{p(x)} I(X; Y) = ?$

0 → 0

1 → 1

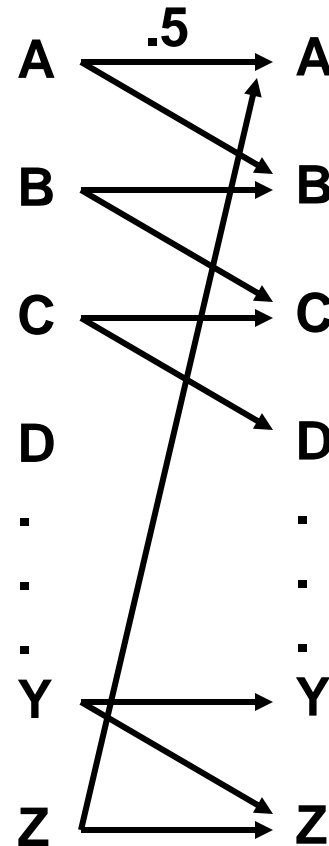
Noisy Channel with Non-Overlapping Output

❖ $C = ?$



Noisy Typewriter

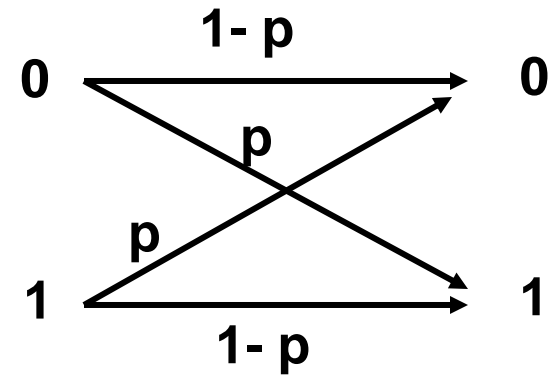
- ❖ The typewriter writes the input letter with prob. $1/2$ or the next letter in the alphabet with prob. $1/2$
- ❖ $C = ?$
- ❖ $C = \max_{p(x)} I(X ; Y)$
 $= \max_{p(x)} [H(Y) - H(Y|X)]$
 $= \max_{p(x)} H(Y) - 1$
 $= \log_2(26) - 1$
 $= \log_2(26/2)$
 $= \log_2(13)$



Binary Symmetric Channel (Off)

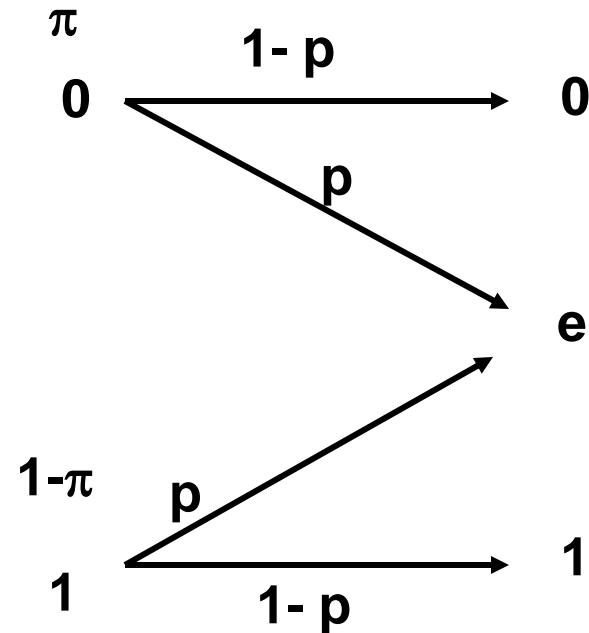
$$\begin{aligned} \blacklozenge I(X; Y) &= H(Y) - H(Y|X) \\ &= H(Y) - \sum p(x) H(Y|X=x) \\ &= H(Y) - H(p) \\ &\leq 1 - H(p) \end{aligned}$$

- ❖ The equality is when $H(Y) = 1$.
- ❖ Y is uniform when X is uniform.
- ❖ $p(x) \sim$ uniform.
- ❖ $C = 1 - H(p)$ bits.



Binary Erasure Channel (Off)

- ❖ Some bits are lost (no decision)
- ❖ $C = \max_{p(x)} I(X ; Y)$
 $= \max H(Y) - H(Y|X)$
 $= \max H(Y) - H(p)$
 $= \max_{\pi} \{H(\pi(1-p), p, (1-\pi)(1-p))$
 $\quad - H(p)\}$
 $= \max_{\pi} (1-p)H(\pi)$
 $\leq 1 - p$
- ❖ Let $E = \{Y = e\}$; $P(E)=p$; $P(E^c)=1-p$
- ❖ $H(Y) = H(Y, E) = H(Y) + H(E|Y)$
 $= H(E) + H(Y|E) = H(p)+(1-p)H(\pi)$



Binary Erasure Channel

- ❖ $C = 1 - p$
- ❖ With prob. p , the bit is lost.
- ❖ Thus, we could recover at most $1-p$ percentages of bits.

HW#4 (Due ?)

❖ Ch.5. 12, 16, 18, 20

❖ Ch. 8. 1, 5

❖ Midterm ?.

Information Theory

Channel Capacity Module

Tentative Schedule

<i>Weekly Course Schedule</i>		
<i>Calendar</i>	<i>Description</i>	<i>*Remarks</i>
1st week, 9/2	<i>Introduction to Information Theory, Entropy</i>	
2nd week, 9/7, 9	Entropy, Relative Entropy and Mutual Information	
3rd week, 9/14, 16	Entropy, Relative Entropy and Mutual Information	
4th week, 9/2,23	Asymptotic Equipartition Property	
5th week, 9/28, 30	Asymptotic Equipartition Property/Entropy Rates of a Stochastic Process	
6th week, 10/5, 7	Entropy rates of Markov Chain	
7th week, 10/12, 14	Data compression	
8th week, 10/19, 21	Data compression/Channel capacity	Midterm 10/21
9th week, 10/26, 28	Channel capacity theorems/forward/reverse	10/28OFF, 보충 Friday(10:30am)
10th week, 11/2, 4	Differential entropy	
11th week, 11/9, 11	Gaussian channel capacity	Selection of papers due M,W OFF, 보충 Friday 9:00am
12th week, 11/16, 18	MIMO channel capacity theorem	보충 Friday(9:00am)?
13th week, 11/23, 25	Multiple access channel capacity theorem	
14th week, 11/30, 12/2	Slepian Wolf	
15th week, 12/6, 9	Student Presentation	
16th week, 12/16	Final Exam (12/16 일)	

Agenda

- ❖ List of Papers
- ❖ Channel Coding Theorem (Chapter 7)

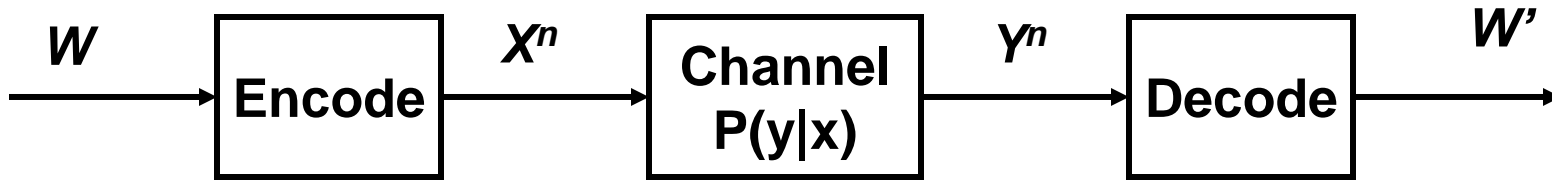
Skip Some Sideline Topics

❖ Chapter 6: Gambling and Data Compression

Channel Capacity

- ❖ Capacity1 = $\log(\text{Size of the Message Set})$
- ❖ T = Time required to transmit a message in the set.
- ❖ Capacity = Capacity1/T, bits/sec
- ❖ Capacity = Capacity in n -use of channel / n
= *bits/channel use*
- ❖ Capacity is the number of bits that can be transferred over a given channel with almost no error.
- ❖ The capacity is achieved in reality in n -channel uses.
 - Operational capacity.
 - Why?

Discrete Channel



- ❖ Input alphabet X , output alphabet Y , and the probability transition matrix $p(y|x) \sim$ prob. of observing y given x transmitted.
- ❖ The channel is memoryless if $P(y^n|x^n) = \prod p(y|x)$.
- ❖ We can transform the physical channel into the discrete channel.
- ❖ Example) A cable with a certain bandwidth F_{\max} and noise power spectral density \rightarrow Sample at twice the bandwidth and obtain the discrete channel (sample IN sample OUT).

“*Information*” Channel Capacity

❖ $C = \max_{p(x)} I(X; Y)$

- The maximum is taken over all input distr. $p(x)$
- Compared with the “*operational*” channel capacity ~ the highest rate in bits/channel-use, at which information can be sent at arbitrary small probability of error
- Shannon proved that the *operational* channel capacity is equal to the *information* channel capacity

Noiseless Binary Channel

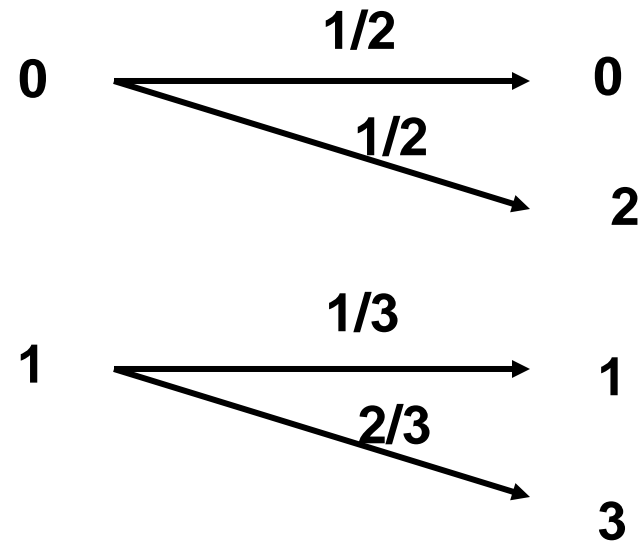
❖ $C = \max_{p(x)} I(X; Y) = ?$

0 → 0

1 → 1

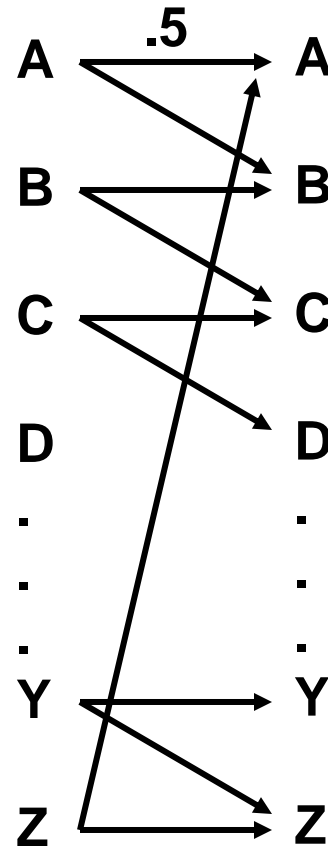
Noisy Channel with Non-Overlapping Output

❖ $C = ?$



Noisy Typewriter

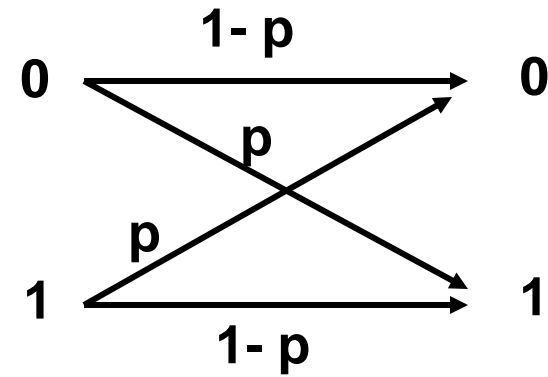
- ❖ The typewriter writes the input letter with prob. $1/2$ or the next letter in the alphabet with prob. $1/2$
- ❖ $C = ?$



Binary Symmetric Channel

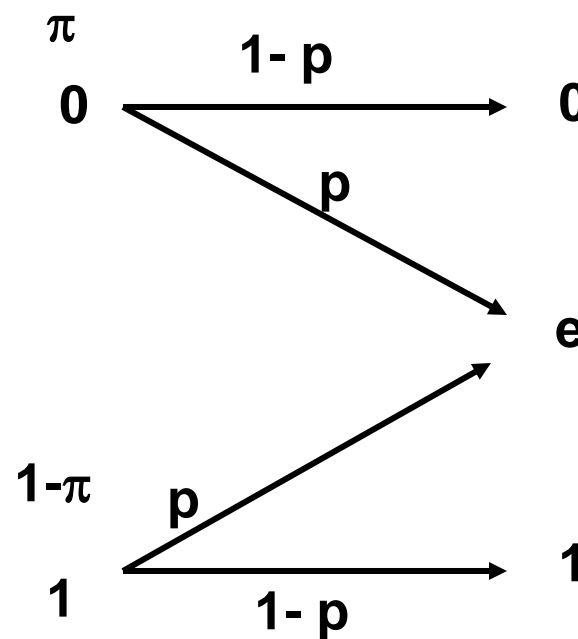
❖ $I(X; Y) = H(Y) - H(Y|X)$

❖ $C = ?$



Binary Erasure Channel

- ❖ Some bits are lost (no decision)
- ❖ $C = \max_{p(x)} I(X ; Y)$
 $= \max H(Y) - H(Y|X)$
 $= ?$
- ❖ Let $E = \{Y = e\}$; $P(E)=p$; $P(E^c)=1-p$
- ❖ $H(Y) = H(Y, E) = H(Y) + H(E|Y)$
 $= H(E) + H(Y|E) = H(p)+(1-p)H(\pi)$



Binary Erasure Channel

- ❖ $C = 1 - p$
- ❖ With prob. p , the bit is lost
- ❖ Thus, we could recover at most $1-p$ percentages of bits

Symmetric Channels

- ❖ $p(y | x) = [0.5 \ 0.1 \ 0.4; 0.4 \ 0.5 \ 0.1; 0.1 \ 0.4 \ 0.5]$
 - x-th row, y-th column
 - All the rows are permutations of each other and so are the columns

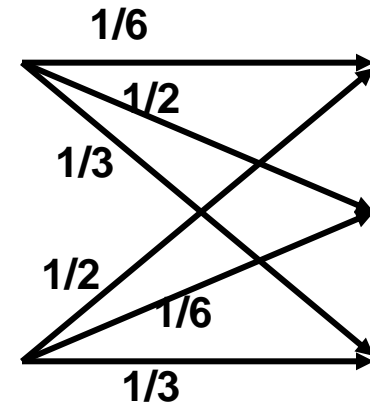
- ❖ $I(X ; Y) = H(Y) - H(Y | X)$
 - $= H(Y) - H(\text{any row})$
 - $\leq \log|\mathcal{Y}| - H(0.1, 0.4, 0.5)$

- ❖ The equality, when?

- ❖ Thus, the answer is?

Weakly Symmetric

- ❖ $p(y | x) = [1/6 \ 1/2 \ 1/3; 1/2 \ 1/6 \ 1/3]$
 - Rows are permutation of each other
 - Column sums are equal
- ❖ $C = \log|\mathcal{Y}| - H(\text{any row})$, which is achieved when?



Properties of Channel Capacity

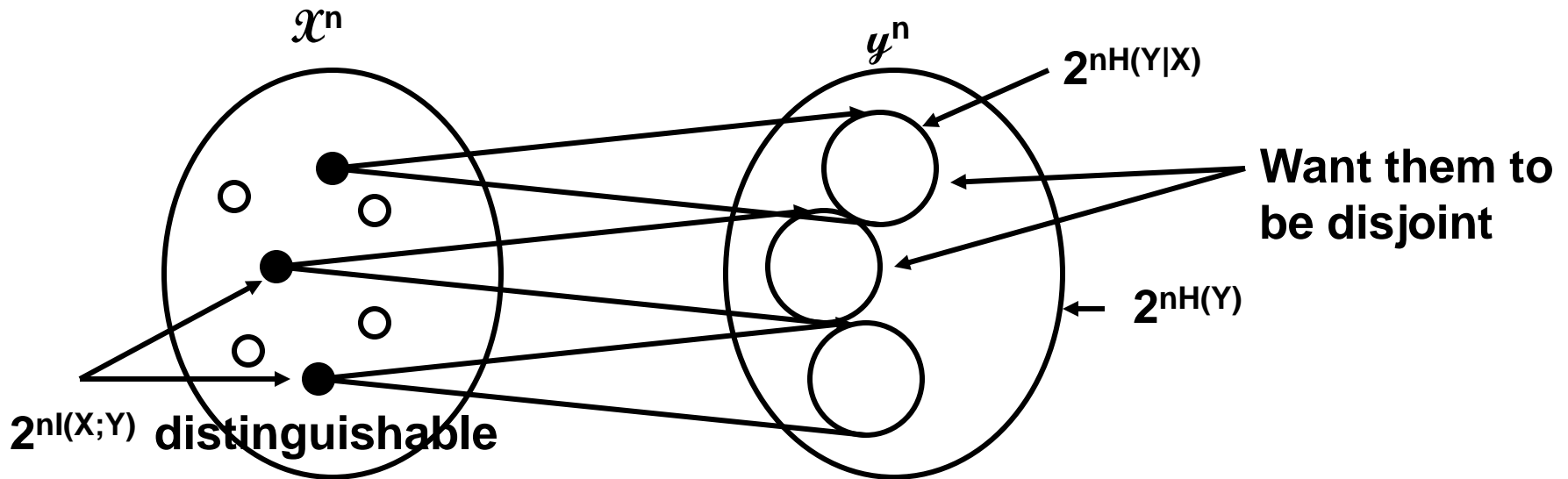
1. $C \geq 0$, since $I(X; Y) \geq 0$
2. $C \leq \log|X|$, why?
3. $C \leq \log|Y|$,
4. $I(X; Y)$ is a continuous function of $p(x)$.
5. $I(X; Y)$ is a concave \cap function of $p(x)$ for fixed $p(y|x)$.
 - ❖ Or $I(X; Y)$ is a convex \cup function of $p(y|x)$ for fixed $p(x)$.

$I(X; Y)$, concave \cap on $p(x)$ for fixed $p(y|x)$

- ❖ $I(X; Y) = H(Y) - \sum_x p(x) H(Y| X = x)$
 - $p(y) = \sum_x p(y|x) p(x)$ is a linear combination of $p(x)$
 - $H(Y)$ is a concave func. of $p(y)$
 - Thus, $H(Y)$ is a concave func. (of a linear combination) of $p(x)$
 - The second term is a linear function of $p(x)$

Channel Coding Theorem (Idea)

- ❖ Recall the typical set argument
- ❖ Recall the noisy type writer problem
- ❖ Consider the set of n symbol sequences
- ❖ $2^{n[H(Y) - H(Y|X)]} = 2^{n I(X; Y)}$, **how many distinguishable seq.'es?**

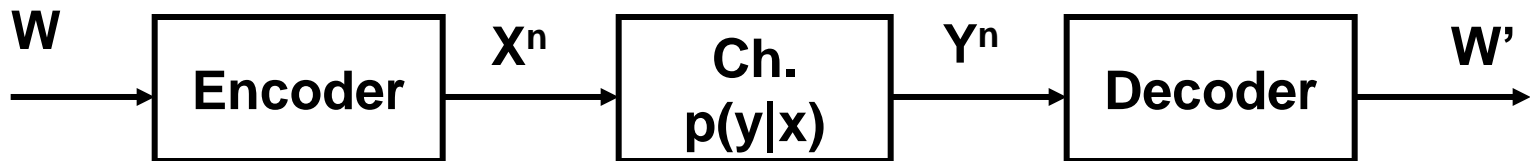


Channel Coding Theorem (Idea) (2)

- ❖ The total number of disjoint sets $\leq 2^{nI(X; Y)}$
- ❖ That's the maximum number of distinguishable sequences you can send at the transmitter while satisfying the close-to-zero decision error requirement.
- ❖ What's the role of a large n ?

Definitions

- ❖ A message set $W = \{1, 2, \dots, M\}$
- ❖ A signal $X^n(W)$
- ❖ A received signal Y^n
- ❖ The channel is $p(y^n|x^n)$
- ❖ $W' = g(Y^n)$, a decision made at the receiver
- ❖ $E = \{W' \neq W\}$, an error event



A discrete channel

- ❖ A discrete channel $(\mathcal{X}, p(y|x), \mathcal{Y})$
- ❖ The n -th extension of the discrete memoryless channel (DMC) is $(\mathcal{X}^n, p(y^n|x^n), \mathcal{Y}^n)$
 - $P(y^n|x^n) = \prod_{i=1}^n p(y_i|x_i)$

An (M, n) code for $(\mathcal{X}, p(y|x), \mathcal{Y})$

- ❖ A message index set $W = \{1, 2, \dots, M\}$
- ❖ An encoder is a function, i.e., $X^n: W \rightarrow \mathcal{X}^n$, yielding a codeword for each message index from a code $\{X^n(1), X^n(2), \dots, X^n(M)\}$.
- ❖ A decoder is a function which maps the received sequence to the index:

$$g: \mathcal{Y}^n \rightarrow \{1, 2, \dots, M\}$$

Probability of Error

- ❖ Probability of error, given i^{th} message:

$$\lambda_i = \Pr(g(Y^n) \neq i | X^n = X^n(i)) = \sum_{y^n} p(y^n | X^n(i)) \mathbf{I}(g(y^n) \neq i)$$

↑
Indicator func.

- ❖ The maximal prob. of error $\lambda^{(n)}$

$$\lambda^{(n)} = \max_{i \in \{1, 2, \dots, M\}} \lambda_i$$

- ❖ Probability of error is the average of λ_i over all equally probable message index i :

$$P_e^{(n)} = \Pr(\text{a wrong decision}) = (1/M) \sum_{i=1}^M \lambda_i \leq \lambda^{(n)}$$

The rate R of an (M, n) code

- ❖ $R = \log_2(M)/n$ [bits/transmission] (bits/channel-use)
- ❖ We say a rate R is achievable if there exists a sequence of $(\lceil 2^{nR} \rceil, n)$ codes with the (maximal) prob. of error tends to 0 as $n \rightarrow \infty$.
- ❖ The capacity is the supremum of all achievable rates.
- ❖ Thus, if $R < C$, then the prob. of error can be made vanishing to zero.

Jointly Typical Sequences

- ❖ Consider a jointly typical set $A_\varepsilon^{(n)}$
- ❖ It is the set of all $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ such that
 1. $\Pr[(X^n, Y^n) \in A_\varepsilon^{(n)}] \rightarrow 1$ as $n \rightarrow \infty$
 2. $|A_\varepsilon^{(n)}| \leq 2^{n(H(X, Y) + \varepsilon)}$
 3. $\Pr((X'^n, Y'^n) \in A_\varepsilon^{(n)}) \leq 2^{-n(I(X; Y) - 3\varepsilon)}$, where X'^n and Y'^n are two independently drawn random variables, one from $p(x^n)$ and the other from $p(y^n)$.

Jointly AEP (Proof)

❖ WLLN

❖ $H'_x := -(1/n) \log(p(x^n)) \rightarrow H(X)$ in probability

– For $\forall \varepsilon > 0, \exists n_1$, i.e., for $n > n_1$,

$$\Pr(|H'_x - H(X)| > \varepsilon) < \varepsilon/3 \quad \text{----- (1)}$$

❖ Similarly for H'_y

– $H'_y := -(1/n) \log(p(y^n)) \rightarrow H(Y)$ in probability

– For $\forall \varepsilon > 0, \exists n_2$, i.e., for $n > n_2$,

$$\Pr(|H'_y - H(Y)| > \varepsilon) < \varepsilon/3 \quad \text{----- (2)}$$

❖ Similarly for $H'_{x,y}$

– $H'_{x,y} = -(1/n) \log(p(x^n, y^n)) \rightarrow H(X, Y)$ in probability

– For $\forall \varepsilon > 0, \exists n_3$, i.e., for $n > n_3$,

$$\Pr(|H'_{x,y} - H(X, Y)| > \varepsilon) < \varepsilon/3 \quad \text{----- (3)}$$

Jointly AEP (Proof) (2)

- ❖ Now, choose $n > \max\{n_1, n_2, n_3\}$
- ❖ The prob. of the union of the three sets in previous page is smaller than ε .

Jointly AEP (Proof) (3)

$$\begin{aligned} \blacklozenge \quad 1 &= \sum p(\mathbf{x}^n, \mathbf{y}^n) \\ &\geq \sum_{\mathcal{A}_\varepsilon^{(n)}} p(\mathbf{x}^n, \mathbf{y}^n) \\ &\geq |\mathcal{A}_\varepsilon^{(n)}| 2^{-n(H(X, Y) + \varepsilon)} \\ \blacklozenge \quad \text{Thus, } |\mathcal{A}_\varepsilon^{(n)}| &\leq 2^{n(H(X, Y) + \varepsilon)}. \end{aligned}$$

X'^n and Y'^n independent

- ❖ X'^n and Y'^n independent, but have the same marginals as X^n and Y^n .

❖ $p(x, y) =$

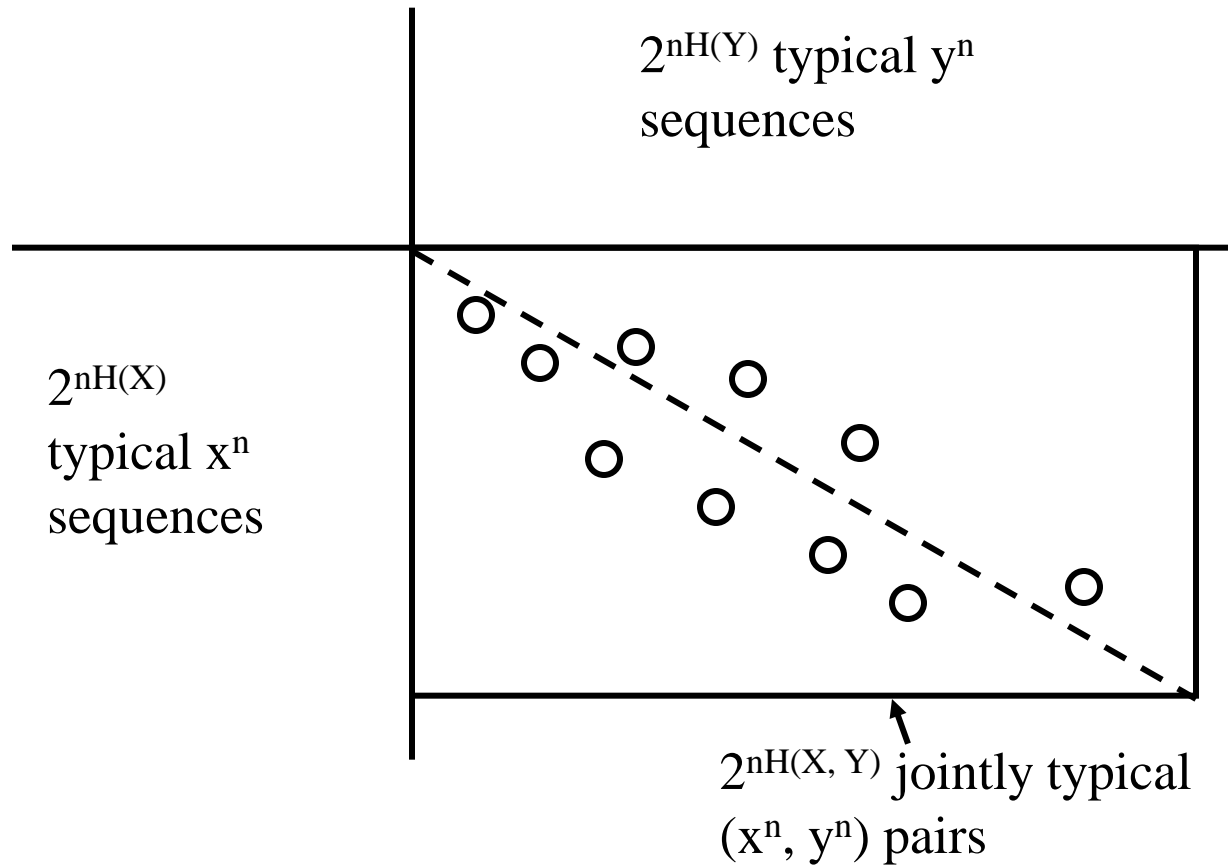
$x \backslash y$	0	1
0	1/4	1/4
1	1/8	3/8

$x \backslash y$	0	1
0	3/16	5/16
1	3/16	5/16

$p(x)p(y) =$

- ❖
$$\begin{aligned} \Pr((X'^n, Y'^n) \in A_\varepsilon^{(n)}) &= \sum_{A_\varepsilon^{(n)}} p(x^n)p(y^n) \\ &\leq 2^{n(H(X, Y) + \varepsilon)} 2^{-n(H(X) - \varepsilon)} 2^{-n(H(Y) - \varepsilon)} \\ &= 2^{-n(I(X; Y) - 3\varepsilon)} \end{aligned}$$

Illustration of Jointly Typical Set



$$\frac{2^{nH(X,Y)}}{2^{nH(X)} \times 2^{nH(Y)}} = 2^{-nI(X;Y)}$$

Now, consider independently drawn pair (x'^n, y'^n) ,
 $\Pr(\text{this pair belongs to } A_\epsilon^{(n)}) = 2^{-n(I(X;Y)-3\epsilon)}$.

Achievability of Channel Capacity

❖ Sketch of new ideas Shannon used

- Allowing arbitrarily small but non-zero error probability
- Successive use of channels (utilize the LLN)
- Calculate the average performance of codebooks, rather than that of a single codebook.

❖ Sketch of proof

- Random code selection
- Calculate the average prob. of error for a random choice of codewords
- Decode by joint-typicality; look for a codeword that is jointly typical with the received word.

Shannon's Channel Coding Theorem

- ❖ All rates below capacity C are achievable. Specifically, for rate $R < C$, there exists a sequence of $(2^{nR}, n)$ codes with maximum probability of error $\lambda^{(n)} \rightarrow 0$.
- ❖ Conversely, any sequence of $(2^{nR}, n)$ codes with $\lambda^{(n)} \rightarrow 0$ must have $R \leq C$.

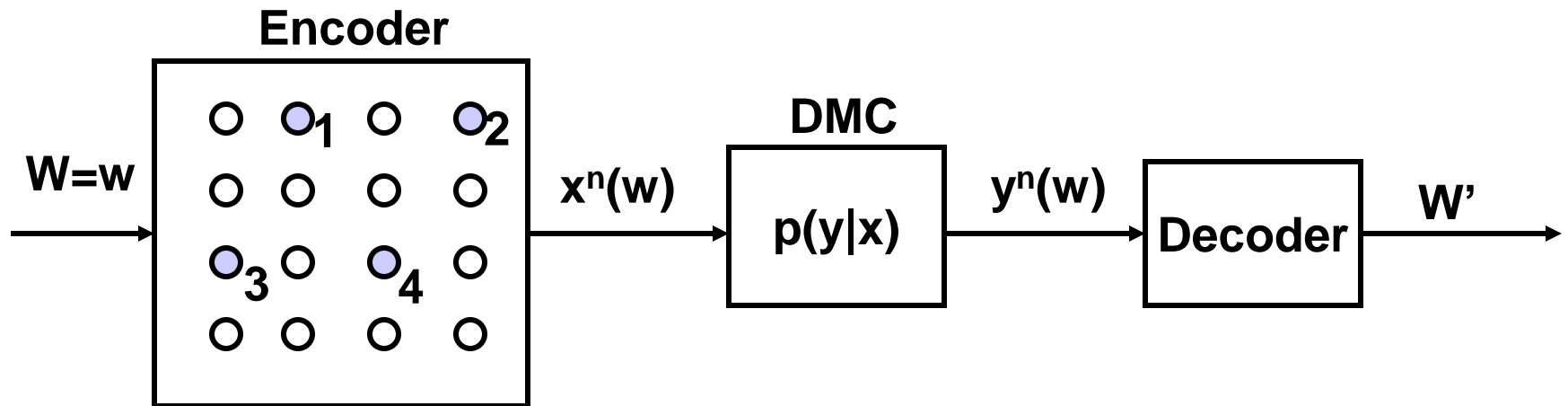
$R < C \Rightarrow (2^{nR}, n)$ codes arbitrary small error (Achievability)

- ❖ $p(x)$ fixed
- ❖ Generate a $(2^{nR}, n)$ code at random $\sim p(x^n) = \prod_{i=1}^n p(x_i)$
 - Select $M=2^{nR}$ codewords randomly
 - Codewords are rows
 - $$\mathcal{C} = \begin{pmatrix} x_1(1) & x_2(1) & \dots & x_n(1) \\ \dots & \dots & \dots & \dots \\ x_1(M) & x_2(M) & \dots & x_n(M) \end{pmatrix}$$
- ❖ $\Pr(\mathcal{C}) = \prod_{m=1}^M \prod_{i=1}^n p(x_i(m))$,
 - All the elements in the matrix are selected iid from $p(x)$.
- ❖ The use of this code \mathcal{C} is known to both transmitter and the receiver.
- ❖ The $p(y|x)$ is known to both transmitter and receiver.
- ❖ A message $W=w$ is chosen uniformly among $\{1, 2, \dots, M\}$ and sent over the channel.

Block Diagram

- ❖ The w -th codeword $x^n(w)$, the w -th row of \mathcal{C} , is sent
- ❖ The receiver receives a sequence y^n , according to the distribution

$$P(y^n|x^n(w)) = \Pr\{Y^n = y^n|x^n(w)\} = \prod p(y_i | x_i(w))$$



Achievability (2)

- ❖ The optimal receiver is the maximum likelihood decoder
- ❖ But here, let's use the typical set decoding, suboptimal but gives easier proof of achievability
- ❖ Jointly typical decoding rule:
 - The receiver decide $W'=q$, if $(x^n(q), y^n)$ is jointly typical and no other codeword is jointly typical with y^n
- ❖ $E = \{W' \neq W\}$, a decoding error event

Probability of Error

❖ Probability of an error

$$\lambda_i = \Pr(g(Y^n) \neq i | X^n = x^n(i)) = \sum_{y^n} p(y^n | x^n(i)) \mathbf{I}(g(y^n) \neq i)$$

↑
Indicator func.

❖ The maximal prob. of error $\lambda^{(n)}$

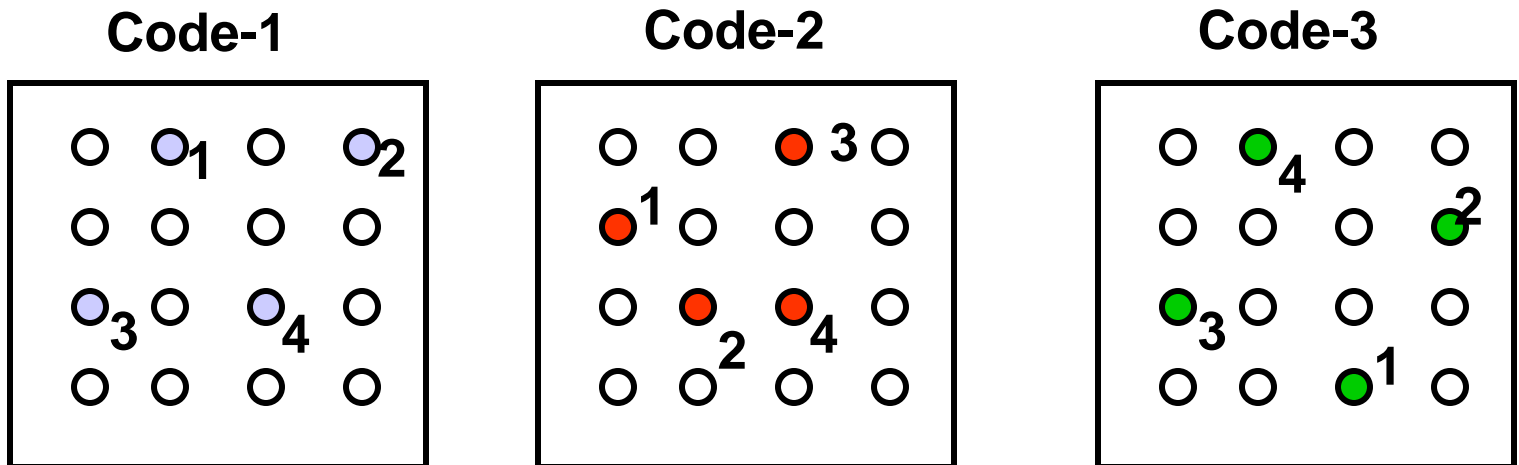
$$\lambda^{(n)} = \max_{i \in \{1, 2, \dots, M\}} \lambda_i$$

❖ Arithmetic average:

$$P_e^{(n)} = \Pr(\text{a wrong decision}) = (1/M) \sum_{i=1}^M \lambda_i \leq \lambda^{(n)}$$

Achievability (3)

- ❖ Consider a prob. of error for a particular word w of a particular code \mathcal{C}
 - It's difficult to calculate $\lambda_w(\mathcal{C})$ because codewords in \mathcal{C} are chosen random
- ❖ But if averaged over all codebooks
 - $\sum_c P(\mathcal{C}) \lambda_i(\mathcal{C}) = \sum_c P(\mathcal{C}) \lambda_j(\mathcal{C})$, for any i and j , by symmetry



Achievability (4)

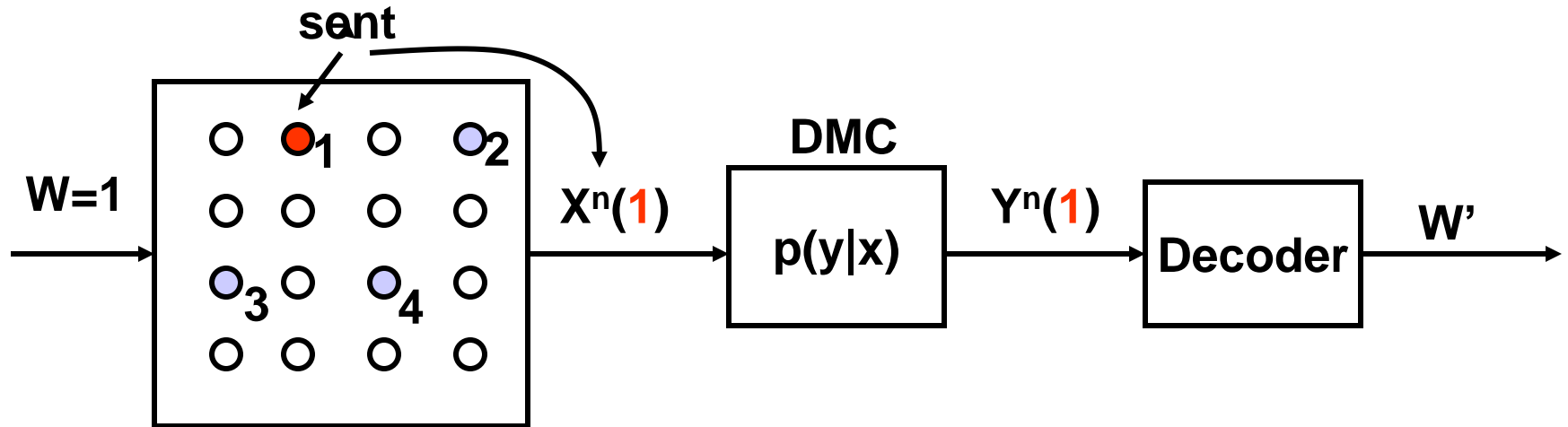
$$\begin{aligned} \diamond \Pr\{E\} &= \sum_{\mathcal{C}} P(\mathcal{C}) P_e(\mathcal{C}) \\ &= \sum_{\mathcal{C}} P(\mathcal{C}) (1/M) \sum_{w=1}^M \lambda_w(\mathcal{C}) \\ &= (1/M) \sum_{w=1}^M \sum_{\mathcal{C}} P(\mathcal{C}) \lambda_w(\mathcal{C}) \\ &= (1/M) M \sum_{\mathcal{C}} P(\mathcal{C}) \lambda_1(\mathcal{C}) \\ &= \sum_{\mathcal{C}} P(\mathcal{C}) \lambda_1(\mathcal{C}) \\ &= \Pr\{E \mid W = 1\} \end{aligned}$$

◇ Now, the problem is reduced to

“given the first message sent, find the error probability.”

Achievability (5)

- ❖ $X^n(1)$ is independent with $X^n(i)$, $i=2, 3, \dots, M$.
- ❖ $Y^n(1)$ is also independent with $X^n(i)$, $i=2, 3, \dots, M$.



Achievability (6)

❖ Let's define the events

$$E_i = \{(X^n(i), Y^n) \in A_\varepsilon^{(n)}\}, i = 1, 2, \dots, M,$$

where the i -th codeword is jointly typical with Y^n .

❖ Now, going back to $\Pr(E | W=1)$

$$\Pr(E | W=1) = P(E_1^c \cup E_2 \cup E_3 \cup \dots \cup E_M)$$

--- by the union of events bound

$$\leq P(E_1^c) + \sum_{i=2}^M P(E_i)$$

We note that as $n \rightarrow \infty$

$$P(E_1^c) \rightarrow 0$$

$$P(E_i) \rightarrow 2^{-nI(X;Y)}$$

$$\leq \varepsilon + (M - 1) 2^{-n(I(X;Y) - 3\varepsilon)}, \text{ for } n \text{ suffi. large}$$

Achievability (7)

$$\begin{aligned} \blacklozenge \Pr(\mathbf{E} \mid W=1) &\leq \varepsilon + (M-1) 2^{-n(I(X;Y) - 3\varepsilon)}, \text{ for } n \text{ suff. large} \\ &= \varepsilon + (2^{nR} - 1) 2^{-n(I(X;Y) - 3\varepsilon)} \\ &\leq \varepsilon + 2^{-n(I(X;Y) - R - 3\varepsilon)} \end{aligned}$$

--- as long as $R < I(X;Y) - 3\varepsilon$

$$\leq \varepsilon$$

\blacklozenge Thus, for n sufficiently large, $P(\mathbf{E})$ can be made very small.

Achievability (8)

- ❖ Now choose $p(x)$ to be the optimal $p^*(x)$ that maximizes the mutual information $I(X; Y)$. Then, the condition $R < I(X; Y)$ becomes $R < C$.
- ❖ Since the prob. of error averaged over all possible codebooks is small ($\leq 2\varepsilon$), there exists at least one codebook \mathcal{C}^* with a small prob. of error $P_e(\mathcal{C}^*)$. Such a code can be found by an exhaustive search.
- ❖ Throw away the worst half codewords
 - The rate becomes $R \approx R - 1/n$
 - $(2/M) \sum_{\text{best half}} \lambda_i < 4\varepsilon$
 - $\lambda_{\text{worst}} < 4\varepsilon$

Achievability (9)

- ❖ Thus, we have constructed a code $R' = R - 1/n$, with maximal probability of error $\lambda < 4\epsilon$

Q.E.D. on Achievability

- ❖ People had believed this random coding method was only for proof—no practical guidance as a coding method.
- ❖ In fact, it was found to be the most powerful coding method which achieves the capacity very closely:
 - Turbo code (1993), Low Density Parity Check code (1962)
 - These codes are long block codes with built-in random component.
 - Refer to Lecture notes in Channel Coding Theory.

Zero Error Codes

❖ Precursor to the converse of the channel coding theorem

❖ Show if $P_e^{(n)} = 0$, then $R \leq C$

❖ $nR = H(W) = H(W|Y^n) + I(W; Y^n)$

$= 0$ --- Zero error codes Y^n fully determines the input W

$$= I(W; Y^n)$$

$$\leq I(X^n; Y^n)$$

--- DPI

$$\leq \sum_{i=1}^n I(X_i; Y_i)$$

--- DMC

$$\leq nC$$

❖ Thus, for any zero error $(2^{nR}, n)$ code, for all n ,

$$R \leq C$$

Fano's Inequality for Converse to CCT

- ❖ Recall the Fano's inequality
- ❖ Let $E = 1$ for $W' = W$, 0 for $W' \neq W$
- ❖ $P(E=1) = P_e^{(n)}$, $P(E = 0) = 1 - P_e^{(n)}$,
- ❖ Consider two expansions of $H(E, W | Y^n)$
- ❖
$$\begin{aligned} H(E, W | Y^n) &= H(E | Y^n) + H(W | E, Y^n) \\ &= H(W | Y^n) + H(E | W, Y^n) \end{aligned}$$
$$\Rightarrow H(E | Y^n) + H(W | E, Y^n) = H(W | Y^n) + \mathbf{0}$$

Fano's Inequality for Converse to CCT

$$\blacklozenge H(E|Y_n) + H(W|E, Y^n) = H(W|Y^n)$$

$$1. H(E|Y_n) \leq H(E) = H(P_e^{(n)}) \leq 1$$

$$\begin{aligned} 2. H(W|E, Y^n) &= P(E=0) H(W|E=0, Y^n) + P(E=1) H(W|E=1, Y^n) \\ &= P(E=0) \times 0 + P_e^{(n)} \log_2(M-1) \\ &\leq P_e^{(n)} \log_2(2^{nR}) \quad \text{--- } M = 2^{nR} \\ &= P_e^{(n)} nR \end{aligned}$$

$$3. H(W|Y^n) \leq H(X^n|Y^n)$$

\blacklozenge Thus, we have

$$H(X^n|Y^n) \leq 1 + P_e^{(n)} nR$$

Converse to the Channel Coding Theorem

❖ Show: Any sequence of $(2^{nR}, n)$ codes with $\lambda^{(n)} \rightarrow 0$
 $\Rightarrow R \leq C$

❖ $nR = H(W) = H(W|Y^n) + I(W; Y^n)$

--- DPI (equality if $X^n(W)$ is one-to-one)

$$\leq H(W|Y^n) + I(X^n(W); Y^n)$$

---Fano's inequality

$$\leq 1 + P_e^{(n)} nR + I(X^n(W); Y^n)$$

--- DMC & Capacity

$$\leq 1 + P_e^{(n)} nR + nC$$

❖ Thus, we have $R \leq (1/n) + P_e^{(n)} R + C$. And,

$$R \leq C \text{ as } n \rightarrow \infty$$

HW 5

❖ Prob. 1 Read the channel capacity proof in Chap 7.

- What is the optimal decoding scheme for Hamming codes defined in 7.11 for BSC? What is the code turned out to be not useful?
- How many errors can the Hamming code in 7.11 correct? How many erasures can the code correct?
- Give an example of two erasures that can be corrected.

❖ Prob. 2 Consider a channel defined as $Y = (X + E) \bmod 2$ where E is Bernoulli($p = 0.1$) with alphabet $\{0, 1\}$. X is the input symbol $\{1, 0\}$ and Y is the output symbol. Consider a set of randomly generated codebooks of length 10.

1. Suppose we choose a rate $R = .4$. Is this choice making sense? If no, choose your own. Construct a single codebook of length $n = 10$.
2. Design a Maximum Likelihood decision receiver for your codebook.
3. Now consider we use Eq. (7.66) in the textbook. How many different codebooks can we obtain?
4. Calculate $\Pr(\mathcal{C})$ exactly in your case.
5. Show in your case how $\Pr(E) = \Pr(E|W = 1)$.
6. Would your ML receiver produce $\Pr(E)$ smaller than p in average sense? Prove. Also show that $\Pr(E)$ is getting smaller as n increases.

HW5

- ❖ Prob.3 Prove the capacity of BSC again with the lead $I(X; Y) = H(X) - H(X|Y)$.

- ❖ Problems from Cover & Thomas
 - P7.2, P7.3, P7.4, P7.8, P7.9, P7.13

HW 6

- ❖ Prob. 1 (Read Chapter 7) Answer the following questions with plain English sentences, three maximum.
- Describe the source-channel separation theorem.
 - Suppose the availability of a feedback channel where the received sample is sent back to the transmitter immediately upon reception. Describe why this feedback does not increase the capacity.

HW#6

❖ **Prob. 2 (Typical set size)** In class, we have discussed a little bit about the size of typical set $2^{nH(Y|X)}$ for a BSC with parameter p .

(a) The following inequality shall provide insight for that discussion. Show that for any d in $\{0, 1, 2, \dots, n\}$

$$\binom{n}{d} \leq 2^{nH(\frac{d}{n})}$$

where $H(*)$ is the binary entropy function. (Hint: use the Binomial expansion with parameter $p = d/n$)

(b) Find the size of typical set for $n = 100$ and $\epsilon = 0.2$ for BSC with $p = 0.1$. Suppose the use of the typical set decoder. Determine the set of number of channel errors that the decoder can correct. Determine the capacity and the probability error (a bound is good enough) at the given length.

(c) Suppose the transmitted codeword is the all-zero sequence. What's obtained from the channel is a sequence of two ones and 98 1s. Is the result jointly typical with the setting given in (b)?

HW#6

❖ Cover & Thomas Problems

P7.13 (Cap for erasures+erros)

P7.16 (Encoder/Decoder as part of the channel)

P7.25 (Bottleneck channel)

P7.26 (Noisy typewriter)

P7.28 (Choice of channels)

P7.34 (Capacity of various channels)

P7.36 (Capacity with memory)

Information Theory

Gaussian Channel Capacity

Agenda

- ❖ Differential Entropy for continuous random variables (Chapter 8)
 - Refer Maximum Entropy (Chapter 11) as well.
- ❖ Fading Channels
- ❖ Gaussian Channel Capacity (Chapter 9)

Differential Entropy $h(X)$

❖ $h(X)$ with a density $f(x)$ is

$$h(X) = -\int_S f(x) \log f(x) dx$$

where S is the support set of the random variable.

❖ Ex1) $X \sim$ uniform distribution on $(0, a)$

$$h(X) = -\int_0^a (1/a) \log(1/a) dx = \log(a).$$

- when $a < 1$, $\log(a) < 0$ (Differential entropy can be negative!)
- But $2^{h(X)} = 2^{\log(a)} = a$ is the volume of the support set, which is always non-negative!

$h(X)$ on Gaussian Distribution

❖ $X \sim \mathcal{N}(0, \sigma^2)$: $f(x) = (1/\sqrt{2\pi \sigma^2}) \exp(-0.5x^2/\sigma^2)$

❖ In “nats”,

$$\begin{aligned}h(X) &= - \int f(x) \ln f(x) dx \\&= - \int f(x) [- 0.5 x^2/\sigma^2 + \ln(\sqrt{2\pi\sigma^2})] dx \\&= 0.5/\sigma^2 \int f(x) x^2 dx + \ln(\sqrt{2\pi\sigma^2}) \\&= 0.5 + \ln(\sqrt{2\pi \sigma^2}) \\&= 0.5 \log_e e + 0.5 \ln(2\pi \sigma^2) \\&= 0.5 \ln(2\pi\sigma^2 e) \quad [\text{nats}]\end{aligned}$$

– Use $\log_e(A) = \log_2(A)/\log_2(e)$ for [bits]

– $h(X) = ?$ [bits]

AEP for continuous r.v.

- ❖ X_1, X_2, \dots, X_n , *i.i.d.* according to $f(x)$
- ❖ $-(1/n) \log f(X_1, X_2, \dots, X_n) \Rightarrow E\{-\log f(X)\} = h(X)$ in probability
- ❖ Let A be the typical set
 - $p(A) \approx 1$
 - $P[(x_1, x_2, \dots, x_n) \text{ in } A] \approx 2^{-nh(X)}$
 - Size of the typical set = $\text{Vol}(A) \approx 2^{nh(X)}$

Vol(A)

$$\begin{aligned} \blacklozenge \quad 1 &= \int f(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n \\ &\geq \int_A f(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n \\ &\geq 2^{-n(h(X) + \varepsilon)} \int_A dx_1 dx_2 \dots dx_n \\ &= 2^{-n(h(X) + \varepsilon)} \text{Vol}(A) \end{aligned}$$

$$\blacklozenge \quad \text{Vol}(A) \leq 2^{n(h(X) + \varepsilon)}$$

Entropy of Multivariate Normal Distribution

- ❖ $(X_1, X_2, \dots, X_n) \sim \mathcal{N}(\mathbf{m}, \mathbf{K})$ where \mathbf{m} is the mean vector and \mathbf{K} is the covariance matrix

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} |\mathbf{K}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^t \mathbf{K}^{-1} (\mathbf{x}-\mathbf{m})}$$

- ❖ $h(X_1, X_2, \dots, X_n) = 0.5 \log(2\pi e)^n |\mathbf{K}|$ [bits]
where $|\mathbf{K}|$ is the determinant of \mathbf{K} .

Relative Entropy and Mutual Information

- ❖ $D(f \parallel g) = \int f \log f/g \geq 0$, equality when $f = g$.
 - Did we ever use entropy is non negative? If we did, then we have to prove it again, with a different approach.
 - What was this proof?
- ❖ $I(X; Y) = \int f(x, y) \log [f(x, y)/f(x)f(y)] dx dy$.
- ❖ Most things stay the same as those for the discrete r.v. case except a few.

Translation Invariant

$$\blacklozenge h(X + c) = h(X)$$

$$h(aX) = h(X) + \log|a|$$

- ❖ cf) Discrete r.v. $H(aX) = H(X)$
- ❖ Since this is **continuous** random variable, we need a compensation term $\log|a|$, just as in the change of variables in integral.
- ❖ Let $Y = aX$. Then, $f_Y(y) = (1/|a|) f_X(y/a)$,
$$\begin{aligned} h(aX) &= - \int f_Y(y) \log f_Y(y) dy \\ &= - \int (1/|a|) f_X(y/a) [\log(1/|a|) + \log f_X(y/a)] dy \\ &= \log|a| - \int (1/|a|) f_X(y/a) \log(f_X(y/a)) dy \\ &= \log|a| - \int f_X(x) \log(f_X(x)) dx \\ &= \log|a| + h(X) \end{aligned}$$

Gaussian has the maximum differential entropy (Under power constraint)

❖ Let the random vector $\mathbf{X} \in \mathbf{R}^n$ have zero mean and covariance, $\mathbf{K} = E\{\mathbf{X}\mathbf{X}^t\}$, i.e., $\mathbf{K}_{ij} = E\{X_i X_j\}$, $1 \leq i, j \leq n$.

❖ Then,

$$h(\mathbf{X}) \leq 0.5 \log(2\pi e)^n |\mathbf{K}|,$$

with equality *iff* $\mathbf{X} \sim \mathcal{N}(0, \mathbf{K})$

❖ Proof:

Consider $g(\mathbf{x})$ be any density with $\int x_i x_j g(\mathbf{x}) d\mathbf{x} = \mathbf{K}_{ij}$

Show $h(g) \leq h(f)$, where $f(\mathbf{x}) \sim \mathcal{N}(0, \mathbf{K})$.

Show $h(g) \leq h(f)$, where $f(\mathbf{x}) \sim \mathcal{N}(0, \mathbf{K})$

$$\blacklozenge 0 \leq D(g \parallel f)$$

$$= \int g \log g/f$$

$$= \int g \log(g) - \int g \log f$$

$$= -h(g) - \int g(\mathbf{x}) \log(f(\mathbf{x})) d\mathbf{x}$$

$$\text{----- } \log(f(\mathbf{x})) = \log(C_1) - 0.5(\mathbf{x}^t \mathbf{K}^{-1} \mathbf{x})$$

$$\text{----- } \int g(\mathbf{x}) (\mathbf{x}^t \mathbf{K}^{-1} \mathbf{x}) d\mathbf{x} = \int f(\mathbf{x}) (\mathbf{x}^t \mathbf{K}^{-1} \mathbf{x}) d\mathbf{x}$$

$$= -h(g) - \int f(\mathbf{x}) \log(f(\mathbf{x})) d\mathbf{x}$$

$$= -h(g) + h(f)$$

$\therefore h(g) \leq h(f)$, equality with $g = f$.

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} |\mathbf{K}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^t \mathbf{K}^{-1} (\mathbf{x}-\mathbf{m})}$$

Wireless Communications Channel

- ❖ To enrich our discussion, I'll briefly take a side step and talk about wireless communications channels a little bit.
- ❖ Namely, wireless channels exhibit signal strength fading phenomenon.
- ❖ We will see why and how to describe it.

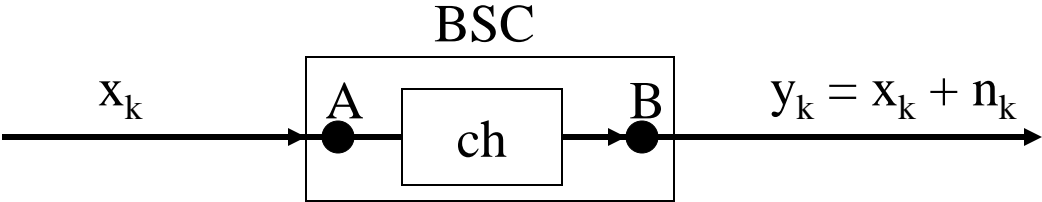
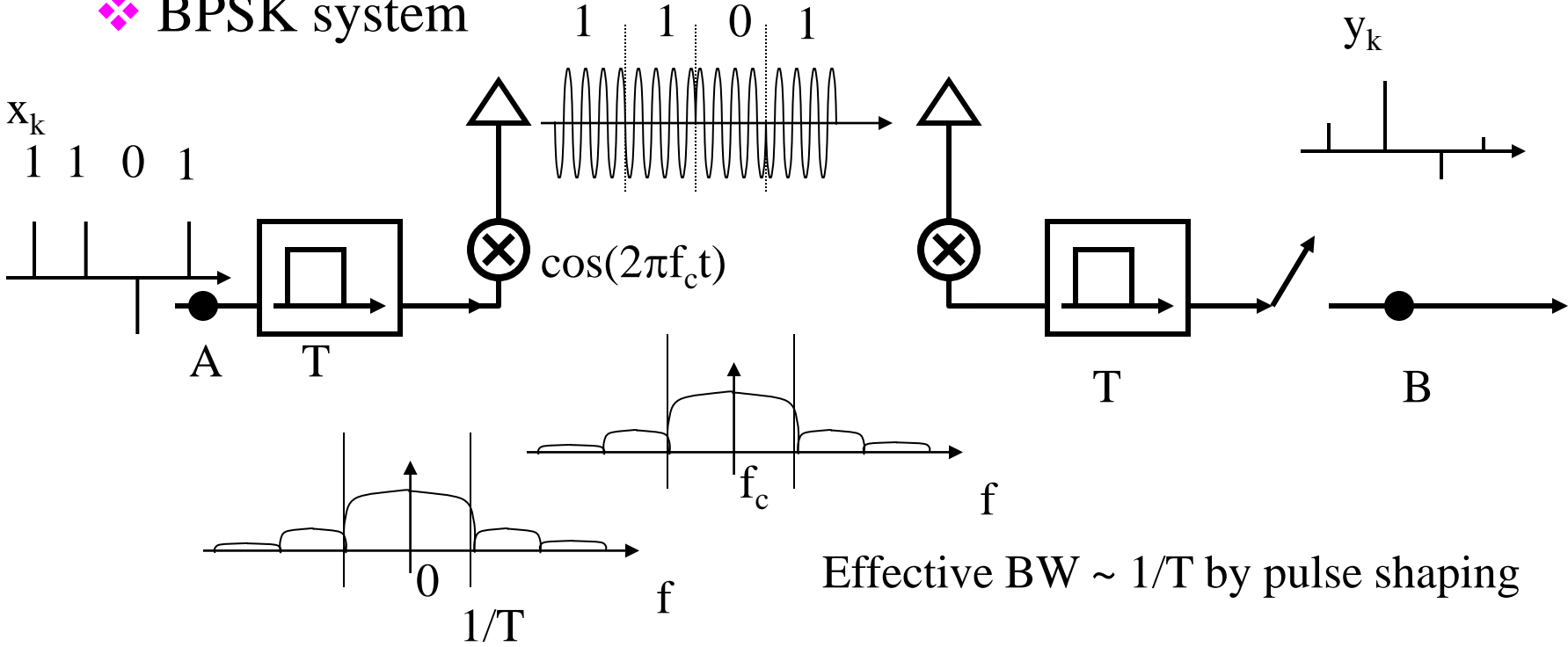
Fading Channels

- ❖ (What is it?) One of the characteristics of terrestrial wireless channels is the signal-strength fading.
- ❖ (How does it happen?) The culprit of fading is multi-path propagation of the radio-wave signals.
- ❖ (When does it happen?) Thus, the fading phenomenon becomes significant in rich scattering environment where a large number of reflectors and scatters are around at the receiver.
- ❖ Good transceiver design shall provide a solution that is not limited by the fading effect but the one that takes advantage of the phenomenon.
- ❖ Recent results show that we can utilize the signal strength fading to boost up the channel capacity.

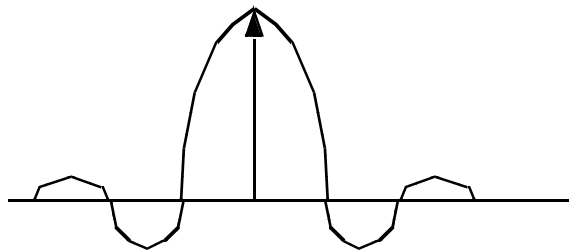
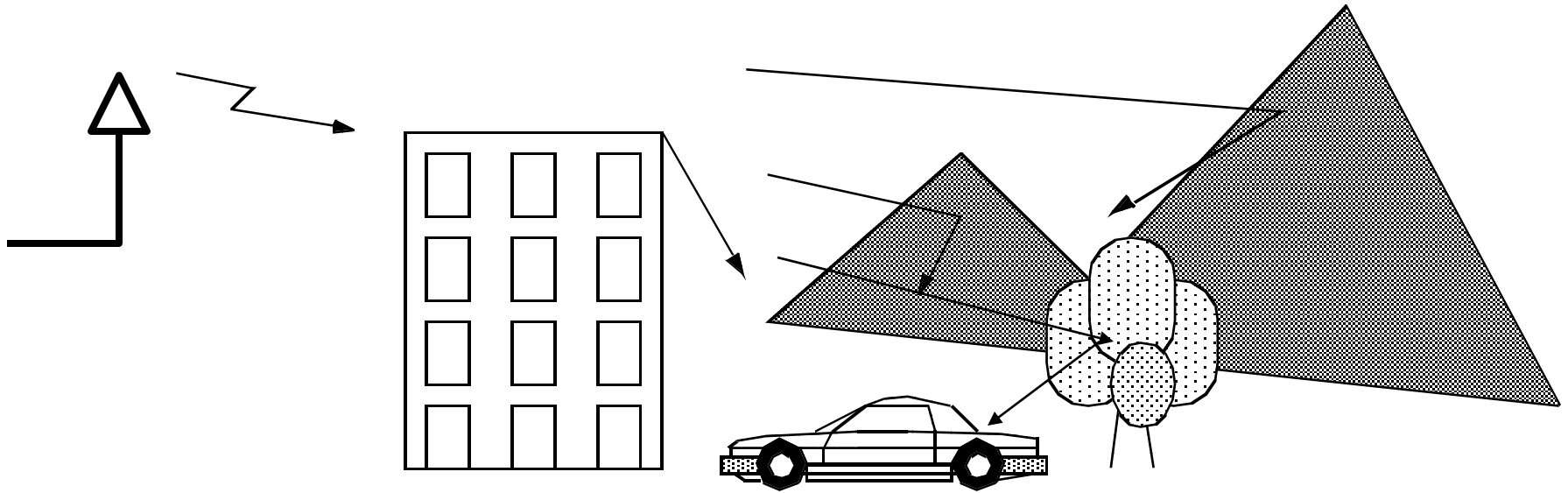
Quick Review on Modulation and Detection Theory

(From EE 1473)

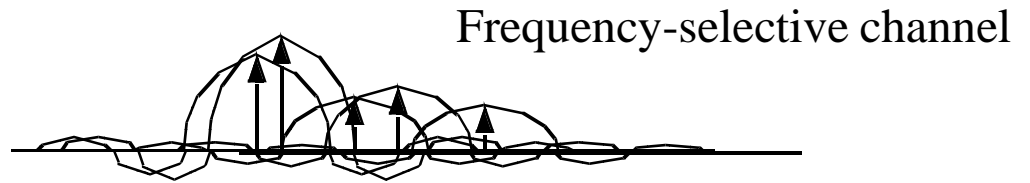
❖ BPSK system



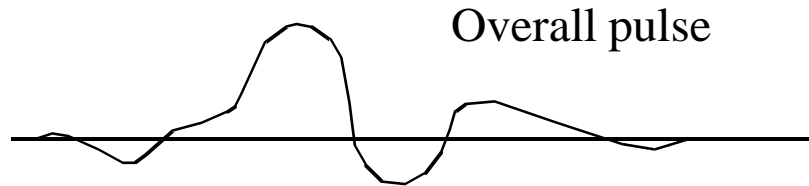
Multipath Fading ISI Channels



Transmitted pulse



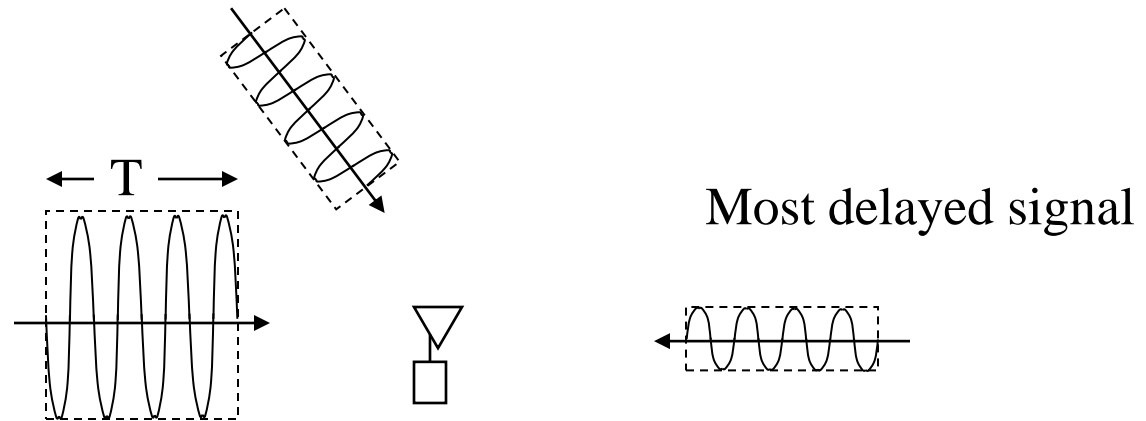
Frequency-selective channel



Overall pulse

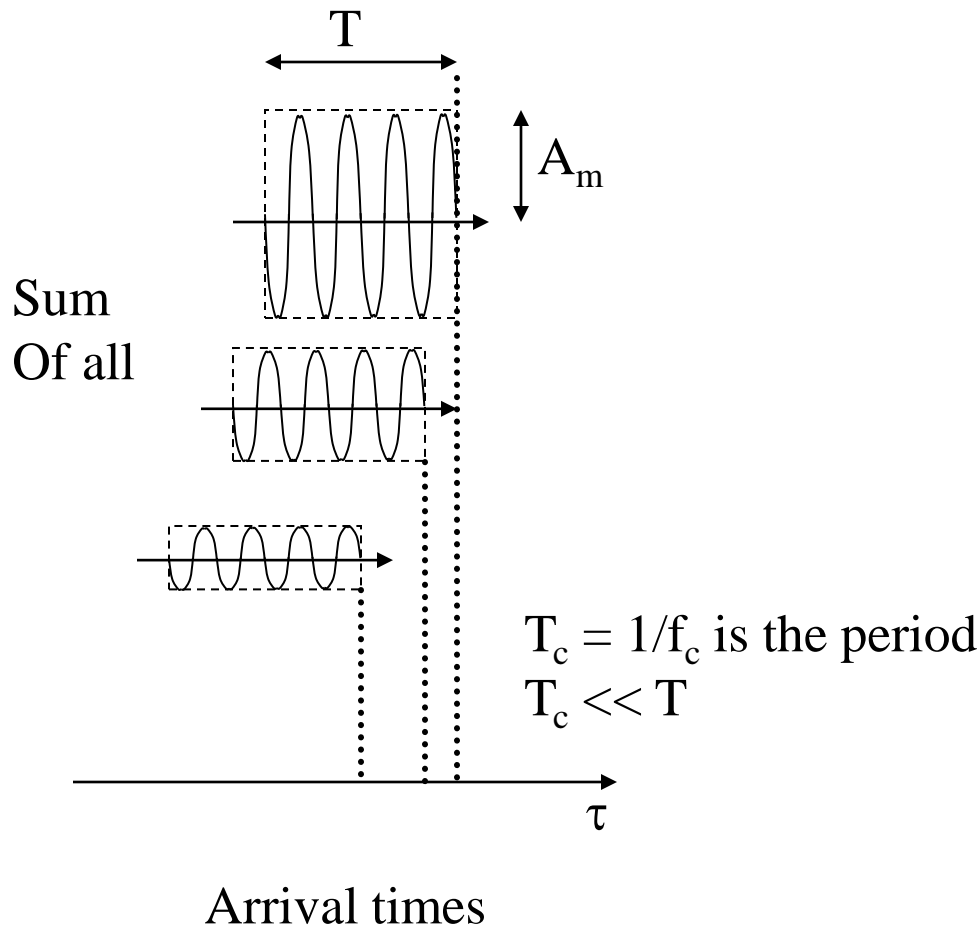
ISI occurs
327

What is happening at the receiver antenna?



- ❖ The phase and strength of the electro-magnetic excitation at the receiver's antenna is proportional to the summation of all the incidental waves arriving to the receiver antenna with different phases and magnitudes due to multipath propagation

Delay Spread vs. Fading Model



- ❖ When arrival-time differences are relatively small ($T_m \ll T$), the signal perceived at the receiver is the addition of all arriving sinusoids
- ❖ $G = \sum_m A_m \exp(j2\pi \theta_m)$
- ❖ Add constructively or destructively, depending on the phase
- ❖ The delay τ_m is translated to the phase difference $\theta_m \in (0, 2\pi)$
- ❖ Invoking the Central Limit Theorem, G is modeled as a complex-valued Gaussian

Fading Channel Statistics

- ❖ Complex-valued Gaussian Process $G(t)$
- ❖ For fixed t , $G(t)$ is complex-valued Gaussian r.v.
 - The distribution of the power, $|G(t)|^2$, is Chi-square
 - The distribution of the mag., $|G(t)|$, is **Rayleigh**
- ❖ PDF of Chi-square r.v. with two deg. of freedom
 - $Y := G(t)^* G(t) = G_r^2 + G_i^2$ where G_r and G_i are independent Gaussians with zero mean and variance σ^2
 - $P(y) = (1/2\sigma^2) \exp(-y/(2\sigma^2)) U(y)$
- ❖ PDF of Rayleigh r.v.
 - $R := Y^{1/2}$ with $P(r) = (r/\sigma^2) \exp(-r^2/(2\sigma^2)) U(r)$
 - The magnitude of the signal

Chi-Square Distributed R.V.

❖ Let

$$Y := \sum_{j=1}^n X_j^2$$

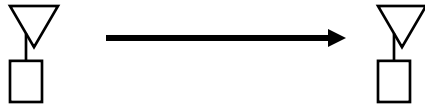
where $\{X_j\}$ are iid Gaussian rvs with zero mean and variance σ^2

❖ It's pdf is

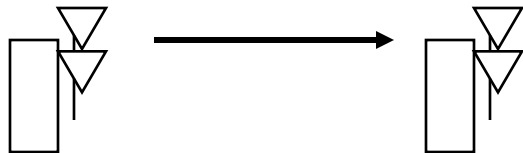
$$p_Y(y) = \frac{1}{\sigma^n 2^n \Gamma(\frac{n}{2})} y^{n/2-1} e^{-y/2\sigma^2} u(y)$$

where $\Gamma(p)$ is the gamma function

The Bottom Line is



- ❖ $\mathbf{r}_k = \mathbf{h}_k \mathbf{x}_k + \mathbf{n}_k$, where \mathbf{h}_k is complex-valued Gaussian (Rayleigh amplitude distribution).



- ❖ $\mathbf{r}_k = \mathbf{h}_k \mathbf{x}_k + \mathbf{n}_k$, where elements of \mathbf{h}_k are i.i.d. complex-valued Gaussian (Rayleigh amplitude distribution)
 \mathbf{h}_k is $[\mathbf{N}_r \times \mathbf{N}_t]$ matrix. This is a MIMO system.

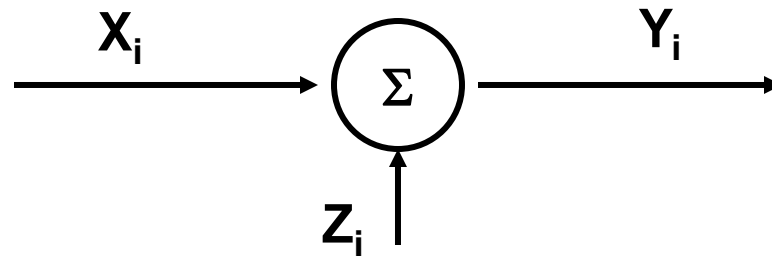
Topics in MIMO

- ❖ MIMO can be used to increase
 - Diversity
 - Capacity
- ❖ MIMO Channel Estimation Errors vs. Capacity
- ❖ MIMO Broadcasting Capacities
- ❖ MIMO Multiple Access Channel Capacities
- ❖ MIMO Beamforming Capacities
- ❖ MIMO Capacities vs. Reduced Complexity Receivers
- ❖ Capacities over Wireless Networks (Ad Hoc Networks)

Chapter 10

❖ The Gaussian Channel

The Additive White Gaussian Noise Channel



- ❖ $Y_i = X_i + Z_i, \quad Z_i \sim \mathcal{N}(0, N)$
- ❖ X_i and Z_i are *independent*.
- ❖ Constraint on the input X_i
 - $E(X_i^2) \leq P$
- ❖ Use of channel
 - Send a codeword (x_1, x_2, \dots, x_n) : $(1/n) \sum_i x_i^2 = P$
 - Successive use of the channel trying to exploit the LLN

Binary Signaling over AWGN

- ❖ $P = E(X^2)$; Let $A = \text{sqrt}(P)$;
- ❖ $X = +A$, prob. $1/2$
 $= -A$, prob. $1/2$
- ❖ $Z \sim \mathcal{N}(0, N)$
- ❖ $Y = X + Z$,
- ❖ Assume Maximum Likelihood Detection
- ❖ $P_e = 0.5 \{ \Pr(Y < 0 \mid X = +A) + \Pr(Y \geq 0 \mid X = -A) \}$
 $= \Pr\{Z > A\}$
 $= 1 - \int_{-\infty}^A (1/\text{sqrt}(2\pi N)) \exp(-0.5z^2/N) dz$
 $= Q(\text{sqrt}(P/N))$

where Q is the Gaussian Q -function

Information Capacity of AWGN

$$\diamond C = \max_{p(x): E(X^2) = P} I(X; Y)$$

$$\diamond I(X; Y) = h(Y) - h(Y|X)$$

$$= h(Y) - h(X+Z|X)$$

$$= h(Y) - h(Z)$$

$$= h(Y) - 0.5 \log(2\pi eN)$$

$$\begin{aligned} \text{--- } E(Y^2) &= E(X^2) + E(Z^2) - X, Z \text{ indep} \\ &= P + N \end{aligned}$$

$$\leq 0.5 \log(2\pi e(P+N)) - 0.5 \log(2\pi eN)$$

$$= 0.5 \log(1+P/N)$$

When do we have the equality?

Shannon Capacity over AWGN

- ❖ The channel capacity theorem gives the following results

$$C = \frac{1}{2} \log_2 \left(1 + \frac{P}{N} \right) \quad [\text{bits/symbol}]$$

- ❖ This results tell you that at a given SNR how many bits can be carried by a single channel-symbol so that almost error-free recovery can be obtained at the receiver
- ❖ Multiplying this number with the maximum symbol rate, [symbols/sec], then you have [bits/sec].

Shannon's Channel Coding Theorem

- ❖ All rates below capacity C are achievable. Specifically, for rate $R < C$, there exists a sequence of $(2^{nR}, n)$ codes with maximum probability of error $\lambda^{(n)} \rightarrow 0$.
- ❖ Conversely, any sequence of $(2^{nR}, n)$ codes with $\lambda^{(n)} \rightarrow 0$ must have $R \leq C$.

Code

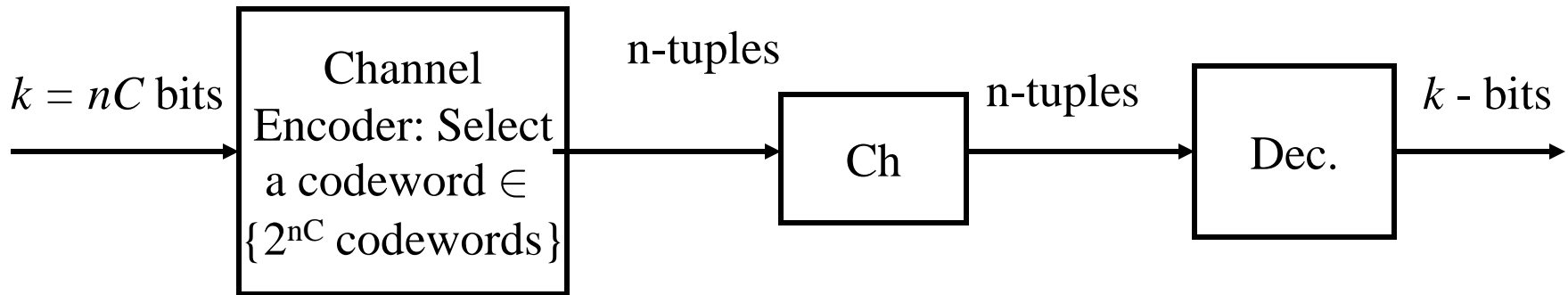
- ❖ A code is a collection of n -tuples, (x_1, x_2, \dots, x_n) , where each x_i is a sample of a real-valued Gaussian with zero mean and variance P .
- ❖ The size of code is the number of codewords, 2^{nC} where C is the rate of the code
 - Via each codeword, we send $k = nC$ bits
- ❖ Note that using n -tuples, using Eq. (1), we note that

$$\mathbf{y} = \mathbf{x}_i + \mathbf{z}, \quad i=1, 2, \dots, 2^{nC}$$

where \mathbf{y} , \mathbf{x}_i and \mathbf{z} are n -tuples:

- The avg. energy of the codeword \mathbf{x}_i is nP
- The avg. energy of the noise \mathbf{z} is nN
- The avg. energy of the received signal \mathbf{y} is $n(P+N)$

The Channel Coding Theorem

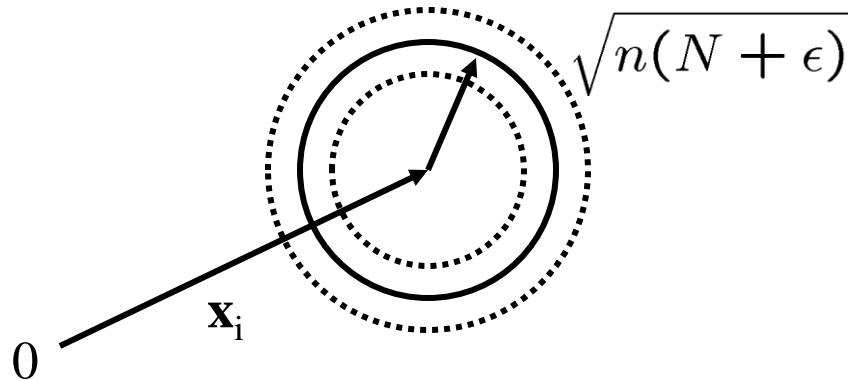


- ❖ Encoding: Select a codeword in a code (There are 2^{nC} codewords}, and send it over the AWGN channel,
- ❖ Decision: decide the codeword whose E. distance with the received vector is the smallest.
- ❖ Ex) Select a codeword $(-1 -1) \in \{(-1 -1), (-1 1), (1 -1), (1 1)\}$, and send it over the channel. Suppose receiving $y=(-.9, 0.1)$. Then, optimum decision is $(-1 1)$.

Code (2)

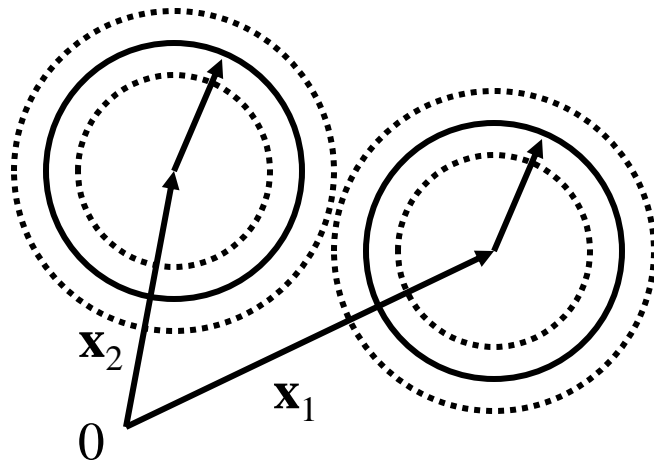
- ❖ We are interested in showing that it's possible to find a code with rate C with which the prob. of making decision errors at the receiver approaches zero as n grows to ∞
- ❖ ML decision rule: Find the i -th codeword whose Euclidean distance with the rec., $\| \mathbf{y} - \mathbf{x}_i \|$ is the minimum, and Decode for i .

Surface Hardening as $n \rightarrow \infty$



- ❖ Consider a received word $\mathbf{y} = \mathbf{x}_i + \mathbf{z}$. The mean of \mathbf{y} is \mathbf{x}_i and variance is nN
- ❖ With high probability, the noise is constrained in a sphere of radius $\sqrt{n(N + \epsilon)}$ (Surface hardening)
- ❖ With a very small probability, the magnitude of the noise vector is larger than the ball of radius, \sqrt{nN} (or smaller).

Two or More Noise Spheres



- ❖ Think about a chance of making decision error--being confused between \mathbf{x}_1 with \mathbf{x}_2 --when the two codewords \mathbf{x}_1 and \mathbf{x}_2 are chosen as shown in the figure

Capacity obtained from Sphere Packing

- ❖ Note that as long as we have the right number of codewords in a code such that the noise-sphere does not overlap with each other, $P(e)$ can be kept very small.
- ❖ Then, the question is “how many noise-spheres can be packed into the volume of the received signal r ”.
- ❖ The volume of n -dimensional sphere of radius q is $A_n q^n$.
- ❖ The radius of the received signal sphere is $(n(P+N))^{1/2}$.
- ❖ The radius of the noise sphere is $(nN)^{1/2}$.
- ❖ The capacity is the maximum number of noise spheres that can be packed into the sphere of the received signal.

Capacity from Sphere Packing Argument

❖ # of noise sphere in the rec-sphere

$$= \frac{A_n(n(P + N))^{n/2}}{A_n(nN)^{n/2}}$$

$$= \left(\frac{P}{N} + 1\right)^{n/2}$$

= # of codewords

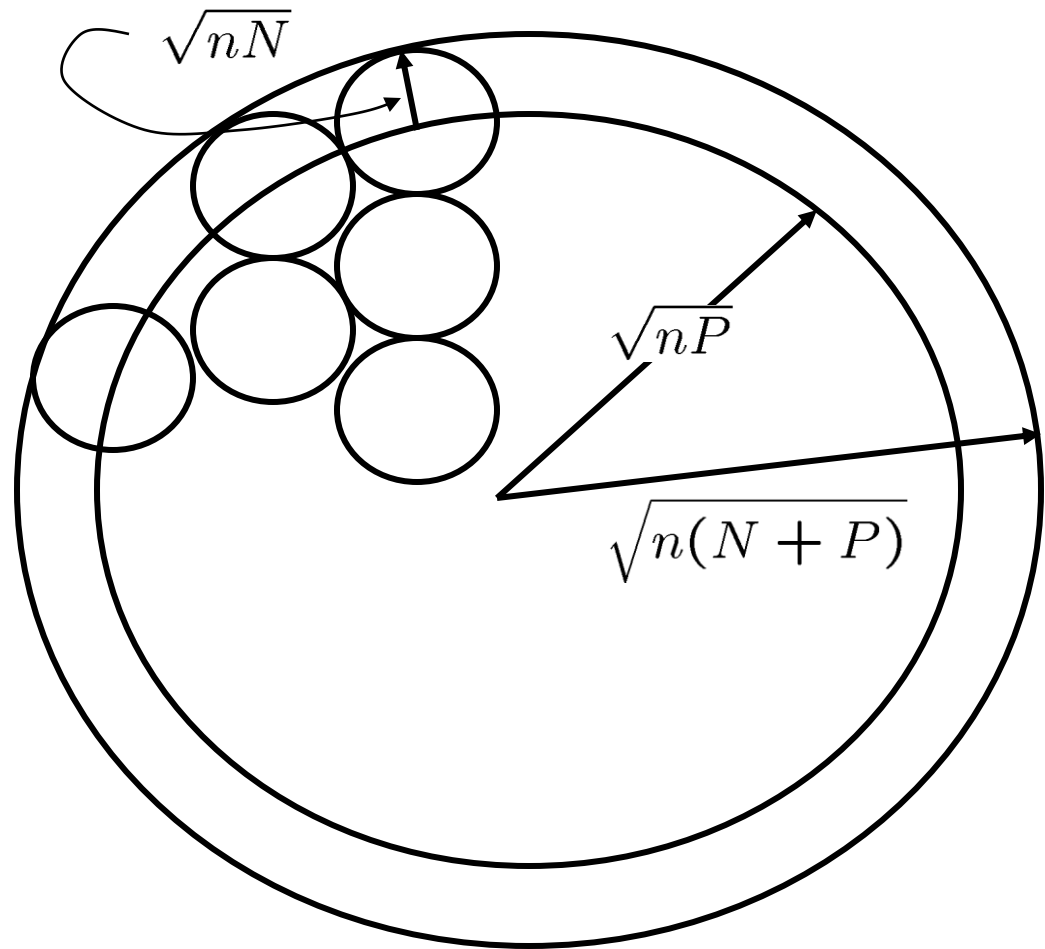
❖ Thus, the capacity is the $\log_2(\# \text{ of codewords})$ which gives $(n/2) \log_2(1+\text{SNR})$.

❖ This is the number of bits we can transmit almost error-free using n -channel symbols.

❖ Dividing it by n , we have the Shannon theoretic capacity per symbol use.

Capacity from Sphere Packing

❖ n-dimensional sphere



Proof of Achievability

- ❖ Sketch of the proof
- ❖ Show that if there is a random code with rate $R < C$, then P_e must go to 0 as $n \rightarrow \infty$
- ❖ Construct the code randomly (Power constraint)
 - Select the n -tuple codewords randomly $\sim \mathcal{N}(0, P)$
 - $(1/n) \sum X_i^2 \Rightarrow P$ (WLLN)
- ❖ Send the first codeword and detect without loss of generality
- ❖ Decode with joint typicality
 - $E_p = \{ |1/n \sum_{i=1}^n X_i^2(1) - P| > \varepsilon \}$: Power constraint violation
 - $E_i = \{ (X^n(i), Y^n) \text{ is in } A_\varepsilon \}, i=2, 3, \dots, M=2^{nR}$, some codeword other than the first one is typical with the received signal

Proof of Achievability (2)

$$\begin{aligned} \blacklozenge \Pr\{\mathbf{E}\} &= \Pr\{\mathbf{E}|\mathbf{W}=1\} = \Pr(E_p \cup E_1^c \cup E_2 \cup \dots \cup E_M) \\ &\leq \Pr(E_p) + \Pr(E_1^c) + \sum_{i=2}^M \Pr(E_i) \\ &\leq \varepsilon + \varepsilon + \sum_{i=2}^M 2^{-n(I(X;Y) - 3\varepsilon)} \\ &\leq 2\varepsilon + 2^{nR} 2^{-n(I(X;Y) - 3\varepsilon)} \\ &= 2\varepsilon + 2^{-n(I(X;Y) - R - 3\varepsilon)} \\ &\leq 3\varepsilon \end{aligned}$$

for n sufficiently large and $R < I(X; Y) - 3\varepsilon$

This proves the existence of a good $(2^{nR}, n)$ code

Converse to the Coding Theorem for AWGN

- ❖ Show if $P_e^{(n)} \rightarrow 0$ for a seq. of $(2^{nR}, n)$ codes for a Gaussian channel with power constraint P ,
then $R < C = 0.5 \log(1+P/N)$

- ❖ Fano's $H(X^n|Y^n) \leq H(W|Y^n) \leq 1 + n R P_e^{(n)} = 1 + n\varepsilon_n$
 - $P_e = P(W \neq g(Y^n))$
 - $\varepsilon_n \rightarrow 0$ as $n \rightarrow \infty$

- ❖ $W \sim$ uniform on $\{1, 2, \dots, M\}$

- ❖ Consider any $(M = 2^{nR}, n)$ code that satisfies the power constraint, i.e.,

$$(1/n) \sum_{i=1}^n x_i^2(w) \leq P, \quad w=1, 2, \dots, M$$

Converse (2)

$$\begin{aligned} \blacklozenge \quad nR &= H(W) = I(W; Y^n) + H(W|Y^n) \\ &\leq I(W; Y^n) + 1 + n\varepsilon_n \\ &\leq I(X^n; Y^n) + 1 + n\varepsilon_n \\ &= h(Y^n) - h(Y^n|X^n) + n\varepsilon_n + 1 \\ &= h(Y^n) - h(Z^n) + n\varepsilon_n + 1 \\ &= h(Y^n) - \sum_i h(Z_i) + n\varepsilon_n + 1 \\ &\leq \sum_{i=1}^n h(Y_i) - \sum_i h(Z_i) + n\varepsilon_n + 1 \end{aligned}$$

Converse (3)

$$nR \leq \sum_{i=1}^n h(Y_i) - \sum_i h(Z_i) + n\varepsilon_n + 1$$

$$\text{--- } Y_i = X_i + Z_i$$

$$R \leq (1/n) \sum_i (1/2) \log(1+P_i/N) + \varepsilon_n + 1/n$$

$$\text{--- Let } P_i = (1/M) \sum_{w=1, \dots, M} x_i(w)^2,$$

the avg. power of i-th symbol of length-n codeword

--- $\log(1+x)$ is concave; apply Jensen's Ineq.

$$\leq (1/2) \log(1 + (1/n) \sum_i P_i/N)$$

$$\text{--- } (1/n) \sum_i P_i \leq P$$

$$\leq (1/2) \log(1+P/N)$$

Q.E.D.

HW #7

❖ P8.1, 8.2, 8.9, 9.2, 9.3, 9.5, 9.12, 9.20

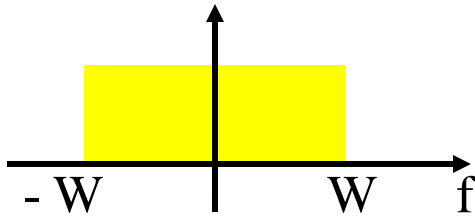
Appendix

- ❖ The rest of the charts are taken from EE 1473 notes, an undergraduate communications course at U Pitt.
- ❖ Take a look at it for the capacity of band-limited channel.

Channel Capacity of Bandlimited Channel

❖ Nyquist's Sampling Theorem and Shannon's Capacity Theorem

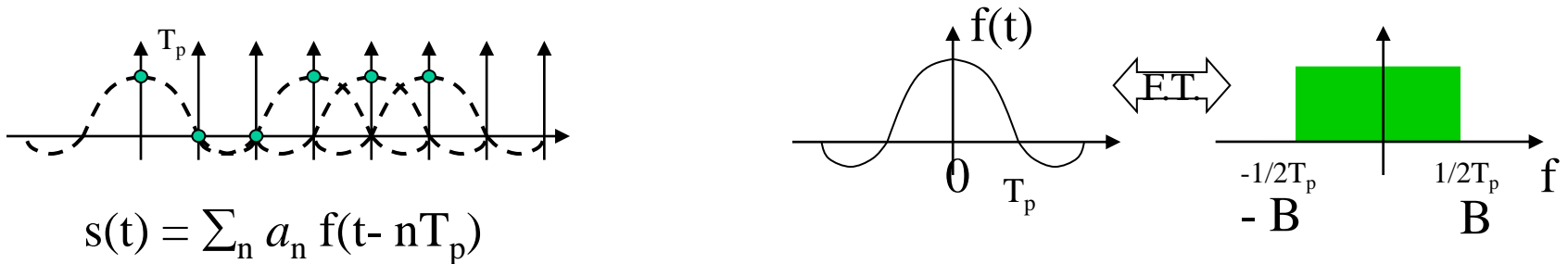
❖ Consider a channel whose bandwidth is limited by W Hz



❖ Nyquist Sampling Theorem: Any function of time bandlimited by B Hz, can be represented with samples of the signal, if $f_s \geq 1/2B$

❖ Now, let's take a look at how fast can we send our channel symbols over the channel, without noise

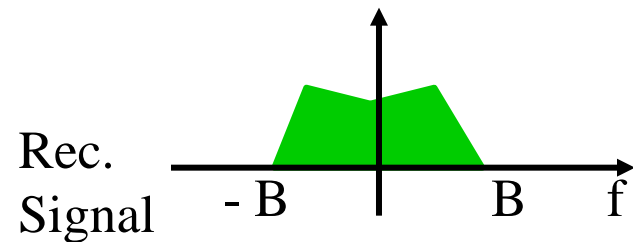
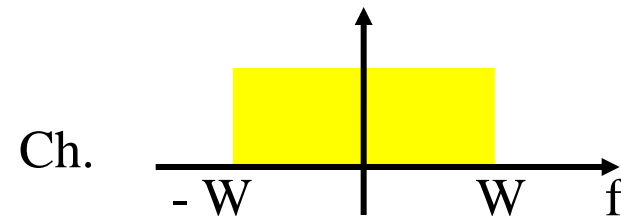
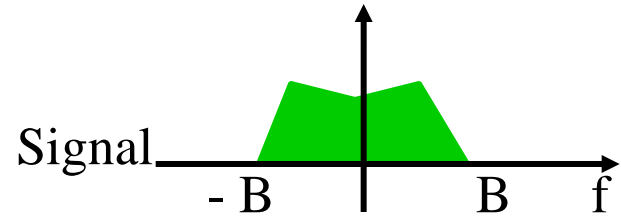
The Maximum Baud = Nyquist Symbol Rate



- ❖ The signal $s(t)$ is modulation of channel-symbol sequence $\{a_n\}$ by the shifted *sinc*'s $f(t - nT_p)$
- ❖ When the Baud (reduce T_p) is increased, the maximum frequency of the digital signal is increased.
- ❖ Baud (The symbol rate) = $2 B$ symbols/sec where $B = 1/(2T_p)$
- ❖ Now, let's consider what happens when $B < W$ and $B > W$

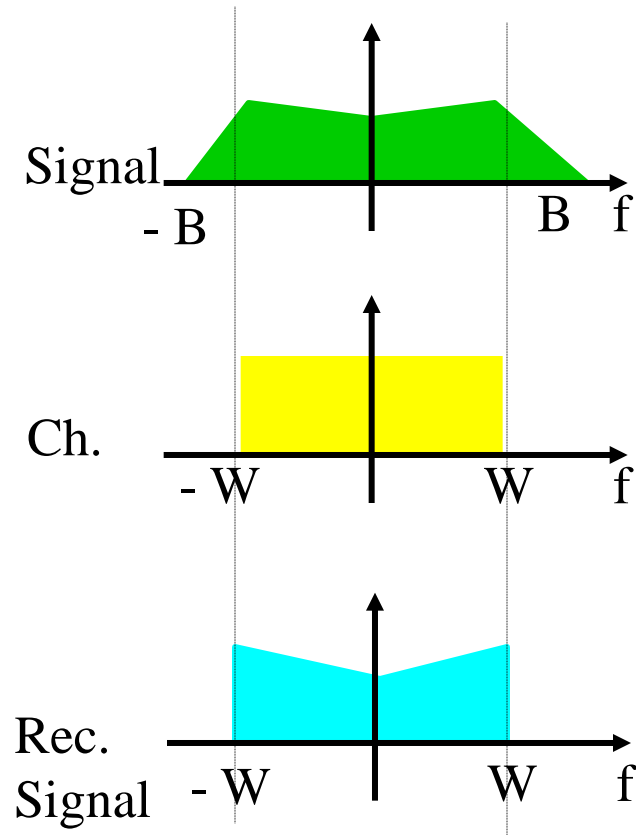
Nyquist Rate

- ❖ What happens when $B < W$?
 - The received signal is the same as the signal sent
 - Thus, applying the matched filter to the received signal and taking samples at $2B$ will get me the channel symbols back
- ❖ What happens when $B > W$?



Nyquist Rate (2)

- ❖ When $B > W$, the received signal is not what you sent at TX?
 - Some of high frequency content of $s(t)$ will get cut off
 - Error free transmission is impossible due to this distortion (Intersymbol Interference)
- ❖ Thus, the maximum possible channel-symbol rate ($1/T_p$) for channel bandlimited by W Hz is $2W$ symbols/sec [Nyquist Rate]



Nyquist Rate (3)

- ❖ Channel with W Hz (Flat, ideal spectrum) and no noise
- ❖ **Maximum Baud without distortion is $2W$ symbols/sec**
- ❖ When $B \leq W$, we can perfectly demodulate the symbols sent, by convolving the receive-signal with the matched filter $f(t)$ and then taking samples at every T_p second.
- ❖ That is, the samples at the receiver is exactly the same as the channel symbols transmitted

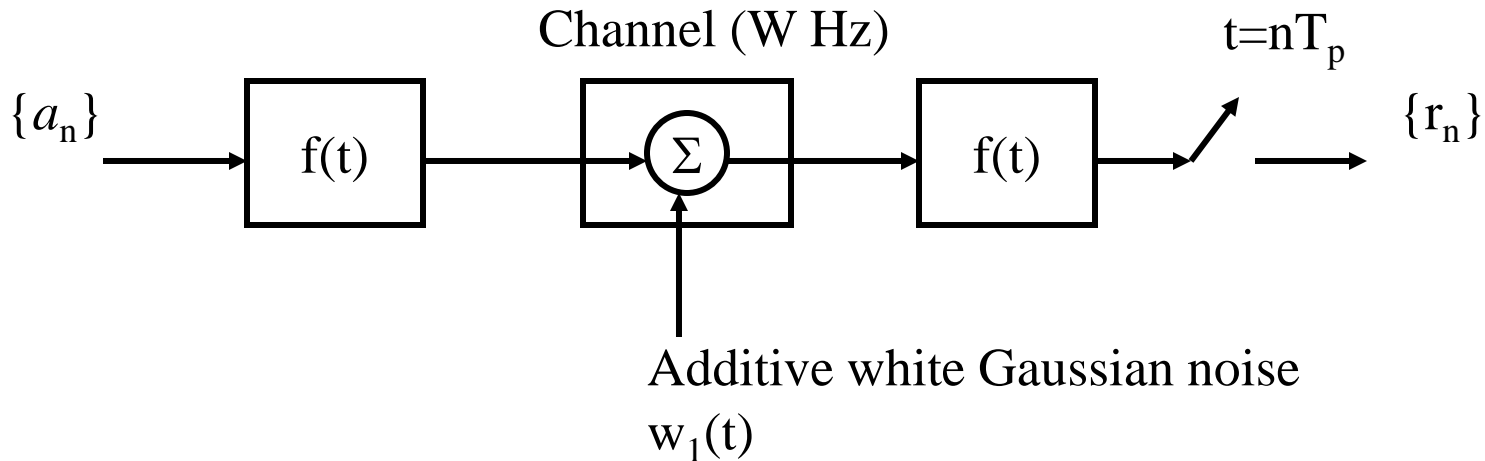
$$r_n = a_n$$

- ❖ But, when $B > W$, r_n contains the distortion and never be equal to a_n

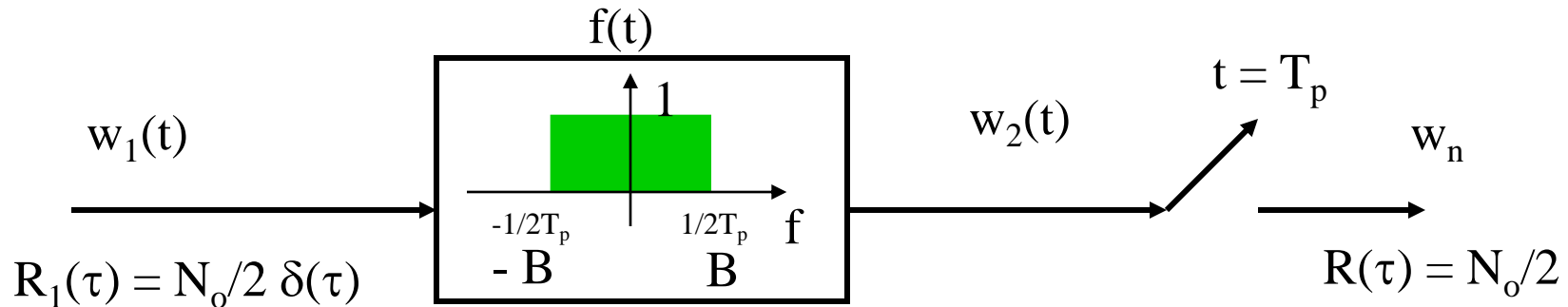
$$r_n = a_n + \text{distortion}$$

Channel Capacity

- ❖ For situations where $B = W$, let's consider now that the channel also adds Gaussian noise to the receive-signal



Bandlimited AWGN



- ❖ What is the characteristic of white Gaussian noise which passes thru a filter bandlimited by B Hz, and sampled at the sampling frequency of $2B$ seconds
- ❖ What is the autocorrelation function of $w_2(t)$?
 - $R_2(\tau) = (N_o B) \text{sinc}(\pi \tau / T_p)$
 - Mean of $w_2(t)$?
- ❖ Finally, what are the mean and the variance of w_n ?

Channel Capacity

- ❖ Now, everything comes down to finding out the channel capacity of the following equation

$$r_n = a_n + w_n$$

where

- $\{a_n\}$ is a random variable representing the channel symbol.
- $\{w_n\}$ is the Gaussian random variable with zero mean and variance WN_0 , representing the bandlimited and sampled thermal noise (choose $B = W$).
- $\{y_n\}$ is the random variable representing the received signal.
- $SNR = (E\{a_n^2\} \int |f(t)|^2 dt) / E\{w_n^2\} = P/N$.

Shannon Capacity

- ❖ The channel capacity theorem gives the following results

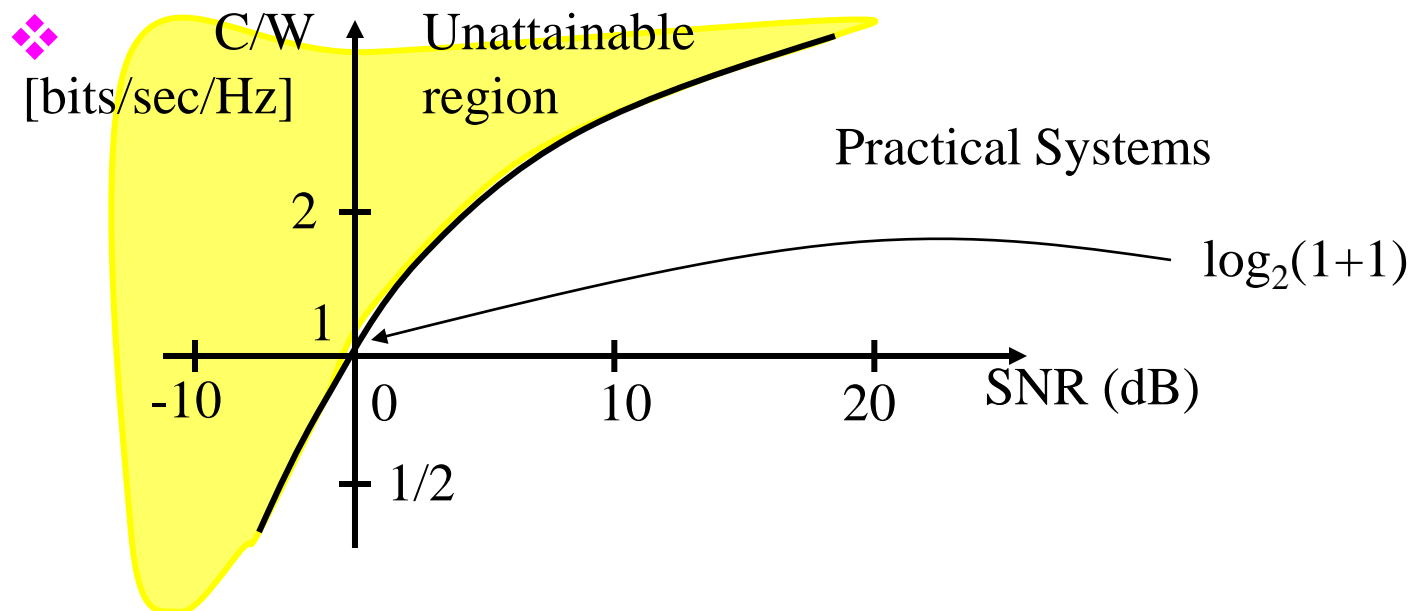
$$C = \frac{1}{2} \log_2 \left(1 + \frac{P}{N} \right) \quad [\text{bits/symbol}]$$

- ❖ This results tell you that at a given SNR how many bits can be carried by a single channel-symbol a_n and almost error-free recovery of a_n can be obtained at the receiver.
- ❖ Multiplying this number with the maximum Baud of $2W$ symbols/sec, we have the channel capacity in terms of bits/sec.

Shannon Capacity (2)

- ❖ The channel capacity in [bits/sec] is

$$C = W \log_2\left(1 + \frac{P}{N}\right) = W \log_2\left(1 + \frac{P}{N_o W}\right)$$



Information Theory

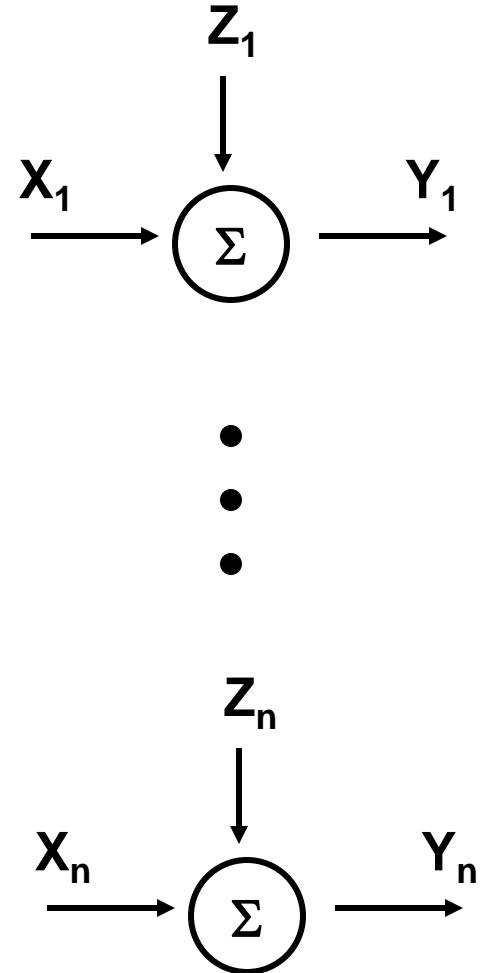
Parallel and MIMO Channel Capacities

Agenda

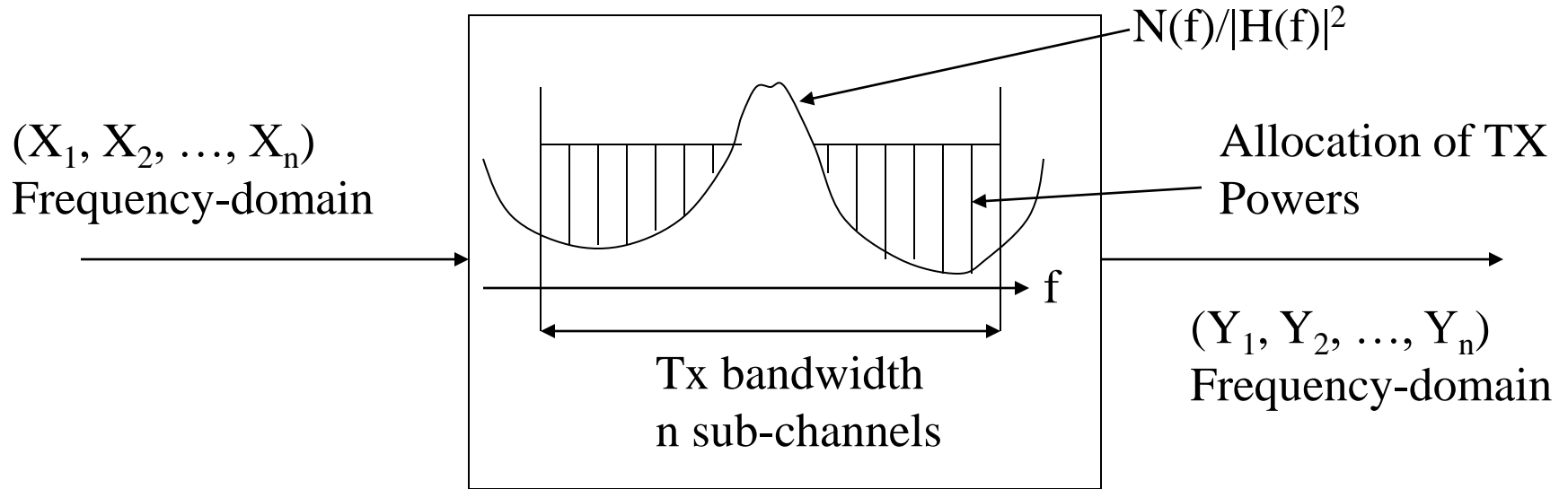
- ❖ Parallel Gaussian Channels
- ❖ Channels with colored Gaussian noise
- ❖ MIMO Capacity

Parallel Gaussian Channels

- ❖ $Y_i = X_i + Z_i, i = 1, 2, \dots, n,$
with $Z_i \sim \text{indep. } \mathcal{N}(0, N_i)$
- ❖ Power constraint: $E\{\sum_{j=1}^n X_j^2\} \leq P.$
- ❖ $C = \max I(X_1, \dots, X_n; Y_1, \dots, Y_n)$
subject to $f(x_1, \dots, x_n): E\{\sum_{j=1}^n X_j^2\} \leq P$
 - Again, this capacity is the supremum of all achievable rates.



Example) Capacity over Frequency-Selective Channels



- ❖ Maximize the channel capacity subject to the input distribution
- ❖ Consider dividing up the channel with small chunks of bandwidth

Parallel Gaussian Channels

- ❖ $Y_k = H_k X_k + W_k$, where W_k is i.i.d. Gaussian with fixed variance.
- ❖ Or equivalently consider $Y_k = X_k + Z_k$, where noise Z_k is independent Gaussian with a different variance.

Water Filling Capacity over PGCs (2)

- ❖ The length of signal should be at least greater or equal to $1/\Delta f$ in order to meet the condition of independent channels
 - cf) In Orthogonal Frequency Division Multiplexing, the transmitter uses a fixed power for all frequency bins (Suboptimal)
- ❖ Set of independent channels
 - $Y_k = X_k + Z_k, k=1, 2, \dots, n$, where $Z_k \sim \mathcal{N}(0, N_k)$
 - Find the optimum distribution $f(X_1, X_2, \dots, X_n)$ that maximizes the channel capacity
 - When Gaussian is proven optimal, it is equivalent to find the optimum power allocation $\{P_k\}$

Water Filling Solution over PGCs (3)

- ❖ Find the maximum mutual information

$$C = \max I(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n ; \mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n)$$

$$\text{subject to } f(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n): E\{\sum_k \mathbf{X}_k^2\} \leq P$$

$$\begin{aligned} \text{❖ } I(\mathbf{X}_1, \dots, \mathbf{X}_n; \mathbf{Y}_1, \dots, \mathbf{Y}_n) &= h(\mathbf{Y}_1, \dots, \mathbf{Y}_n) - \sum_k h(\mathbf{Z}_k) \\ &\leq \sum_k h(\mathbf{Y}_k) - h(\mathbf{Z}_k) && \text{--- (1)} \\ &\leq (1/2) \sum_k \log(1 + P_k/N_k) && \text{--- (2)} \end{aligned}$$

where $P_k := E(\mathbf{X}_k^2)$ and thus $P = \sum P_k$ and $P_k \geq 0$

- (1) Independent assumption (higher entropy)
- (2) Entropy of Gaussian is maximum
- The equality achieved when $\mathbf{Y}_k = \mathbf{X}_k + \mathbf{N}_k$ is indeed a Gaussian
 $(\mathbf{X}_1, \dots, \mathbf{X}_K) \sim \mathcal{N}(0, \text{COV}(\mathbf{X}) = \text{diag}(P_1, P_2, \dots, P_n))$

Water Filling Solution over PGCs (4)

- ❖ Now find the optimum power allocation vector

$$\mathbf{P} = (P_1, \dots, P_n)$$

which maximizes the R.H.S. of inequality (2).

- ❖ Consider the Lagrange multiplier

$$J(P_1, \dots, P_n) = (1/2) \sum_k \log(1 + P_k/N_k) + \lambda(P - \sum_k P_k)$$

with constraints $P_k \geq 0$ for $\forall k$, and $\sum_k P_k = P$.

- ❖ Partial differentiation wrt P_k

$$(1/2)/(P_k + N_k) - \lambda = 0 \text{ for all } k$$

- ❖ Thus, $P_k + N_k = 1/2\lambda = L$ or

$$P_k = L - N_k \text{ for all } k$$

- ❖ In fact, since power ≥ 0 , we choose $P_k = [L - N_k]^+$.

Water Filling Solution over PGCs (5)

- ❖ The solution $P_k = [L - N_k]^+$ where L is chosen so that

$$\sum_k [L - N_k]^+ = P$$

- This is called the water-filling solution.

- ❖ The condition $P_k \geq 0$ makes the problem difficult.

- ❖ We use the Kuhn Tucker conditions to verify the solution is indeed the assignment that maximizes the capacity.

Kuhn-Tucker Conditions:

Constraint Optimization with inequality constraints

❖ $\max f(\mathbf{x})$

- subject to $g_1(\mathbf{x}) \leq c_1, \dots, g_m(\mathbf{x}) \leq c_m$
- $x_i \geq 0$ for $i=1, \dots, n$.

❖ First form the Lagrangian J

$$J = f(\mathbf{x}) + \lambda_1 (c_1 - g_1(\mathbf{x})) + \dots + \lambda_m (c_m - g_m(\mathbf{x}))$$

❖ For a point \mathbf{x} to be a maximum, the following must be satisfied (necessary conditions for a point to be max)

- $\partial J / \partial x_i \leq 0$ & $x_i \geq 0$ & $x_i(\partial J / \partial x_i) = 0$, for all $i = 1, \dots, n$
- $g_j(\mathbf{x}) \leq c_j$ & $\lambda_j \geq 0$ & $\lambda_j(c_j - g_j(\mathbf{x})) = 0$ for all $j = 1, \dots, m$

❖ Sufficiency conditions include

- $f(\mathbf{x})$ is differentiable and concave in the region $x_i \geq 0$
- Each $g_j(\mathbf{x})$ is differentiable and convex in the region $x_i \geq 0$

Kuhn-Tucker Example

❖ Problem: find $\max f(x, y) = 2x + 3y$,
subject to $g(x, y) = x^2 + y^2 \leq 2$ and $x, y \geq 0$.

❖ Solution:

$$J = 4x + 3y + \lambda(10 - x^2 - y^2)$$

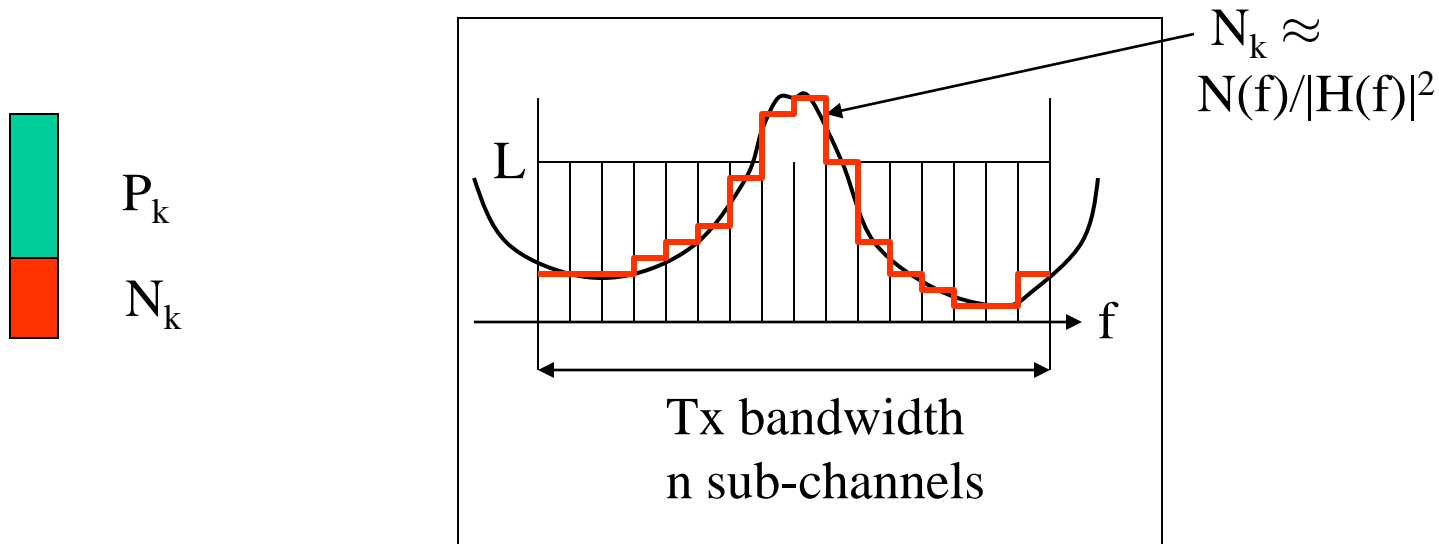
- Note that f and g satisfy the suff. cond's
- Check the necessary conditions for a point to be a maximum
- (set 1) $2 - 2\lambda x \leq 0$, $x \geq 0$, $x(2 - 2\lambda x) = 0$
- (set 2) $3 - 2\lambda y \leq 0$, $y \geq 0$, $y(3 - 2\lambda y) = 0$
- (set 3) $x^2 + y^2 \leq 2$, $\lambda \geq 0$, $\lambda(2 - x^2 - y^2) = 0$
- (1)(iii) implies that either $x = 0$ or $\lambda x = 1$. But (1)(i) cannot be satisfied with $x=0$. Thus, $\lambda x = 1$.
- Similarly $\lambda y = 3/2$ from set 2; $y > 0$; $\lambda > 0$.
- Substituting $x = 1/\lambda$ and $y = 3/2\lambda$ to set 3 (iii), $2 - 1/\lambda^2 - 9/4\lambda^2 = 0$
- $8\lambda^2 - 13 = 0$, thus $\lambda = \sqrt{13/8}$; $x = \sqrt{8/13}$; $y = (3/2) \sqrt{8/13} = \sqrt{18/13}$; $\max f(x, y) = 5.099$, and $x^2 + y^2 = 2$.

Kuhn-Tucker on Capacity

- ❖ $J(P_1, \dots, P_K) = (1/2) \sum_k \log(1+P_k/N_k) + \lambda (P - \sum_k P_k)$
with constraints $P_k \geq 0$ for any k , and $\sum_k P_k \leq P$.
- ❖ Sufficiency cond.'s are o.k. and $\log(1+x)$ is concave/diff.
- ❖ Necessary cond.
 - (Set 1) for all k
 - $J_k = (1/2)/(P_k + N_k) - \lambda \leq 0, \quad P_k \geq 0, \quad P_k [(1/2)/(P_k + N_k) - \lambda] = 0$
 - (Set 2)
 - $\sum_k P_k \leq P, \quad \lambda \geq 0, \quad \lambda(P - \sum_k P_k) = 0$
- ❖ Solution:
 - Using 1, $\lambda = (1/2)/(P_k + N_k)$ for all k . $P_k = [1/2\lambda - N_k]^+$
 - Using 2, we have $\sum_k [1/2\lambda - N_k]^+ = P$,
 - Let $L = 1/2\lambda$ be **the water-level** satisfying $\sum_k [L - N_k]^+ = P$

Reference: Convex Optimization by Boyd/Vandenberghe

Water Filling Solution over PGCs (6)



- ❖ Note over the region where $N_k > L$, no power is allocated; no transmission of information takes place over such a region.
- ❖ Thus, $C = (1/2) \sum_k \log(1 + P_k/N_k)$
 $= (1/2) \sum_k (\log(L/N_k))^+$

Channels with Colored Gaussian Noise

- ❖ Capacity of $(Y_1, \dots, Y_n) = (X_1, \dots, X_n) + (Z_1, \dots, Z_n)$
 - with a power constraint $(1/n) \sum_i E(X_i^2) \leq P$ (or $\text{tr}(\mathbf{K}_x) \leq nP$)
 - with correlated noise, i.e., $Z \sim \mathcal{N}(0, \mathbf{K}_z)$ where \mathbf{K}_z is not diagonal
- ❖ $I(X_1, \dots, X_n; Y_1, \dots, Y_n)$
 - $= h(Y_1, \dots, Y_n) - h(Z_1, \dots, Z_n)$
 - $\leq 0.5 \log((2\pi e)^n |\mathbf{K}_x + \mathbf{K}_z|) - h(Z_1, \dots, Z_n) \quad \text{---- (1)}$
- ❖ Find the max $|\mathbf{K}_x + \mathbf{K}_z|$ subject to the trace constraint on \mathbf{K}_x

Channels with Colored Gaussian Noise

- ❖ \mathbf{K}_z is non-negative definite; thus, \mathbf{K}_z can be decomposed into

$$\mathbf{K}_z = \mathbf{Q} \Lambda \mathbf{Q}^t$$

where \mathbf{Q} is unitary such that $\mathbf{Q}\mathbf{Q}^t = \mathbf{I}$ ($|\mathbf{Q}| = 1$)

- ❖ $|\mathbf{K}_x + \mathbf{K}_z| = |\mathbf{K}_x + \mathbf{Q} \Lambda \mathbf{Q}^t|$
 $= |\mathbf{Q} (\mathbf{Q}^t \mathbf{K}_x \mathbf{Q}) \mathbf{Q}^t + \mathbf{Q} \Lambda \mathbf{Q}^t|$
 $= |\mathbf{Q} (\mathbf{Q}^t \mathbf{K}_x \mathbf{Q} + \Lambda) \mathbf{Q}^t|$
 $\quad \text{--- } \det(\mathbf{BC}) = \det(\mathbf{B}) \det(\mathbf{C})$
 $= |\mathbf{Q}^t \mathbf{K}_x \mathbf{Q} + \Lambda|$
 $= |\mathbf{A} + \Lambda|$

Channels with Colored Gaussian Noise

$$\diamond |\mathbf{K}_x + \mathbf{K}_z| = | \mathbf{Q}^t \mathbf{K}_x \mathbf{Q} + \Lambda | = | \mathbf{A} + \Lambda |$$

$$\diamond \text{tr}(\mathbf{A}) = \text{tr}(\mathbf{Q}^t \mathbf{K}_x \mathbf{Q})$$

$$= \text{tr}(\mathbf{Q}^t \mathbf{Q} \mathbf{K}_x) \quad \text{--- } \text{tr}(\mathbf{BC}) = \text{tr}(\mathbf{CB})$$

$$= \text{tr}(\mathbf{K}_x)$$

\diamond Thus, the problem is now to maximize $|\mathbf{A} + \Lambda|$ with the constraint that $\text{tr}(\mathbf{A}) = \sum_i^n A_{ii} \leq nP$

\diamond Now we note that

$$|\mathbf{A} + \Lambda| \leq \prod_i (A_{ii} + \lambda_i) \quad \text{--- chain rule on Gaussian } \mathbf{X}$$

with equality iff \mathbf{A} is diagonal.

\diamond Now, going back to (1)

Channels with Colored Gaussian Noise

- $$\begin{aligned}
 \diamond I(X_1, \dots, X_n; Y_1, \dots, Y_n) & \\
 & \leq 0.5 \log((2\pi e)^n |K_x + K_z|) - h(Z_1, \dots, Z_n) \\
 & = 0.5 \log((2\pi e)^n |A + \Lambda|) \\
 & \quad - 0.5 \log((2\pi e)^n |K_z|) \\
 & \leq 0.5 \sum_{i=1}^n \log((2\pi e) (A_{ii} + \lambda_i)) \\
 & \quad - 0.5 \log((2\pi e)^n |Q \Lambda Q^t|) \\
 & \leq 0.5 \sum_{i=1}^n \log(1 + A_{ii}/\lambda_i)
 \end{aligned}$$
- $$\diamond \text{The water-filling solution again } (A_{ii} \geq 0; \sum_i A_{ii} \leq nP)$$

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

where L is chosen such that $\sum_{i=1}^n (L - \lambda_i)^+ = nP$
- $$\diamond C = 0.5 \sum_{i=1}^n [\log(L/\lambda_i)]^+$$

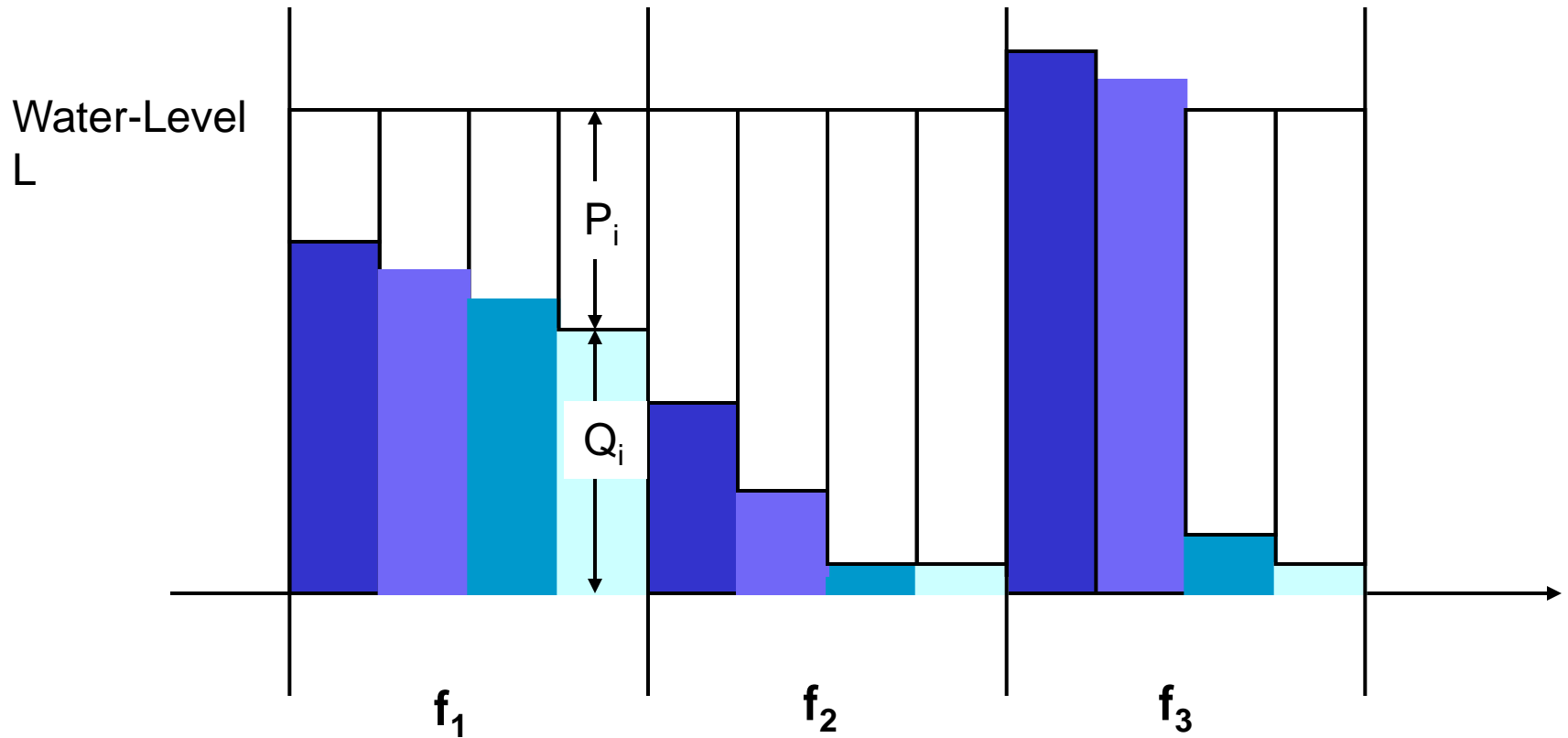
Channels with Colored Gaussian Noise

- ❖ The following two problems are equivalent
- ❖ $Y = X + Z$, where $Z \sim N(0, K_Z = Q\Lambda Q^t)$; find the best K_X with constraint that $\text{tr}(K_X) \leq nP$
 \Leftrightarrow
- ❖ $Q^t Y = Q^t X + Q^t Z \equiv Y' = X' + Z'$, where $Z' \sim N(0, K_Z')$
 - $K_Z' = E(Q^t Z Z^t Q) = Q^t E(ZZ^t) Q = \Lambda$
 - Find the best $\text{tr}(E(X'X'^t)) = \text{tr}(K_X') = \text{tr}(K_X) \leq nP$
 - $K_X' = Q^t K_X Q$ (=A in previous pages)
 - Note that the channel is simply the parallel channels to which we can apply the water-filling solution.
 - Thus, the optimal K_X' = a diagonal matrix with water-filling powers.
- ❖ Once we have found K_X' , which is diagonal, we can find K_X from $K_X = QK_X'Q^t$ which is not diagonal in general.

Water-filling Solutions in Different Applications

- ❖ Water-filling over time-varying fading channels
- ❖ Water-filling over frequency-selective channels
- ❖ Water-filling over space-selective channels
- ❖ Water-filling over selective eigenchannels
- ❖ Water-filling over combination-selective channels

Water-Filling Over Selective Space-Frequency Gains ($1/\text{SNR}_i$)

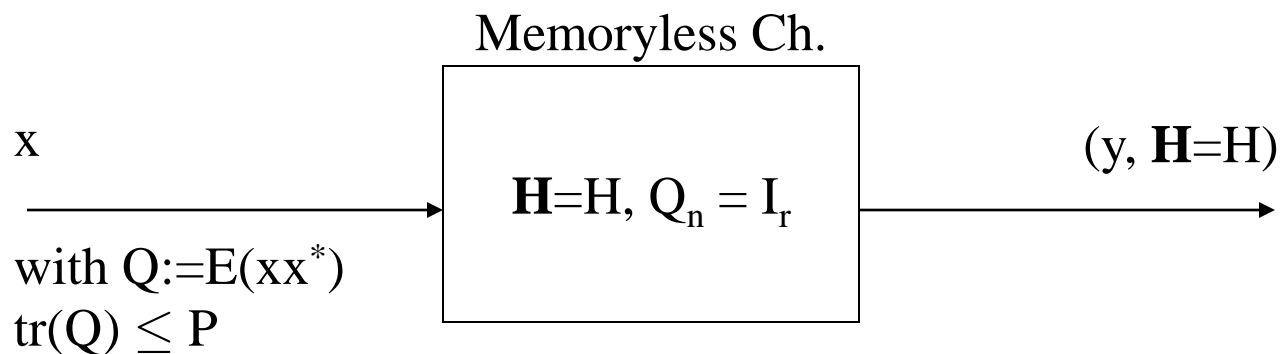


MIMO Capacity

Agenda

- ❖ Two papers dealing with the same MIMO problem
 - Telatar
 - This note
 - Ergodic capacity
 - Foschini
 - The Outage Capacity and practical algorithms
 - Upper and Lower Bounds on the Ergodic Capacity

The Channel



- ❖ A memory-less channel which draws an independent n_k and \mathbf{H}_k for every new choice of x_k

$$y_k = \mathbf{H}_k x_k + n_k.$$

For each k , \mathbf{H}_k , y_k , and x_k are independent:

$$I(\mathbf{x}; (y, \mathbf{H})) = I(\mathbf{x}; \mathbf{H}) + I(\mathbf{x}; y | \mathbf{H})$$

Mutual Information

$$\begin{aligned} \blacklozenge I(\mathbf{x}; (y, \mathbf{H})) &= I(\mathbf{x}; \mathbf{H}) + I(\mathbf{x}; y | \mathbf{H}) \\ &= 0 + I(\mathbf{x}; y | \mathbf{H}) \\ &= h(\mathbf{x} | \mathbf{H}) - h(\mathbf{x} | y, \mathbf{H}) \\ &= \sum p(H) [h(\mathbf{x} | \mathbf{H} = H) - h(\mathbf{x} | y, \mathbf{H} = H)] \\ &= E_{\mathbf{H}} I(\mathbf{x}; y | \mathbf{H}) \end{aligned}$$

System Equation

❖ We have $y = \mathbf{H} x + n$,

where n is $[r \times 1]$ circularly symmetric complex Gaussian vector with zero mean and covariance $E(nn^*) = \mathbf{I}_r$, and x and n are mutually independent

- y, n are $[r \times 1]$
- x is $[t \times 1]$
- \mathbf{H} is $[r \times t]$ matrix

❖ For a fixed $\mathbf{H} = H$, find the channel capacity subject to

$$E(x^* x) = \text{tr}(E(xx^*)) < P$$

❖ Find the ensemble average of the capacity for \mathbf{H}

Entropy of Gaussian $\mathbf{x} \in \mathbb{C}^n$ is maximum
under power constraint

❖ For $\mathbf{x} \in \mathbb{C}^n$ with zero mean and covariance $\mathbf{Q} := E(\mathbf{x}\mathbf{x}^*)$,

$$H(\mathbf{x}) = E(-\log(p(\mathbf{x}))) \leq \log \det(\pi e \mathbf{Q})$$

with equality attained iff \mathbf{x} is CSC Gaussian with $E(\mathbf{x}\mathbf{x}^*) = \mathbf{Q}$

❖ First, show that for CSC Gaussian \mathbf{x} , the entropy is

$$E(-\log(p(\mathbf{x}))) = \log \det(\pi e \mathbf{Q})$$

❖ Then, show the inequality (see Telatar's Lemma-2)

– We have proved this already in Chapter 9 (Theorem 9.6.5.)

Show $E(-\log(p(x))) = \log \det(\pi e Q)$ for CSCG x

❖ First note that $p(x) = \det(\pi Q)^{-1} \exp(-x^* Q^{-1} x)$ where x is $[t \times 1]$ vector (thus Q is $[t \times t]$ matrix)

❖ $E(-\log(p(x))) = \log \det(\pi Q) - E(-x^* Q^{-1} x)$

$$\text{-- } E(-x^* Q^{-1} x) = - E\{\text{tr}(x x^* Q^{-1})\}$$

-- try to see it with 2x2 example

$$= \log \det(\pi Q) + \text{tr}(E(xx^*)Q^{-1})$$

$$= \log \det(\pi Q) + (\log e) \text{tr}(I)$$

$$= \log \det(\pi e Q)$$

❖ cf) For $x \in \mathbb{R}^n$, $\sim \mathcal{N}(0, K)$,

$$E(-\log(p(x))) = 0.5 \log[(2\pi e)^n |K|] = 0.5 \log |2\pi e K|$$

$$I(\mathbf{x}; \mathbf{y} | \mathbf{H}) =: I(\mathbf{x}; \mathbf{y}) = H(\mathbf{y}) - H(\mathbf{y} | \mathbf{x})$$

❖ $I(\mathbf{x}; \mathbf{y}) = H(\mathbf{y}) - H(\mathbf{y} | \mathbf{x})$ (definition of the mutual information)

$$= H(\mathbf{y}) - H(\mathbf{x} + \mathbf{n} | \mathbf{x})$$

$$= H(\mathbf{y}) - H(\mathbf{n} | \mathbf{x})$$

$$= H(\mathbf{y}) - H(\mathbf{n}) \quad (\text{independence of } \mathbf{n} \text{ and } \mathbf{x})$$

-- We know $H(\mathbf{n}) = \log \det(\pi \mathbf{e} \mathbf{Q}_n)$ with $\mathbf{Q}_n = \mathbf{I}_r$

$$\leq \log \det(\pi \mathbf{e} \mathbf{Q}_y) - \log \det(\pi \mathbf{e} \mathbf{Q}_n) \quad \text{--- (0)}$$

-- “=” achieved iff \mathbf{y} is CSCG

$$= \log \det(\pi \mathbf{e} [\mathbf{H}\mathbf{Q}\mathbf{H}^* + \mathbf{I}_r]) + \log \det(\mathbf{I}_r / (\pi \mathbf{e}))$$

$$= \log \det(\mathbf{H}\mathbf{Q}\mathbf{H}^* + \mathbf{I}_r) \quad \text{--- (1)}$$

❖ $\max I(\mathbf{x}; \mathbf{y})$ is achieved when \mathbf{x} is CSCG—thus, \mathbf{y} is CSCG

$$\det(\mathbf{I} + \mathbf{AB}) = \det(\mathbf{I} + \mathbf{BA})$$

$$\blacklozenge \mathbf{I}(\mathbf{x}; \mathbf{y}) = \log \det(\mathbf{I}_r + \mathbf{HQH}^*) = \log \det(\mathbf{I}_t + \mathbf{QH}^*\mathbf{H})$$

Finding Q that maximizes $\log \det(\mathbf{I}_r + \mathbf{H}\mathbf{Q}\mathbf{H}^*)$

- ❖ Again, this should be the water-filling solution
- ❖ Note that the covariance matrix Q is non-negative definite
- ❖ $\mathbf{H}^*\mathbf{H} = \mathbf{U}^*\mathbf{\Lambda}\mathbf{U}$ ($\because \mathbf{H}^*\mathbf{H}$ is non-negative definite)
 - \mathbf{U} is unitary matrix
 - $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_t)$
- ❖ $(\mathbf{I}_t + \mathbf{Q}\mathbf{H}^*\mathbf{H}) = (\mathbf{I}_r + \mathbf{Q}\mathbf{U}\mathbf{S}\mathbf{V}^* \mathbf{V}\mathbf{S}\mathbf{U}^*)$

$$= (\mathbf{I}_r + \mathbf{Q}\mathbf{U}\mathbf{\Lambda}^{1/2} \mathbf{\Lambda}^{1/2} \mathbf{U}^*) \quad \text{--- note } \mathbf{S}=\mathbf{\Lambda}^{1/2}$$
- ❖ $\det(\mathbf{I}_t + \mathbf{Q}\mathbf{H}^*\mathbf{H}) = \det(\mathbf{I}_r + \mathbf{\Lambda}^{1/2} \mathbf{U}^* \mathbf{Q} \mathbf{U}\mathbf{\Lambda}^{1/2})$

$$\text{---- Let } \mathbf{Q}'=\mathbf{U}^*\mathbf{Q}\mathbf{U}$$

$$= \det(\mathbf{I}_r + \mathbf{\Lambda}^{1/2} \mathbf{Q}' \mathbf{\Lambda}^{1/2})$$

Finding Q that maximizes $\log \det(I_r + HQH^*)$

- ❖ $\det(I_r + \Lambda^{1/2} Q' \Lambda^{1/2}) \leq \prod_{k=1}^r (1 + Q'_{kk} \lambda_k)$
- Equality attained when Q' is diagonal (mutually independent $\{x_k\}$).
 - Note while Q' is diagonal, Q may not be diagonal.
 - Ex) $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$: when $b=c$, $\det(A) = ad - bb$
- Q1) Why did I assume (Hermitian) symmetry?

Continuing the maximization from (1)

$$\blacklozenge I(\mathbf{x}; \mathbf{y}) \leq \log \det(\mathbf{H}\mathbf{Q}\mathbf{H}^* + \mathbf{I}_r) \quad \text{--- (1)}$$

$$\leq \log \prod_{k=1}^r (1 + \mathbf{Q}'_{kk} \lambda_k)$$

--- Equality attainable when \mathbf{Q}' is diagonal

$$= \sum_{k=1}^r \log(1 + \mathbf{Q}'_{kk} \lambda_k) \quad \text{--- (2)}$$

**Note λ_k are
channel gains**

--- Now, find the best $\{\mathbf{Q}'_{kk}\}$ subject to
 $\text{tr}\{\mathbf{Q}'\} = \text{tr}\{\mathbf{Q}\} \leq P$ (Water-filling solution)

--- $\mathbf{Q}'_{kk} = (\mu - 1/\lambda_k)^+$, where μ is a const.
chosen to satisfy $\sum_{k=1}^r \mathbf{Q}'_{kk} = P$.

$$= \sum_{k=1}^r \log(1 + (\mu - 1/\lambda_k)^+ \lambda_k)$$

$$= \sum_{k=1}^r [\log(\mu \lambda_k)]^+ \quad \text{--- (3)}$$

--- This is the channel capacity for H

The Channel Capacity for $\mathbf{H} = \mathbf{H}$

❖ Finally, we have $C = \sum_{k=1}^r \log(\mu \lambda_k)^+$

❖ Steps of the derivation

- Step (0) is obtained with the assumption of the channel that the noise is CSC Gaussian with zero mean and covariance \mathbf{Q}_n
- Step (1) $\log \det(\mathbf{H}\mathbf{Q}\mathbf{H}^* + \mathbf{I}_r)$ is obtained by choosing CSC Gaussian \mathbf{x}
- **Step (2), $\sum_{k=1}^r \log(1 + \mathbf{Q}'_{kk} \lambda_k) = \sum_{k=1}^r \log(1 + (\mathbf{P}/t) \lambda_k)$, is achievable by choosing CSCG \mathbf{x} with diagonal $\mathbf{Q} = \mathbf{E}(\mathbf{x}\mathbf{x}^*) = (\mathbf{P}/t) \mathbf{I}_t$ which implies that $\mathbf{x}_1, \dots, \mathbf{x}_t$ are mutually independent and have uniform power.**
- Step (3) is achievable with the water-filling power distribution which is possible to obtain only when \mathbf{H} is available also at the transmitter.

The Channel Capacity for $\mathbf{H} = \mathbf{H}$

- ❖ It may not be reasonable—for fast time-varying channels — that the channel matrix \mathbf{H} is also known at the transmitter (It would perhaps be reasonable for indoors)
- ❖ Telatar and Foschini papers investigate the situation where \mathbf{H} is known only at the receiver (the transmitter does not know the channel \mathbf{H})
- ❖ Without knowing the channel, the best power allocation Q_{kk} should be P/t for all $k=1, \dots, t$
 - $\mathbf{Q} = \text{diag}(P/t, P/t, \dots, P/t)$
 - $\mathbf{Q}' = \mathbf{U}\mathbf{Q}\mathbf{U}^* = \text{diag}(P/t, P/t, \dots, P/t)$
- ❖ Let's call this capacity $C(\mathbf{H})$

$$C(\mathbf{H}) = \sum_{k=1}^r \log(1 + (P/t) \lambda_k)$$

Rayleigh Fading Channel Matrix $H \in \mathbb{C}^{r \times t}$

- ❖ Let's denote H_{ij} for the i -th row and j -th column of H .
- ❖ It's modeled as a Rayleigh fading tap.
- ❖ $|H_{ij}|$ is the Rayleigh r.v.
- ❖ The power of each tap, $|H_{ij}|^2$, is the Chi-square r.v. with two degree of freedom and unit power.
- ❖ $\{H_{ij}\}$ are iid Complex Gaussian with zero mean and unit variance. The real and imaginary parts are independent real Gaussians with zero mean and 0.5 variance.
- ❖ Note under unitary transformation, $\text{Cov}(UH)$ stays the same as that of H , and so thus that of UHV^* (Lemma-5).

The Channel Capacity [Thm 1]

❖ Thus, the capacity is

$$C = E_{\mathbf{H}}\{\log \det(\mathbf{I}_r + \mathbf{Q}\mathbf{H}\mathbf{H}^*)\}$$

--- We already know $\mathbf{Q}=(P/t) \mathbf{I}_t$ due to \mathbf{H} unknown at Tx

$$= E_{\mathbf{H}}\{\log \det(\mathbf{I}_r + (P/t) \mathbf{H}\mathbf{H}^*)\} \quad \text{---- (4)}$$

$$= E_{\mathbf{H}}\{\log \det(\mathbf{I}_t + (P/t) \mathbf{H}^*\mathbf{H})\} \quad \text{---- (5)}$$

$$= E_{\mathbf{H}}\{\sum_{k=1}^r \log(1 + (P/t) \lambda_k)\} \quad \text{---- (6)}$$

--- note $\{\mathbf{H}=\mathbf{H}\} \equiv \{\lambda_k = \lambda_k, k=1, \dots, \min\{t,r\}\}$

❖ Note that the distribution of the eigenvalues of \mathbf{H} is the key information needed for evaluation of the capacity

❖ Note that this capacity is achievable with the choice of

$$\mathbf{Q}_x = E(\mathbf{x}\mathbf{x}^*) = (P/t)\mathbf{I}_t$$

Capacity ($t = 1$ and r is arbitrary)

- ❖ Using (5), we have $C = E_{\mathbf{H}}\{\log \det(\mathbf{I}_t + (P/t) \mathbf{H}^* \mathbf{H})\}$
 - Note $\mathbf{H}=(H_{11}; H_{21}; \dots; H_{r1})$ where each H_{k1} is complex Gaussian with indep. real. and imag. parts with zero mean and variance $\frac{1}{2}$
 - Thus, power of each tap $|H_{k1}|^2$ is Chi-square with two degrees of freedom with mean 1
- ❖ Thus $\mathbf{H}^* \mathbf{H} = \sum_{k=1}^r |H_{k1}|^2 = \sum_{k=1}^r \chi^2_2 = \chi^2_{2r}$
 - χ^2_{2r} is Chi-square with $2r$ degree of freedom with mean r
- ❖ $C(\mathbf{H}) = \log \det(\mathbf{I}_t + (P/t) \mathbf{H}^* \mathbf{H})$
 $= \log(1 + (P/t) \chi^2_{2r})$
- ❖ $C = \int_0^\infty p(x) \log(1 + (P/t) x) dx$
where $p(x) = (1/\Gamma(r)) x^{r-1} e^{-x} u(x)$, is the pdf of χ^2_{2r}

Capacity ($r = 1$ and t is arbitrary)

❖ Similarly we have

$$C(H) = \log \det (\mathbf{I}_r + \mathbf{P} \mathbf{H} \mathbf{H}^*) = \log(1 + \mathbf{P} \chi^2_{2t})$$

❖ The pdf of the Chi-square r.v. with $2t$ degree of freedom with mean t is given by

$$p(x) = (1/\Gamma(t)) x^{t-1} e^{-x} u(x)$$

❖ Thus, the averaged capacity C is given by

$$C = (1/\Gamma(t)) \int_0^\infty \log(1+\mathbf{P} x) x^{t-1} e^{-x} dx$$

where $\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx = (t-1)!$ for integer t .

Limit Capacity (fixed r and $t \rightarrow \infty$)

❖ Consider the inner product of i -th and j -th rows of H

$$(1/t) H(i, :) H(j, :)^* = (1/t) \sum_{n=1}^t H_{in} H_{jn}^* \\ \rightarrow \delta(i=j) \quad \text{as } t \rightarrow \infty$$

--- Note this is due to the law of large numbers

--- H_{in} and H_{jn} are mutually independent

❖ Thus, using (4) the capacity in the limit of large t is

$$C \rightarrow r \log(1+P) \quad \text{as } t \rightarrow \infty$$

Limit Capacity (fixed t and $r \rightarrow \infty$)

❖ Using (5),

$$(1/r) \mathbf{H}^* \mathbf{H}_{(ij)} = (1/r) \sum_{n=1}^r \mathbf{H}_{in} \mathbf{H}_{jn}^* \\ \rightarrow \delta(i=j) \text{ as } r \rightarrow \infty$$

❖ $\therefore \log \det(\mathbf{I}_t + (P/t) r (1/r) \mathbf{H}^* \mathbf{H})$

$$\rightarrow \log \det(\mathbf{I}_t + (P/t) r \mathbf{I}_t) \text{ as } r \rightarrow \infty \\ = \log \prod_{n=1}^t (1+rP/t) \\ = t \log(1+rP/t)$$

$$C = t \log(1+rP/t) \quad \text{as } r \rightarrow \infty$$

Rigorous Proof of Thm-1

❖ From (1), we have $C(\mathbf{Q}, \mathbf{H}) := \log \det(\mathbf{I}_r + \mathbf{H}\mathbf{Q}\mathbf{H}^*)$ which is achievable with CSCG \mathbf{x} with covariance \mathbf{Q}

❖ $C(\mathbf{Q}) = E_{\mathbf{H}}\{C(\mathbf{Q}, \mathbf{H})\}$

❖ We want to maximize $C(\mathbf{Q})$ over all possible choice of \mathbf{Q}

❖ \mathbf{Q} is a covariance matrix. Thus, we could have

$$\mathbf{Q} = \mathbf{U}\mathbf{D}\mathbf{U}^*$$

where \mathbf{U} is unitary and \mathbf{D} is diagonal

❖ $\therefore C(\mathbf{Q}) = E_{\mathbf{H}}\{\log \det(\mathbf{I}_r + \mathbf{H}\mathbf{U}\mathbf{D}\mathbf{U}^*\mathbf{H}^*)\}$

---- distribution of $\mathbf{H}\mathbf{U}$ is the same as that of \mathbf{H}

❖ $\therefore C(\mathbf{Q}) = C(\mathbf{D})$ which means that we can restrict our attention to *diagonal covariance matrices* for optimization

Rigorous Proof of Thm-1 (2)

- ❖ Thus, the problem now is to find the optimal distribution of the diagonal terms
- ❖ Let Q_0 be the average of all the possible permutation of a best non-negative diagonal Q with non-uniform diagonal terms, such that

$$\begin{aligned}
 Q_0 &= (1/t!) \sum_{k=1}^{t!} (\Pi_k Q \Pi_k^*) \\
 &= \text{Const. } I_t && \text{(since } Q \text{ is diagonal)} \\
 &= (P/t) I_t && \text{(since } \text{tr}\{Q_0\} = \text{tr}\{Q\} = P)
 \end{aligned}$$

- ❖ We note that the function $C(Q)$ is concave \cap of Q
 - $Q \mapsto (I_r + HQH^*)$, this mapping is linear
 - $Q \mapsto C(Q, H)$, this mapping is concave
 - $Q \mapsto C(Q)$, this mapping is concave

Rigorous Proof of Thm-1 (3)

- ❖ Applying the Jensen's inequality ($f(E(Q)) \geq E\{f(Q)\}$) to the concave mapping $C(Q)$, we have

$$C(Q_0) \geq (1/t!) \sum_{k=1}^{t!} C(\Pi_k Q \Pi_k^*) = C(Q)$$

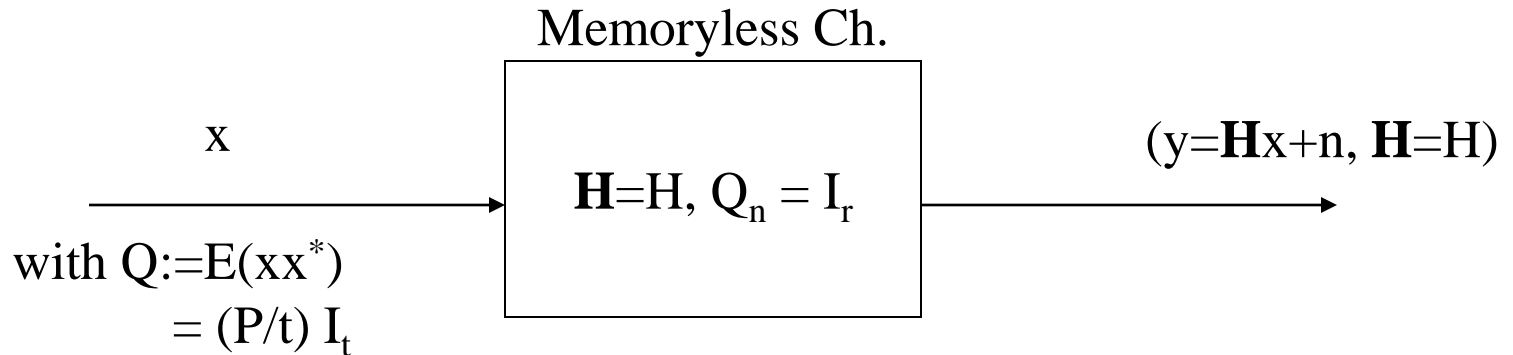
--- Since $C(\Pi_k Q \Pi_k^*) = C(Q)$

--- This tells us that **uniform power distr. is optimal**

- ❖ Thus, the capacity is achieved when x is CSCG with covariance $Q = (P/t) I_t$. The capacity is given by

$$\boxed{E_{\mathbf{H}}\{\log \det(I_r + (P/t) \mathbf{H}\mathbf{H}^*)\}} \quad \text{---- (7)}$$

Summary of Theorem 1



Theorem 1: The capacity

$E_{\mathbf{H}}\{\log \det(I_r + (P/t) \mathbf{H}\mathbf{H}^*)\} = E_{\mathbf{H}}\{\log \det(I_t + (P/t)\mathbf{H}^*\mathbf{H})\}$
 is achieved with the input \mathbf{x} which is CSC Gaussian with zero mean and covariance $(P/t) I_t$, over memoryless channel with CSC Gaussian noise \mathbf{n} with $Q_n = I_r$ and \mathbf{H} is $[r \times t]$ Gaussian with

Evaluation of the Capacity

- ❖ Using (6), we can carry out the calculation of the capacity once we know the distribution of eigenvalues of H^*H matrix

$$C = E\{\sum_{k=1}^r \log(1 + (P/t) \lambda_k)\}$$

--- Let $m = \min\{r, t\}$

$$= \sum_{k=1}^m E\{\log(1 + (P/t) \lambda_k)\}$$

$$= m E\{\log(1 + (P/t) \lambda_1)\}$$

- ❖ Thus, obtain $p(\lambda_1, \dots, \lambda_m)$, where $\{\lambda_k\}$ are unordered, and then obtain for the marginal $p(\lambda_1)$

- ❖ Thus, we need to evaluate

$$\int \log(1 + (P/t)\lambda_1) p(\lambda_1) d\lambda_1 \quad \text{---- (8)}$$

Theorem 2

- ❖ The evaluation of the capacity based on the integral (8)

$$\begin{aligned}
 C &= m \int_0^{\infty} \log(1 + P\lambda/t) p(\lambda) d\lambda \\
 &= \int_0^{\infty} \log(1 + P\lambda/t) \sum_{k=0}^{m-1} \frac{k!}{(k+n-m)!} [L_k^{n-m}(\lambda)]^2 \lambda^{n-m} e^{-\lambda} d\lambda
 \end{aligned}$$

where $m = \min\{r, t\}$, $n = \max\{r, t\}$ and the Laguerre poly.

$$L_k^l(x) = \sum_{i=0}^k (-1)^i \frac{(l+k)!}{(k-i)! (l+i)! i!} x^i$$

- ❖ The capacity is achievable with CSC Gaussian \mathbf{x} with $\mathbf{Q}_x = (P/t) \mathbf{I}_r$, over memoryless ergodic channel \mathbf{H}

HW#8

- ❖ One problem from Telatar's paper
- ❖ 7 Problems from Cover & Thomas
 - P9.7, 9.8, 9.9, 9.10, 9.11, 9.14, 9.15

Telatar's Paper

- ❖ Consider 2x2 MIMO System, given by

$$\mathbf{Y} = \mathbf{H} \mathbf{X} + \mathbf{Z},$$

where elements of the noise vector \mathbf{Z} are i.i.d. real-valued Gaussian with zero mean and variance 1. \mathbf{X} and \mathbf{Y} is real-valued signals. The power budget on transmitted signal \mathbf{X} is P , i.e., $\text{trace}(\mathbf{E}\{\mathbf{X}\mathbf{X}^H\}) \leq P$.

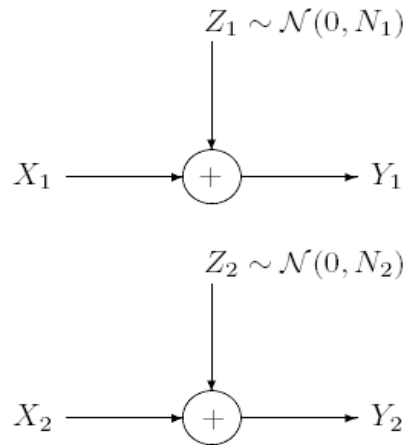
1. For a fixed $\mathbf{H} = \mathbf{H}$, find the general expression of the channel capacity with the assumption that the channel is known both at the transmitter and the receiver.
2. For $\mathbf{H} = \begin{bmatrix} 3/4 & 1/4 \\ 1/4 & 3/4 \end{bmatrix}$, evaluate the capacity for $P = 4$ and for $P = 6$.
3. For $P = 6$, give a description of the signaling scheme which would realize the capacity you obtained. In particular, how would you transmit your signals at antenna-1 and antenna-2?

Cover & Thomas

❖ P9.7, 9.8, 9.9, 9.10, 9.11, 9.14, 9.15

Not Required

Consider the following parallel Gaussian channel



where $Z_1 \sim \mathcal{N}(0, N_1)$ and $Z_2 \sim \mathcal{N}(0, N_2)$ are independent Gaussian random variables and $Y_i = X_i + Z_i$. We wish to allocate power to the two parallel channels. Let β_1 and β_2 be fixed. Consider a total cost constraint $\beta_1 P_1 + \beta_2 P_2 \leq \beta$, where P_i is the power allocated to the i^{th} channel and β_i is the cost per unit power in that channel. Thus $P_1 \geq 0$ and $P_2 \geq 0$ can be chosen subject to the cost constraint β .

- For what value of β does the channel stop acting like a single channel and start acting like a pair of channels?
- Evaluate the capacity and find P_1, P_2 that achieve capacity for $\beta_1 = 1, \beta_2 = 2, N_1 = 3, N_2 = 2$ and $\beta = 10$.

Not Required

❖ *Discrete input continuous output channel.* Let $\Pr\{X = 1\} = p$, $\Pr\{X = 0\} = 1 - p$, and let $Y = X + Z$, where Z is uniform over the interval $[0, a]$, $a > 1$, and Z is independent of X .

(a) Calculate

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

(b) Now calculate $I(X; Y)$ the other way by

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

(c) Calculate the capacity of this channel by maximizing over p

Singular Value Decomposition

❖ Any $[r \times t]$ matrix can be factored into

$$H = USV^*$$

--- the columns of $U \in \mathbb{C}^{r \times r}$ are the eigenvectors of A

--- the columns of $V \in \mathbb{C}^{t \times t}$ are the eigenvectors of B

--- $S \in \mathbb{R}^{r \times t}$ matrix with singular values, $s_1, \dots, s_{\min(t,r)}$, on the diagonal which are the square roots of the eigenvalues of either A or B (the rest are all zeros)

where $A := H H^*$ and $B := H^* H$

--- Unitary matrices U and V have the following property

$$U^*U = I_r, UU^* = I_r, V^*V = I_t, \text{ and } VV^* = I_t$$

Singular Value Decomposition (2)

- ❖ $A := HH^* = USV^*VS^*U^* = U SS^* U^* = U\Lambda_A U^*$
- ❖ $B := H^*H = VS^*U^* USV^*$
 $= V S^*S V^* = V\Lambda_B V^*$
- ❖ $q = \text{rank of } H = \min\{\# \text{ of indept. cols, } \# \text{ of indept. rows}\}$
 $\leq \min\{t, r\}$
- ❖ Full rank system is when $q = \min\{t, r\}$
- ❖ Note that A and B are non-negative definite, such that,
 $\forall \mathbf{x}, \quad \mathbf{x}A\mathbf{x} \geq 0$
- ❖ Singular values and eigenvalues are non-negative

Example

$$\diamond H = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}$$

$\diamond B=HH^* = [2 \ -1; -1 \ 2]$ has eigenvalues 3 and 1

$$\diamond H = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \sqrt{3} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{pmatrix} (1 \ -2 \ 1) \frac{1}{\sqrt{6}} \\ (-1 \ 0 \ 1) \frac{1}{\sqrt{2}} \\ (1 \ 1 \ 1) \frac{1}{\sqrt{3}} \end{pmatrix}$$

\diamond Check if $UU^*=U^*U=I_3$ and $VV^*=V^*V=I_2$
– Use “svd” in MATLAB

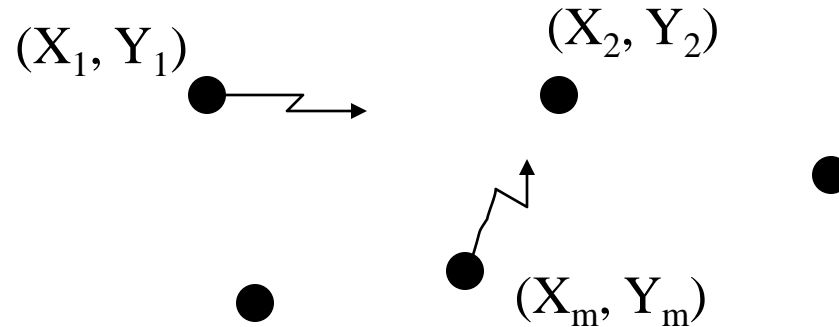
Information Theory

MAC + SW

Agenda

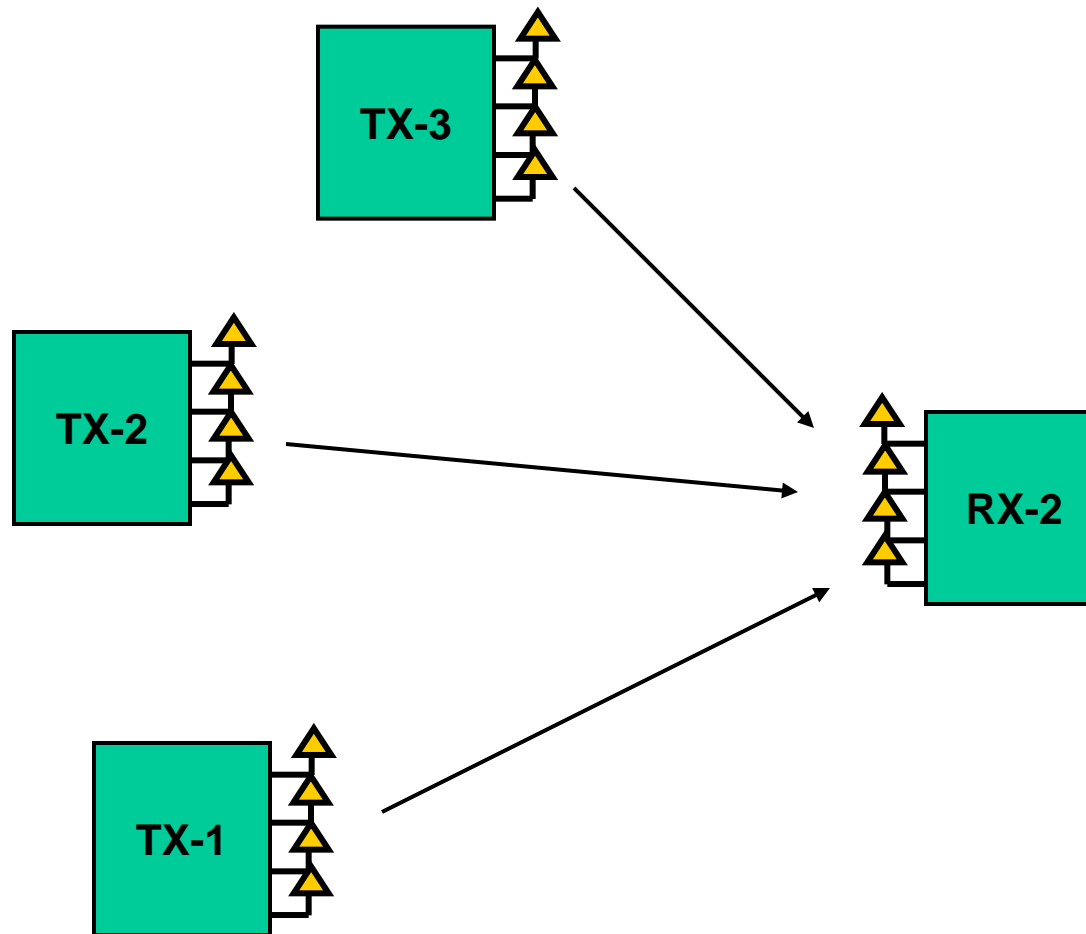
❖ Network Information Theory

The Most General Communication Network

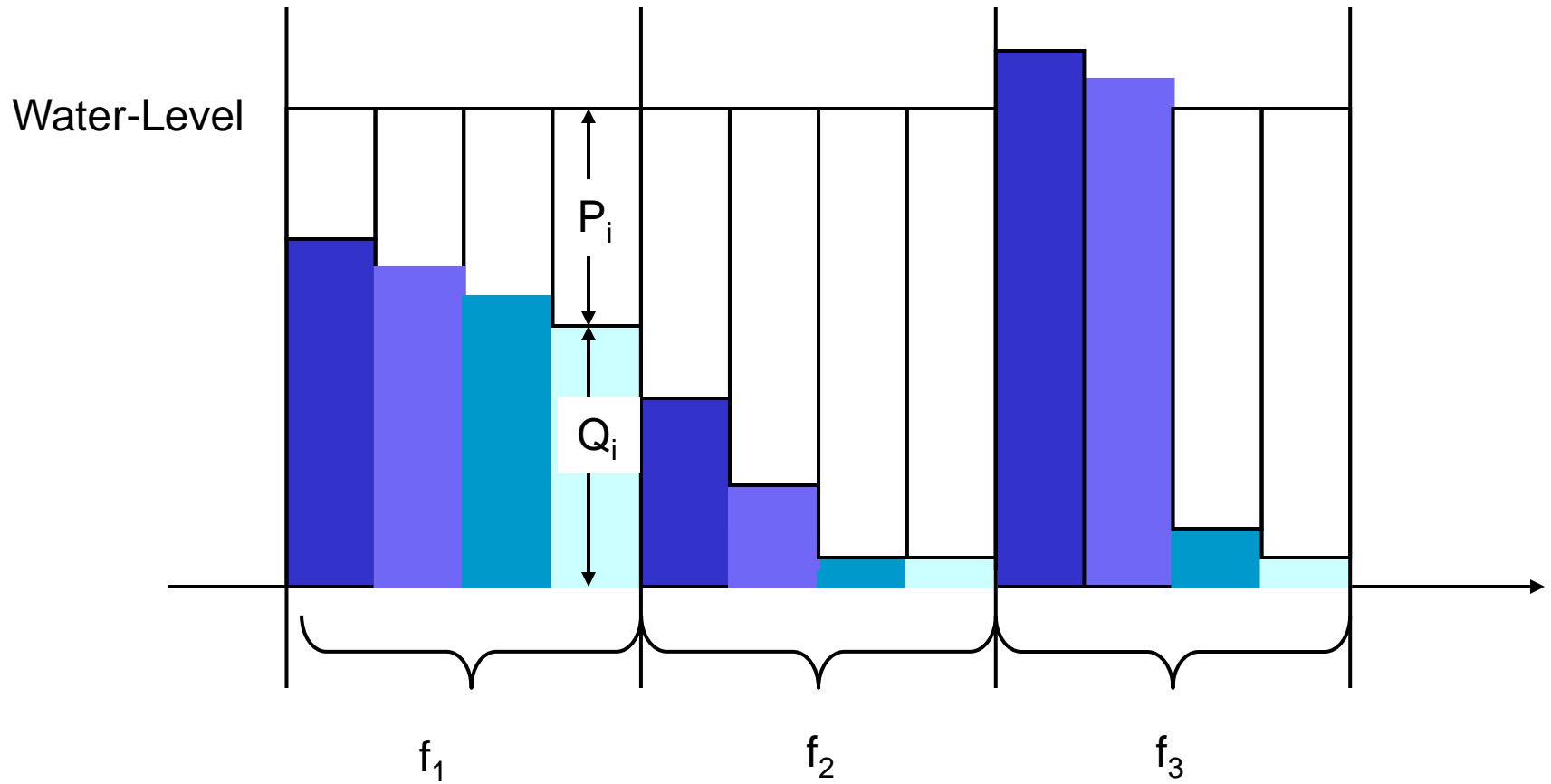


- ❖ M senders and N receivers
- ❖ Wants to communicate
 - Some are sources
 - Some are relays
 - Some are receivers

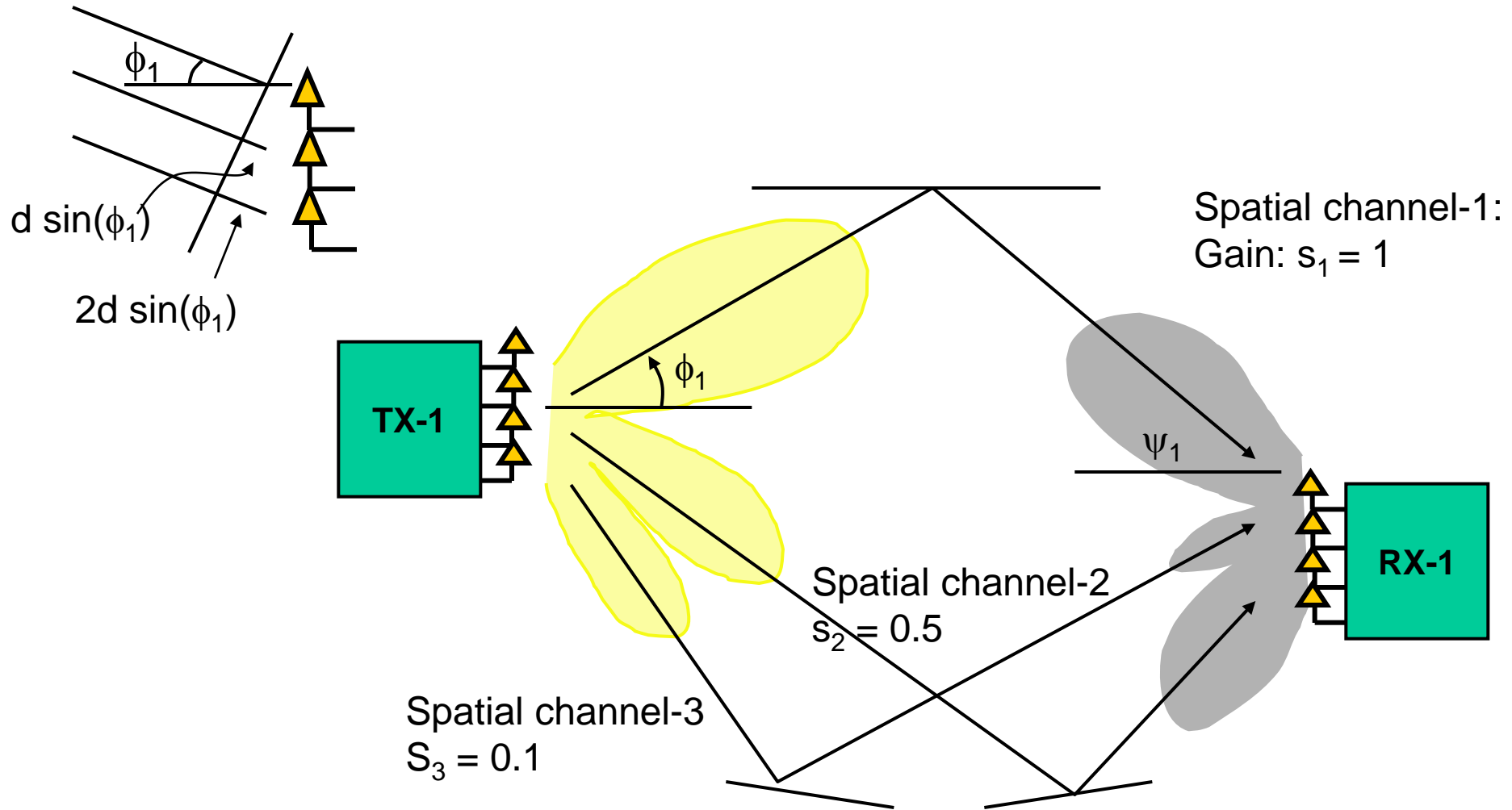
Multiple Access Channels and Multi-User Detection



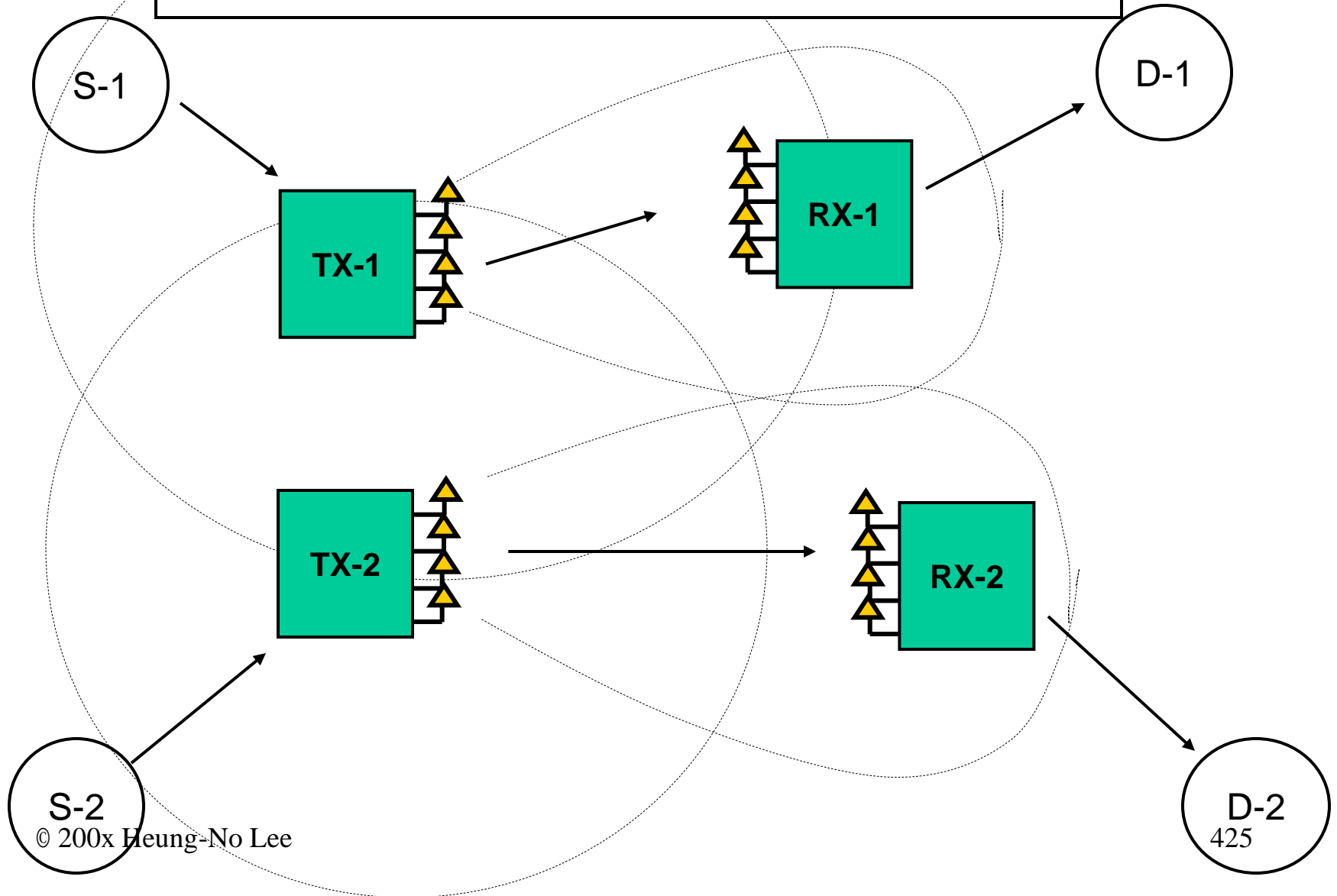
Water-Filling Over Selective Space-Frequency Gains ($1/\text{SNR}$)



SVD Beam-Forming over Spatial Channels



Space-Time-Frequency Agile OFDM Transceivers



S-2

© 200x Heung-No Lee

D-2

425

A system with many senders/many receivers

❖ Network channels

- Interference, noises

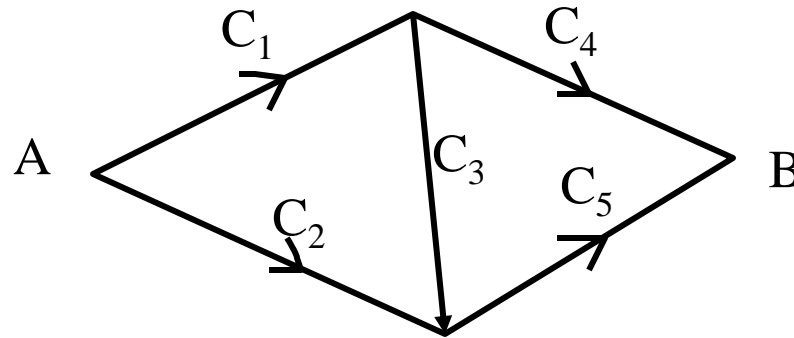
❖ Examples: Satellite network, broadcasting networks, cellular phone networks, sensor networks

- Many to one (Satellite network, star network)
- One to many (Broadcast network)
- Many to many (Computer network—the Internet, adhoc network)

❖ Problem:

- Find the channel capacity
- Find the optimal distributed source and channel coding strategy

Ford-Fulkerson Theorem (Max-flow min-cut solution)



- ❖ $C_{AB} = \min\{C_1+C_2, C_2+C_3+C_4, C_4+C_5, C_1+C_3+C_5\}$
- ❖ The maximum flow across any cut-set cannot be greater than the sum of the capacities of the cut edges.
- ❖ Thus, min. of max. flow across cut-sets is an upper bound on the capacity
- ❖ The Theorem shows that this capacity can be achieved.
- ❖ Only applicable to some information theoretic problems

Jointly Typical Sequences

❖ Let S_i denote an ordered subset of random variables X_1, X_2, \dots, X_k

❖ Let S denote n independent copies of S_i

$$\Pr\{S=s\} = \prod_{i=1}^n \Pr\{S_i = s_i\}, s_i \in \mathcal{X}^n$$

❖ For example, if $S_i := (X_1, X_2)$, then

$$\begin{aligned} \Pr\{S=s\} &= \Pr\{(\mathbf{X}_1, \mathbf{X}_2) = (\mathbf{x}_1, \mathbf{x}_2)\} \\ &= \prod_{i=1}^n p(x_{1i}, x_{2i}) \end{aligned}$$

❖ By the weak law of large numbers, for any subset of random variables,

$$-1/n \log p(S_1, \dots, S_n) = -1/n \sum_{i=1}^n \log p(S_i) \Rightarrow H(S_1) \text{ --- (1)}$$

for all $2^k - 1$ subsets $S \subset \{X_1, X_2, \dots, X_k\}$

Definition of Jointly Typical Set $A_\varepsilon^{(n)}$

- ❖ $A_\varepsilon^{(n)} = A_\varepsilon^{(n)}(X_1, \dots, X_k)$
 $= \{(\mathbf{x}_1, \dots, \mathbf{x}_k) : |-1/n \log p(\mathbf{s}) - H(S)| < \varepsilon, \forall S \subset \{X_1, \dots, X_k\}\}$
- ❖ S is a subset of $\{X_1, \dots, X_k\}$
- ❖ For example, if $\{X_1, X_2\}$
 - $S = \{X_1\}, S = \{X_2\}, S = \{X_1, X_2\}$
 - $A_\varepsilon^{(n)}(X_1, X_2) = \{(\mathbf{x}_1, \mathbf{x}_2) :$
 - $| -1/n \log p(\mathbf{x}_1, \mathbf{x}_2) - H(X_1, X_2)| < \varepsilon,$
 - $| -1/n \log p(\mathbf{x}_1) - H(X_1)| < \varepsilon,$
 - $| -1/n \log p(\mathbf{x}_2) - H(X_2)| < \varepsilon\}$

Theorem 14.2.1: For any $\varepsilon > 0$, for sufficiently large n

1. $P(A_\varepsilon^{(n)}(\mathbf{S})) \geq 1 - \varepsilon, \forall \mathbf{S} \subset \{X_1, \dots, X_n\}$
2. $\mathbf{s} \in A_\varepsilon^{(n)}(\mathbf{S}) \Rightarrow p(\mathbf{s}) \stackrel{\circ}{=} 2^{-n(H(\mathbf{S}) \pm \varepsilon)}$
3. $|A_\varepsilon^{(n)}(\mathbf{S})| \stackrel{\circ}{=} 2^{-n(H(\mathbf{S}) \pm \varepsilon)}$
4. Let $S_1, S_2 \subset \{X_1, \dots, X_k\}$. If $(s_1, s_2) \in A_\varepsilon^{(n)}(S_1, S_2)$, then $p(s_1|s_2) \stackrel{\circ}{=} 2^{-n(H(S_1|S_2) \pm 2\varepsilon)}$

$$a_n \stackrel{\circ}{=} 2^{n(b \pm \varepsilon)}$$

❖ notation $\stackrel{\circ}{=}$ is to mean

$$|1/n \log a_n - b| < \varepsilon$$

❖ $1/n \log a_n - b < \varepsilon$

$$a_n < 2^{n(b+\varepsilon)}$$

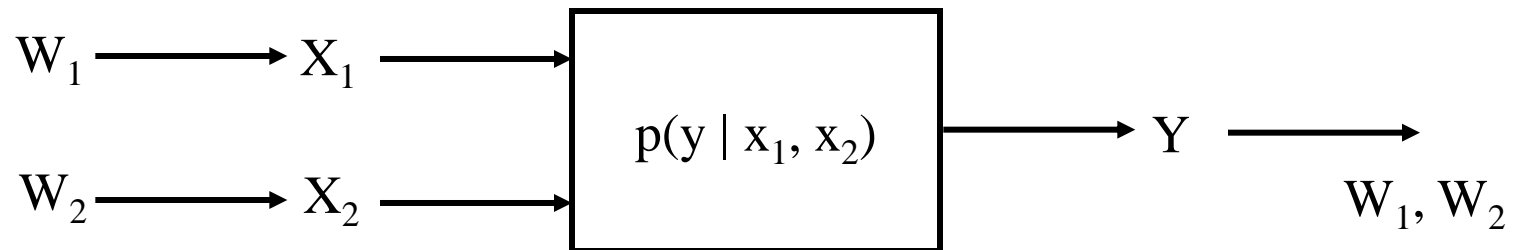
❖ $1/n \log a_n - b > -\varepsilon$

$$a_n > 2^{n(b-\varepsilon)}$$

The Multiple Access Channel

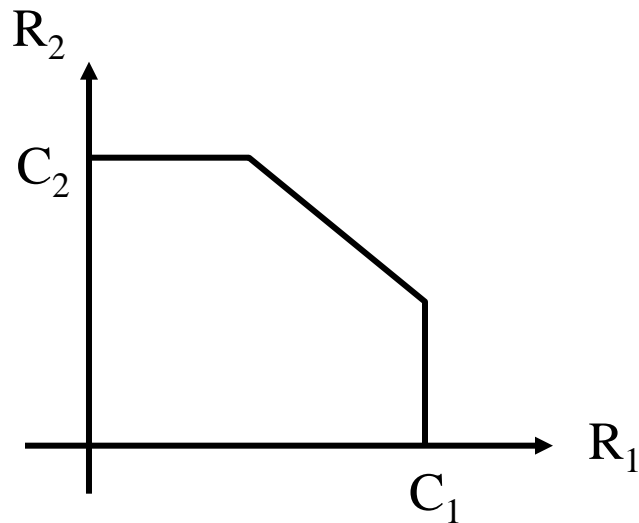
- ❖ Two or more senders and one receiver (Satellite or base station)
- ❖ The channel gives the relationship between the two input alphabets \mathcal{X}_1 and \mathcal{X}_2 and the output alphabet \mathcal{Y}
- ❖ $(2^{nR_1}, 2^{nR_2}, n)$ code has two sets of integers.
- ❖ $W_1 = \{1, 2, \dots, 2^{nR_1}\}$ and $W_2 = \{1, 2, \dots, 2^{nR_2}\}$ message sets
- ❖ Encoder is map
$$X_1: W_1 \rightarrow \mathcal{X}_1^n$$
$$X_2: W_2 \rightarrow \mathcal{X}_2^n$$
- ❖ Decoder is a map
$$g: \mathcal{Y}^n \rightarrow W_1 \times W_2$$

The Multiple Access Channel



- ❖ $P_e^{(n)} = 2^{-n(R_1+R_2)} \sum_{(w_1, w_2)} \Pr(g(Y^n) \neq (w_1, w_2) | (w_1, w_2) \text{ sent})$
- ❖ A rate pair (R_1, R_2) is achievable for the multiple access channel if there exists a sequence of $((2^{nR_1}, 2^{nR_2}, n)$ codes with $P_e^{(n)} \rightarrow 0$

The Multiple Access Channel Capacity [Thm14.3.1]



Hull: the outer covering of a fruit or seed

- ❖ The capacity of a multiple access channel is the closure of the *convex hull* of all (R_1, R_2) satisfying

$$R_1 < I(X_1; Y|X_2)$$

$$R_2 < I(X_2; Y|X_1)$$

$$R_1 + R_2 < I(X_1, X_2; Y)$$

for some product distribution $Q(x_1) Q(x_2)$

Proof of Thm 14.3.1

- ❖ Codebook: generate 2^{nR} independent codewords of length n
 - $\mathbf{X}_1(i), i \in \{1, 2, \dots, 2^{nR_1}\}$
 - $\mathbf{X}_2(j), j \in \{1, 2, \dots, 2^{nR_2}\}$
- ❖ Decoding: Let $A_\varepsilon^{(n)}$ the set of typical $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{y})$ sequences. The decoder chooses the pair (i, j) such that
$$(\mathbf{x}_1(i), \mathbf{x}_2(j), \mathbf{y}) \in A_\varepsilon^{(n)}$$
if such a pair (i, j) exists and is **unique**; otherwise, an error is declared.

Proof of Thm 14.3.1

- ❖ Error (assuming (1,1) was sent without loss of generality)
 - Either $(\mathbf{x}_1(1), \mathbf{x}_2(1))$ is not typical with the rec. sequence \mathbf{y}
 - Or there is a pair of incorrect codewords $(\mathbf{x}_1(i), \mathbf{x}_2(j))$ that is typical with the received sequence
- ❖ Note that $(W_1=1, W_2=1)$ are the message indices that were sent. We still need to choose the message $X_1(1)$ and $X_2(1)$.
 - Let's use $Q(x_1)$ and $Q(x_2)$ to make distinction
- ❖ $E_{ij} = \{(\mathbf{X}_1(i), \mathbf{X}_2(j), \mathbf{Y}) \in A^{(n)}_\epsilon\}$

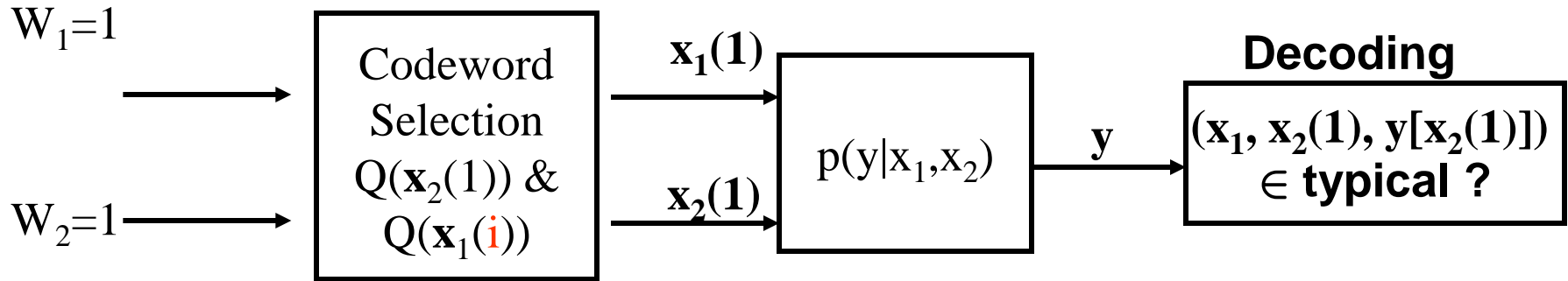
The Union Bound on Prob. Error

(Proof of Thm 14.3.1)

$$\begin{aligned} \blacklozenge P_e^{(n)} &= P(E_{11}^c \cup \cup_{(i,j) \neq (1,1)} E_{ij}) \\ &\quad - \text{Applying the union bound} \\ &\leq P(E_{11}^c) + \sum_{i \neq 1, j=1} P(E_{i1}) + \sum_{i=1, j \neq 1} P(E_{1j}) + \sum_{i \neq 1, j \neq 1} P(E_{ij}) \end{aligned}$$

- \blacklozenge Here, P is the conditional probability that $(W_1=1, W_2=1)$ was sent
- \blacklozenge $E_{11}^c = \{\text{codewords } \mathbf{x}_1(1) \text{ and } \mathbf{x}_2(1) \text{ are not typical with } \mathbf{y}\}$
- \blacklozenge $E_{i1} = \{\mathbf{x}_1(i), i \neq 1, \text{ is selected, and is jointly typical with } (\mathbf{x}_2(1), \mathbf{y})\}$
- \blacklozenge $E_{1j} = \{\mathbf{x}_2(j), j \neq 1, \text{ is selected, and is jointly typical with } (\mathbf{x}_1(1), \mathbf{y})\}$.

$$P(E_{i1})$$



- ❖ Marginal-1: $\mathbf{x}_2(1) = \mathbf{z}$ selected and sent over the channel
 - $p(\mathbf{y}|\mathbf{x}_1, \mathbf{x}_2) Q(\mathbf{x}_2(1) = \mathbf{z}) = q(\mathbf{y}, \mathbf{x}_2(1)=\mathbf{z} | \mathbf{x}_1)$
 - $q(\mathbf{y}, \mathbf{x}_2(1) = \mathbf{z}) = \sum_{\mathbf{x}_1} Q(\mathbf{x}_1) q(\mathbf{y}, \mathbf{x}_2(1) = \mathbf{z} | \mathbf{x}_1)$
- ❖ Marginal-2: Select the codeword $\mathbf{x}_1(i)$ later, independently drawn from $Q(\mathbf{x}_1)$, for the purpose of decoding
- ❖ Form a joint density using the two independent marginals:

$$Q(\mathbf{x}_1) q(\mathbf{y}(\mathbf{x}_2(1)), \mathbf{x}_2(1))$$
- ❖ Thus, the probability of drawing a pair is $q(\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}) = Q(\mathbf{x}_1) q(\mathbf{y}(\mathbf{x}_2(1)), \mathbf{x}_2(1))$
- ❖ Find the probability that the pair belongs to the jointly typical set

$$P(E_{i1}) = P\{(X_1(i), X_2(1), Y) \in A_{\varepsilon}^{(n)}\}$$

(Proof of Thm 14.3.1)

$$\diamond P(E_{i1}) = P\{(\mathbf{X}_1(i), \mathbf{X}_2(1), \mathbf{Y}) \in A_{\varepsilon}^{(n)}\}$$

$$= \sum_{(\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}) \in \text{typical}} Q(\mathbf{x}_1) q(\mathbf{y}, \mathbf{x}_2)$$

--- the joint prob. that the selections \mathbf{x}_1 and $(\mathbf{y}, \mathbf{x}_2)$ are drawn independently from $Q(\mathbf{x}_1)$ and $q(\mathbf{y}, \mathbf{x}_2)$

$$\leq |A_{\varepsilon}^{(n)}| \times 2^{-n(H(X_1) - \varepsilon)} \times 2^{-n(H(X_2, Y) - \varepsilon)}$$

$$\leq 2^{n(H(X_1, X_2, Y) - \varepsilon)} 2^{-n(H(X_1) - \varepsilon)} 2^{-n(H(X_2, Y) - \varepsilon)}$$

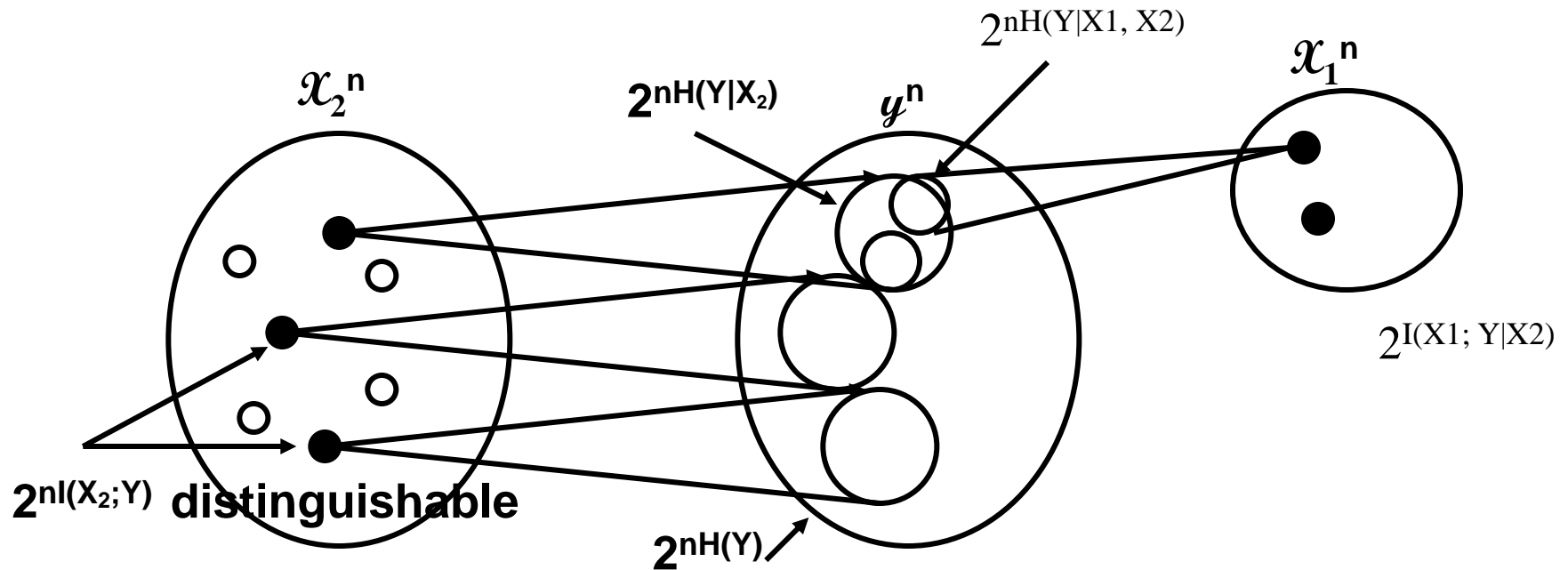
$$= 2^{-n[I(X_1; X_2, Y) - 3\varepsilon]}$$

$$= 2^{-n[I(X_1; Y|X_2) - 3\varepsilon]}$$

X_1 and X_2 independent, then $I(X_1; X_2, Y) = I(X_1; X_2) + I(X_1; Y|X_2) = I(X_1; Y|X_2)$

Proof of Thm 14.3.1 (4)

$$\begin{aligned}
 \diamond \sum_{i \neq 1, j=1} \mathbf{P}(E_{1j}) &\leq \sum_{i \neq 1, j=1} 2^{-n[I(X_1; Y|X_2) - 3\epsilon]} \\
 &\leq 2^{nR_1} 2^{-n[I(X_1; Y|X_2) - 3\epsilon]}
 \end{aligned}$$



Proof of Thm 14.3.1 (5)

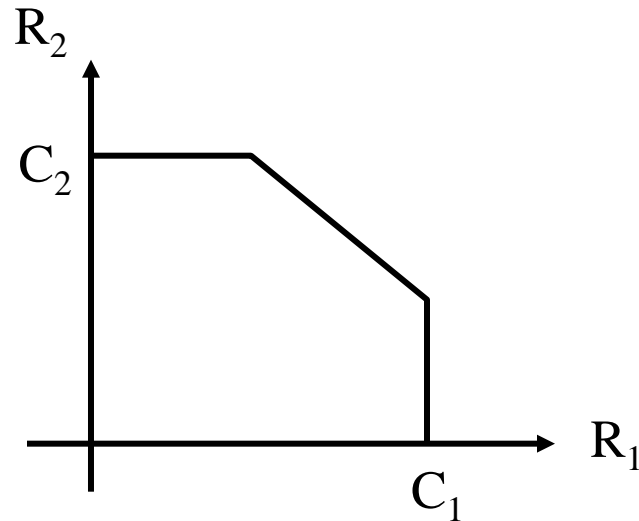
$$\begin{aligned}
 \text{❖ } P_e^{(n)} &= P(E_{11}^c \cup \cup_{(i,j) \neq (1,1)} E_{ij}) \\
 &\quad - \text{Applying the union bound} \\
 &\leq P(E_{11}^c) + \sum_{i \neq 1, j=1} P(E_{1j}) + \sum_{i=1, j \neq 1} P(E_{1j}) \\
 &\quad + \sum_{i \neq 1, j \neq 1} P(E_{ij}) \\
 &\leq P(E_{11}^c) + 2^{nR_1} 2^{-n[I(X_1; Y|X_2) - 3\epsilon]} + 2^{nR_2} 2^{-n[I(X_2; Y|X_1) - 3\epsilon]} \\
 &\quad + 2^{n(R_1+R_2)} 2^{-n[I(X_1, X_2; Y) - 4\epsilon]}.
 \end{aligned}$$

- ❖ All terms go to zero when the condition of the theorem is met.
- ❖ This bound shows that the average probability of error can be made arbitrarily small as $n \rightarrow \infty$, where the averaging is over all choices of codebooks in the random code construction.
- ❖ There exists at least one code that achieves arbitrary small probability of error

Two User Multiple Access Channel

- ❖ The channel gives the relationship between the two input alphabets \mathcal{X}_1 and \mathcal{X}_2 and the output alphabet \mathcal{Y}
- ❖ $(2^{nR_1}, 2^{nR_2}, n)$ code has two sets of integers.
- ❖ $W_1 = \{1, 2, \dots, 2^{nR_1}\}$ and $W_2 = \{1, 2, \dots, 2^{nR_2}\}$ message sets
- ❖ Encoder is map
$$X_1: W_1 \rightarrow \mathcal{X}_1^n$$
$$X_2: W_2 \rightarrow \mathcal{X}_2^n$$
- ❖ Decoder is a map
$$g: \mathcal{Y}^n \rightarrow W_1 \times W_2$$

The Multiple Access Channel Capacity [Thm14.3.1]



Hull: the outer covering of a fruit or seed

Convex?

- ❖ The capacity of a multiple access channel is the closure of the *convex hull* of all (R_1, R_2) satisfying

$$R_1 < I(X_1; Y|X_2)$$

$$R_2 < I(X_2; Y|X_1)$$

$$R_1 + R_2 < I(X_1, X_2; Y)$$

for some product distribution $Q(x_1)Q(x_2)$

Convexity of MAC Capacity Regions

❖ The capacity region \mathcal{C} for MAC is **convex**, i.e.

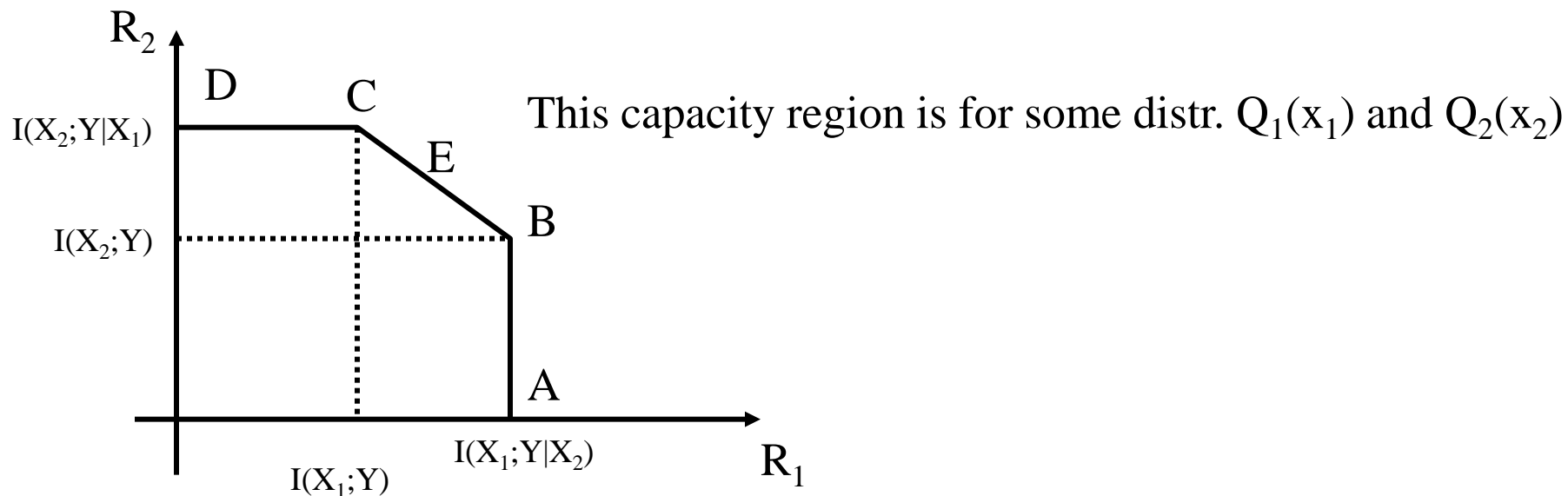
If $(R_1, R_2) \in \mathcal{C}$, and $(R_1', R_2') \in \mathcal{C}$,

then $(\lambda R_1 + (1-\lambda) R_1', \lambda R_2 + (1-\lambda) R_2') \in \mathcal{C}$ for $0 \leq \lambda \leq 1$

A little bit of Math Definitions

- ❖ Given a set $S \subset \mathbb{R}^n$
- ❖ A **convex combination** of elements of S is a vector of the form $\sum_{i=1}^m \lambda_i s_i$, where $s_i \in S$, $\lambda_i \geq 0$ and $\sum_{i=1}^m \lambda_i = 1$
- ❖ The **convex hull** of S , $\text{conv}(S)$, is the intersection of all convex sets containing S . Also, it is the set of all convex combinations from S .

Capacity Region for Multiple Access Channel



- ❖ Point A: maximum rate achievable for sender 1 when sender 2 is not sending any information at all.

$$\max R_1 = \max_{Q(x_1), Q(x_2)} I(X_1; Y|X_2) \quad \text{--- (2)}$$

- ❖ $I(X_1; Y|X_2) = \sum_{x_2} Q_2(x_2) I(X_1; Y|X_2=x_2) \leq \max_{x_2} I(X_1; Y|X_2 = x_2)$

Capacity Region for Multiple Access Channel

- ❖ Thus, the maximum in (2) can be achieved by selecting the input distribution $Q_1(x_1)$ and $Q_2(x_2)$ that maximize the conditional mutual information between X_1 and Y .
- ❖ Point B:
 - Now sender 2 also sends information.
 - The sender 2 achieves $I(X_2; Y)$, which is the rate obtained by treating X_1 as noise.
 - Once decoded, X_2 can be subtracted from the received
 - Then, we can obtain the channel, $p(y|x_1, x_2=x)$.
 - The sender 1 can achieve the rate $I(X_1; Y|X_2)$.

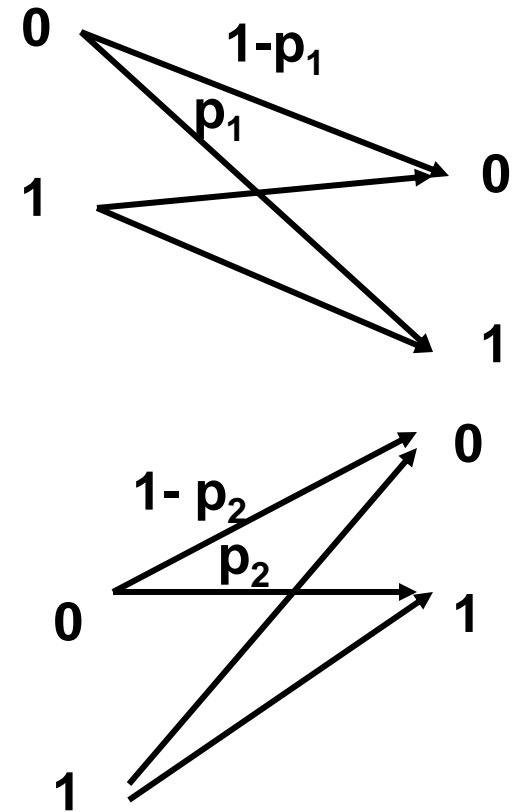
$I(X_1, X_2; Y)$: Where, in the region?

❖ Using the chain rule, we have

$$\begin{aligned} I(X_1, X_2; Y) &= I(X_1; Y) + I(X_2; Y|X_1) \\ &= I(X_2; Y) + I(X_1; Y|X_2) \end{aligned}$$

MAC for Independent Binary Symmetric Channels

- ❖ What is the MAC capacity region for this channel?



Binary Multiplier Channel

- ❖ $Y = X_1 X_2$, where X_1 and X_2 are binary
- ❖ What's the MAC capacity region?

Binary Erasure MAC

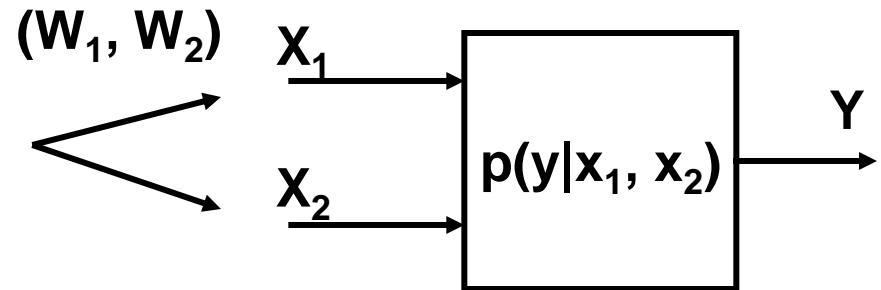
- ❖ Binary input $\{1, 0\}$ and ternary output $\{0, 1, 2\}$

$$Y = X_1 + X_2$$

- ❖ Ambiguity for $(X_1 = 1, X_2 = 0)$ or $(X_1 = 0, X_2 = 1)$ since for both $Y = 1$
- ❖ Obviously, $(R_1, R_2) = (1, 0)$ or $(0, 1)$ are achievable.
- ❖ How about when both transmitting?
 - Choose $R_1 = 1$, and determine the maximum rate at which the user-2 can send?
- ❖ Draw the capacity region?
- ❖ How do we achieve the point E?

The Cooperative Capacity of a Multiple Access Channel

- ❖ Both X_1 and X_2 are allowed to carry message W_1 and W_2
- ❖ Capacity region of this channel?
- ❖ Calculate the capacity for the binary erasure MAC, i.e., $Y = X_1 + X_2$ and compare this with the non-cooperative capacity



Gaussian Multiple User Channels

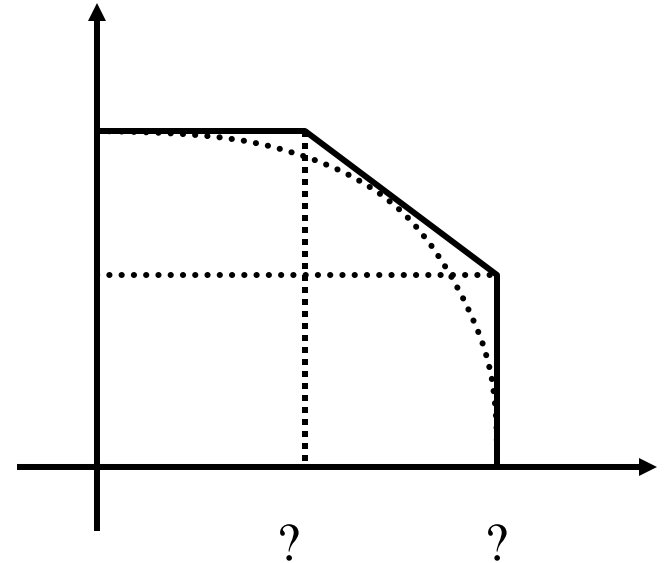
- ❖ Single user channel
 - $Y = X + Z, \quad Z \sim \mathcal{N}(0, N)$
 - $C(P/N) = 1/2 \log(1+P/N)$
- ❖ m transmitters, each with a power P , to one receiver
 - $Y = \sum_{i=1}^m X_i + Z$
 - The achievable rate region for the Gaussian channel is
 - $R_1 < C(P/N)$
 - $R_1 + R_2 < C(2P/N)$
 - ...
 - $R_1 + \dots + R_m < C(mP/N)$
 - The right hand side keeps increasing as m (logarithmic though)
- ❖ At high SNR, $C(mP/N) \approx 0.5 \cdot \log(m) + C(P/N)$
- ❖ What happens to the average rate per sender as $m \rightarrow \infty$?

Gaussian Multiple User Channels (2)

- ❖ The optimal code
 - m codebooks
 - the i -th codebook has 2^{nR_i} codewords of power P
- ❖ ML decoding
 - Compare all $2^{n(\sum R_i)}$ possible combinations and select the one that minimizes the likelihood (Min. Euclidean distance)
- ❖ As long as (R_1, \dots, R_m) is in the capacity region, then $P(e) \rightarrow 0$

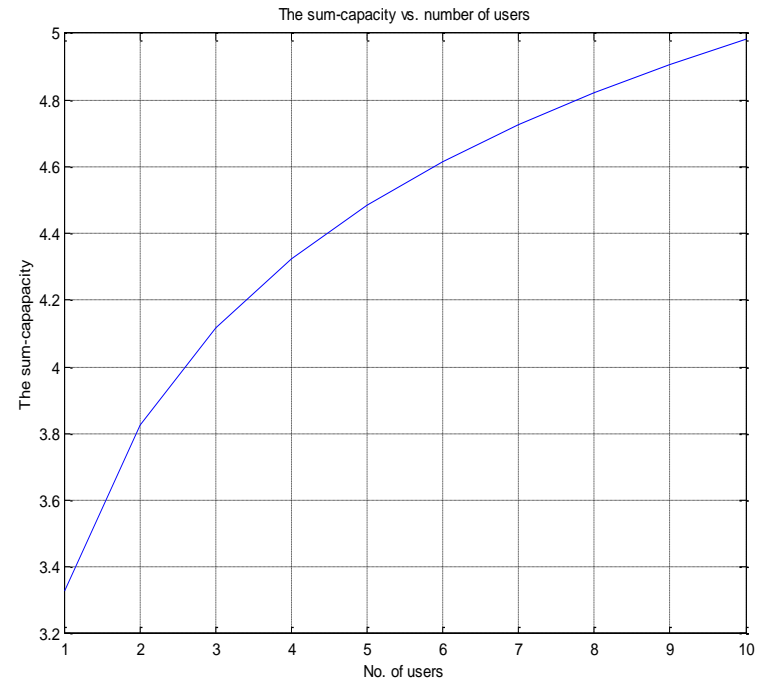
CDMA, TDMA, FDMA

- ❖ CDMA (Code division multiple access)
 - Simultaneous transmissions and receptions
 - Capacity region we obtained previous page
- ❖ TDMA/FDMA: hard division of the channel
- ❖ FDMA example:
 - Total RF spectrum $W = W_1 + W_2$ Hz, and Noise PSD = N_o [J/sec/Hz]
 - $R_1 = W_1 \log_2(1 + P_1/N_o W_1)$
 - $R_2 = W_2 \log_2(1 + P_2/N_o W_2)$



Gaussian Multiple User Channels (3)

- ❖ At high SNR, say 20 dB, let's investigate the behavior of the sum-capacity with the number of users increased
 - The capacity grows to infinity
 - The complexity of the receiver grows with $2^{n \cdot \text{sum-cap}}$
- ❖ Which one better?
 - MUD vs. Time-Sharing
- ❖ 2 x 1 case example
 - Time-Sharing: 3.3 bits/sec, and thus each user sending at a rate 1.65 bits/sec
 - MUD: 3.8 bits/sec, each user sending at a rate 1.9 bits/sec



Gaussian Multiple User Channels (4)

- ❖ What if the maximum rate B and the transmit power P are fixed ?
- ❖ For 2 users and 1 receiver case, then $2B$ bits/sec is achievable
- ❖ Time-Sharing: $B/2$ bits/sec per user
- ❖ MUD: B bits/sec per user

HW#8

- ❖ P15.1, P14.2, P14.3, P14.6, P14.9
- ❖ Compare CDMA, TDMA and FDMA systems by drawing the capacity regions for each of them in one figure
 - Use $N_0 = 0.1$ J/sec/Hz, $P_1 = P_2 = 10$ [J/sec], and $W = 10$ Hz
 - Do you think you need any convex combination operation (to make sure you have the convex hull)? Justify your answer.
- ❖ No need to turn in.

Binary Erasure MAC

- ❖ Binary input $\{1, 0\}$ and ternary output $\{0, 1, 2\}$

$$Y = X_1 + X_2$$

- ❖ Ambiguity for $(X_1 = 1, X_2 = 0)$ or $(X_1 = 0, X_2 = 1)$ since for both $Y = 1$

- ❖ Obviously, $(R_1, R_2) = (1, 0)$ or $(0, 1)$ are achievable.

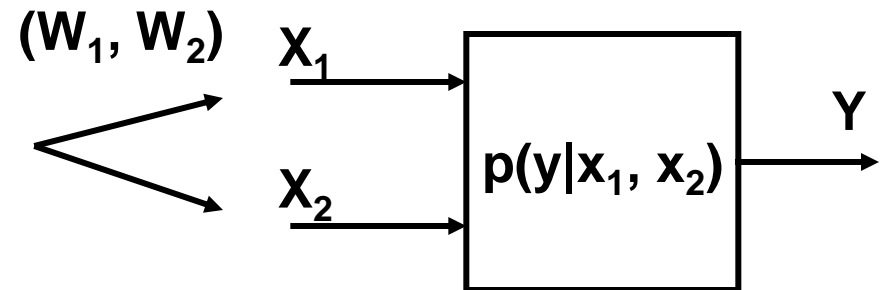
- ❖ How about when both transmitting?

- Choose $R_1 = 1$, and determine the maximum rate at which the user-2 can send?
- $Y = X_2 + X_1$, where X_1 is now an equally likely binary noise

Cooperative Capacity of Multiple Access Channel

- ❖ Calculate the capacity region for the binary erasure MAC, i.e., $Y = X_1 + X_2$ and compare this with the non-cooperative capacity
- ❖ Design $p(x_1, x_2)$ that maximizes the capacity
- ❖ $C = \max_{Q(x_1, x_2)} I(X_1, X_2; Y)$, instead of

$$\max_{Q(x_1)Q(x_2)} I(X_1, X_2; Y)$$



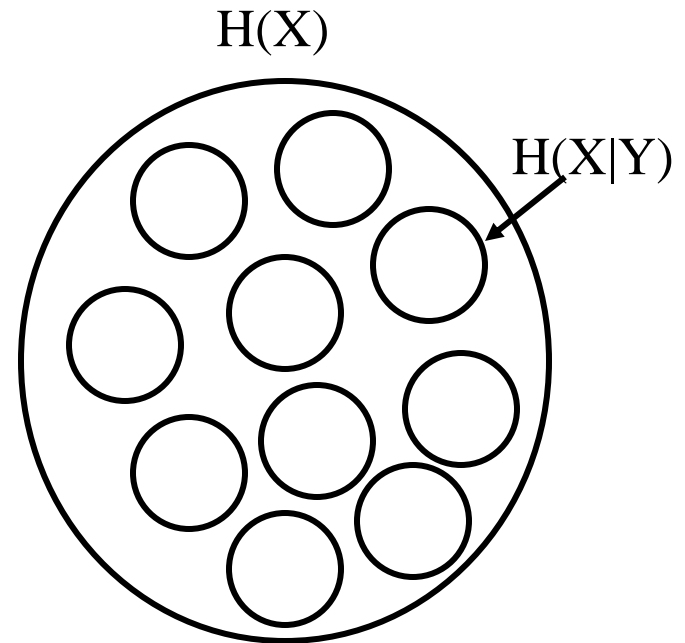
Cooperative Capacity of Multiple Access Channel

- ❖ Both X_1 and X_2 are allowed to carry both messages W_1 and W_2

Idea

$$\begin{aligned} \diamond 2^{nI(X; Y)} &= 2^{nH(X)} / 2^{nH(X|Y)} \\ &= 2^{n[H(X) - H(X|Y)]} \end{aligned}$$

- \diamond Now, the problem is that
- There are $2^{nI(X;Y)}$ number of slots.
 - I have 2^{nR} number of distinct balls numbered from 1 to 2^{nR} .
 - Suppose that the number of slots are a lot more than the number of balls, such that $2^{nI(X;Y)} / 2^{nR} = 2^{n\varepsilon}$
 - You can always choose a large enough n for every ε so that the ratio is large, i.e., $2^{n\varepsilon} = 2^{10}$.
 - Find the probability that I assign two or more balls to a same slot.



Idea (2)

- ❖ Let $B = 2^{nR}$ and $S = 2^{nI(X; Y)}$. $B \ll S$
- ❖ Let the index $b = 1, 2, \dots, B$; $s \in \{1, 2, \dots, S\}$.
- ❖ Balls are drawn from an urn B times with replacement.
What's the probability that you draw a ball more than once?
- ❖ $\Pr\{\text{distinct } B \text{ balls}\}$
 - $= \sum_{\text{first draw}} \Pr\{\text{distinct } B \text{ balls, first draw}\}$
 - $= \sum_{s_1=1}^S \Pr\{\text{distinct } B \text{ balls} | \text{first draw} = s_1\} \Pr\{\text{first draw} = s_1\}$
 - $= \sum_{s_1=1}^S (1/S) \Pr\{\text{distinct Balls} | \text{first draw} = s_1\}$
 - $= \Pr\{\text{distinct Balls} | s_1 = 1\}$
 - $= \sum_{s_2=1}^S \Pr\{\text{distinct } B \text{ balls, second draw} = s_2 | s_1 = 1\}$

Idea (3)

$$\begin{aligned} &= \sum_{s_2=1}^S \Pr\{\text{distinct balls, second draw}=s_2 \mid s_1 = 1\} \\ &= \Pr\{\text{distinct balls, } s_2=1 \mid s_1=1\} + \sum_{s_2=2}^S \Pr\{\text{distinct balls, } s_2 \\ &\quad \mid s_1=1\} \\ &= 0 + \sum_{s_2=2}^S \Pr\{\text{distinct Balls} \mid s_2, s_1=1\} \Pr\{s_2 \mid s_1=1\} \\ &= 0 + (S-1)(1/S) \Pr\{\text{distinct Balls} \mid s_1=1, s_2 = 2\} \\ &\dots \\ &= \frac{S-1}{S} \frac{S-2}{S} \dots \frac{S-B+1}{S} \end{aligned}$$

When $S \gg B$, the last term dominates

$$< (S - B) / S$$

Idea (4)

- ❖ $\Pr\{\text{Ambiguity}\} = 1 - \Pr\{\text{distinct balls}\} > 1 - (S-B)/S = B/S = 2^{-n(I(X;Y) - R)} = 2^{-n\epsilon}$
- ❖ Thus, we now have proved that with a random assignment of message indices to a vector in a slot, an arbitrary small decoding error can be achieved.
- ❖ This may not be a rigorous proof but will be very helpful for giving us insight in to what we are doing.

Idea (5)

❖ Instead, calculate the $\Pr\{\text{Ambiguity}\}$ directly

❖ $\Pr\{\text{Amb}\} = \Pr\{\text{Amb}|s_1 = 1\}$

$$= \sum_{s_2}^B \Pr\{\text{Amb}, s_2|s_1\}$$

$$= \sum_{s_2 \neq 1} \Pr\{\text{Amb}, s_2|s_1=1\} + \Pr\{\text{Amb}, s_2=1|s_1=1\}$$

$$= \Pr\{\text{Amb}|s_2=2, s_1=1\} + 1/S$$

$$= \Pr\{\text{Amb}|s_1=1, s_2=2, s_3=3\} + 2/S + 1/S$$

$$= (1/S) * [1 + 2 + \dots + B-1] = (1/S) * ((B-1)/2) * (1+B-1) \sim \approx B^2/2S$$

❖ $\Pr\{\text{Decoding error due to Amb}\}$

$$= \sum_{c \in \{\text{codewords}\}} P(c) \Pr\{\text{Amb}, \text{decoding error}|c\}$$

$$= \sum P(c) \Pr\{\text{Amb}\} \Pr\{\text{decoding error}|\text{Amb.}, c\}$$

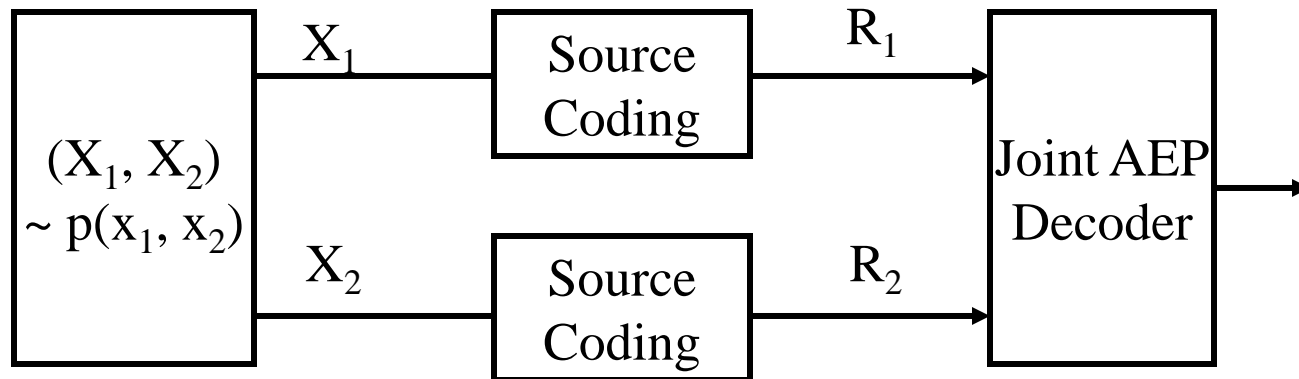
$$= (2/B) * \Pr\{\text{Amb}\} = B/S$$

← The indicator func. with 1 if c is one of the two confused words.

The Slepian-Wolf *Source* Coding

- ❖ A *distributed* source coding theorem where two sources must be encoded separately, but decoded together at a common node
 - For noise-less channels without interference
- ❖ Source Coding
 - To encode X_1 , a rate $R > H(X_1)$ is sufficient
 - To encode two sources (X_1, X_2) , a rate $H(X_1, X_2)$ is sufficient if we encode them together
 - If we encode them separately a sum rate $R > H(X_1) + H(X_2)$ is sufficient
 - But, Slepian Wolf showed that $R=H(X_1, X_2)$ is sufficient even for separate encoding of correlated sources

The Slepian-Wolf Source Coding (2)



- ❖ For the distributed source coding problem for the source (X_1, X_2) drawn iid $\sim p(x_1, x_2)$, the achievable rate region is given by

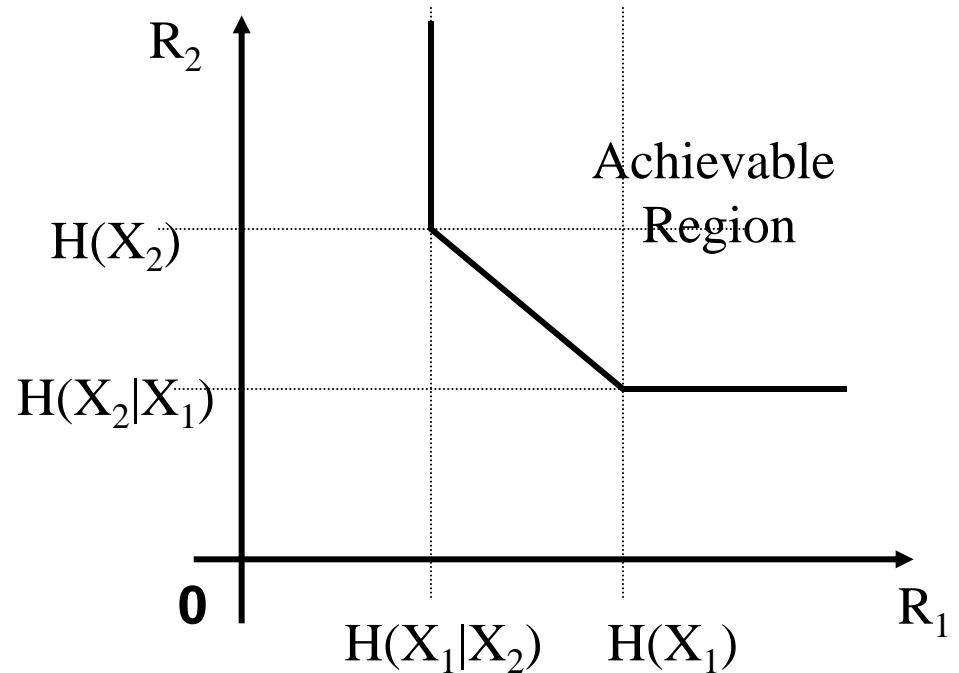
$$R_1 \geq H(X_1|X_2)$$

$$R_2 \geq H(X_2|X_1)$$

$$R_1 + R_2 \geq H(X_1, X_2)$$

Rate Regions for Slepian Wolf Encoding

- ❖ $H(X_1, X_2)$
 $= H(X_1) + H(X_2|X_1)$
 $= H(X_2) + H(X_1|X_2)$
 $\leq H(X_1) + H(X_2)$
- ❖ Equality, when?



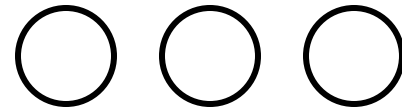
Example of Slepian-Wolf

$p(u, v)$	$u=0$	1
$v=0$	$1/3$	$1/3$
$v=1$	0	$1/3$

- ❖ $H(U, V) = \log_2(3) = 1.58$ bits
- ❖ $H(U) = -[(1/3) \log_2(1/3) + (2/3) \log_2(2/3)] = 0.9183$
- ❖ $H(V) = 0.9183$
- ❖ $H(U)+H(V) = 1.84$ bits
- ❖ You need only 1.58 bits, instead of 1.84 bits, with Slepian-Wolf encoding

Idea

- ❖ There are 3 balls and 7 different color paints—the rainbow.
- ❖ Take one color at random and color the balls
- ❖ Find the probability that there are two or more balls with the same color = $1 - \text{Pr}(\text{distinct colors})$.
- ❖ $\text{Pr}(\text{distinct colors}) = \sum_7 (1/7)(6/7)(5/7) = (6/7)(5/7)$
- ❖ Thus, if the number of colors are sufficiently large, the chance that the probability of having an equivocation is small, especially when $n \rightarrow \infty$

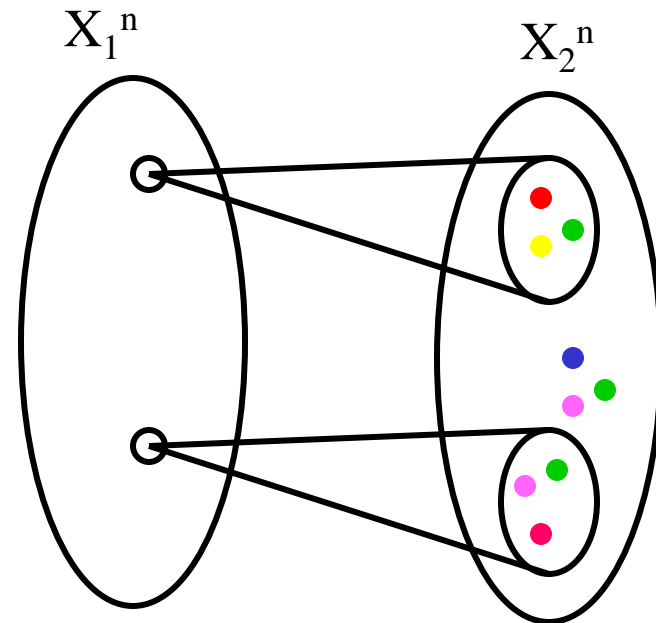


Idea (2)

- ❖ Number of balls: $B = 2^{nH(X_2|X_1)}$
- ❖ Number of colors: $C = 2^{nR}$
- ❖ $C \gg B$, $C/B = 2^{n\epsilon}$
- ❖ Then, the probability of decoding error due to non-distinctive colors in B balls $= 2^{-n\epsilon}$

Idea (3)

- ❖ Encode X_1 with rate $R_1 \geq H(X_1)$ and then the decoding should be no problem.
- ❖ Now, how to encode/decode X_2 ?
- ❖ Encode X_2 with a rate $R_2 \geq H(X_2|X_1)$: randomly color every elements in X_2^n with 2^{nR_2} colors
 - $\Pr(\text{decoding error due to color equivocation}, X_1 = i) = 2^{-nR_2} = 2^{-n\epsilon}$
 - $R_2 = H(X_2|X_1) + \epsilon$
 - Thus, we can choose a big enough n so that $n\epsilon$ be very big



The Slepian-Wolf Source Coding (3)

- ❖ Proof of existence of the code with any rate pair $(R_1, R_2) \geq H(X_1, X_2)$ of the code which maps
 - $f_1: X^n \rightarrow \{1, 2, \dots, 2^{nR_1}\}$
 - $f_2: X^n \rightarrow \{1, 2, \dots, 2^{nR_2}\}$
 - With AEP decoding the $P(e) \downarrow 0$ as $n \uparrow \infty$.
- ❖ Proof again uses
 - The typical set argument for the existence of a code
 - The Fano's inequality for the achievability
 - Refer to Cover and Thomas for proofs

Slepian Wolf Source-Encoding Idea

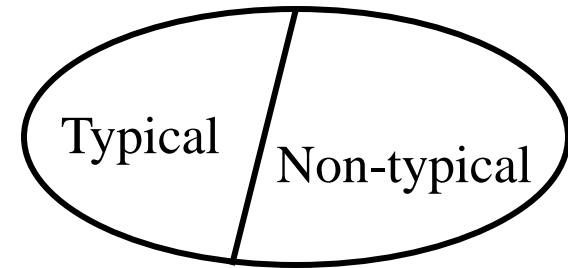
❖ Random code:

- For each $x^n \in X^n$, draw a uniform random number from 1 to 2^{nR_1} . Let the assignments f_1 and f_2 known to the encoder and decoder.

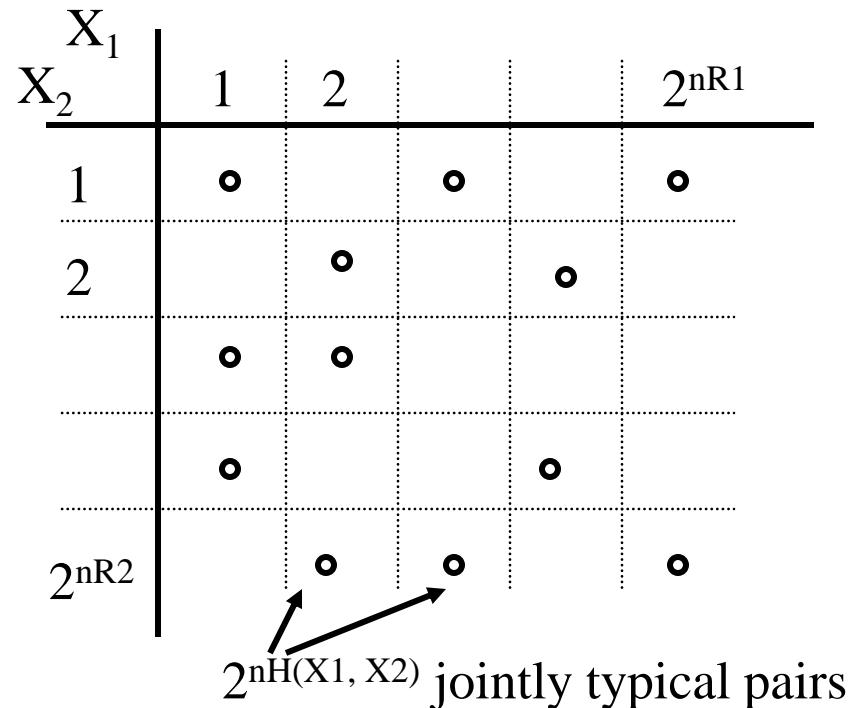
❖ Encoding (x_1^n, x_2^n) : indices $i = f_1(x_1^n)$ and $j = f_2(x_2^n)$ are sent.

❖ Decoding: Find if there is a **unique** pair $(x_1', x_2') \in A^{(n)}_\epsilon$ where

- $x_1' \in \{x_1^n: f_1(x_1^n) = i\}$ and
- $x_2' \in \{x_2^n: f_2(x_2^n) = j\}$, i.e., the inverse images of i and j respectively.



$$f_1: \mathcal{X}^n \mapsto \{1, 2, \dots, 2^{nR_1}\}$$



Slepian Wolf for Many Sources

❖ $(X_1, X_2, \dots, X_m) \sim p(x_1, \dots, x_m)$

❖ The rate region is

$$R(S) > H(X(S)|X(S^c))$$

where

- $S \subseteq \{1, 2, \dots, m\}$,
- $X(S) = \{X_j: j \in S\}$
- $R(S) = \sum_{i \in S} R_i$

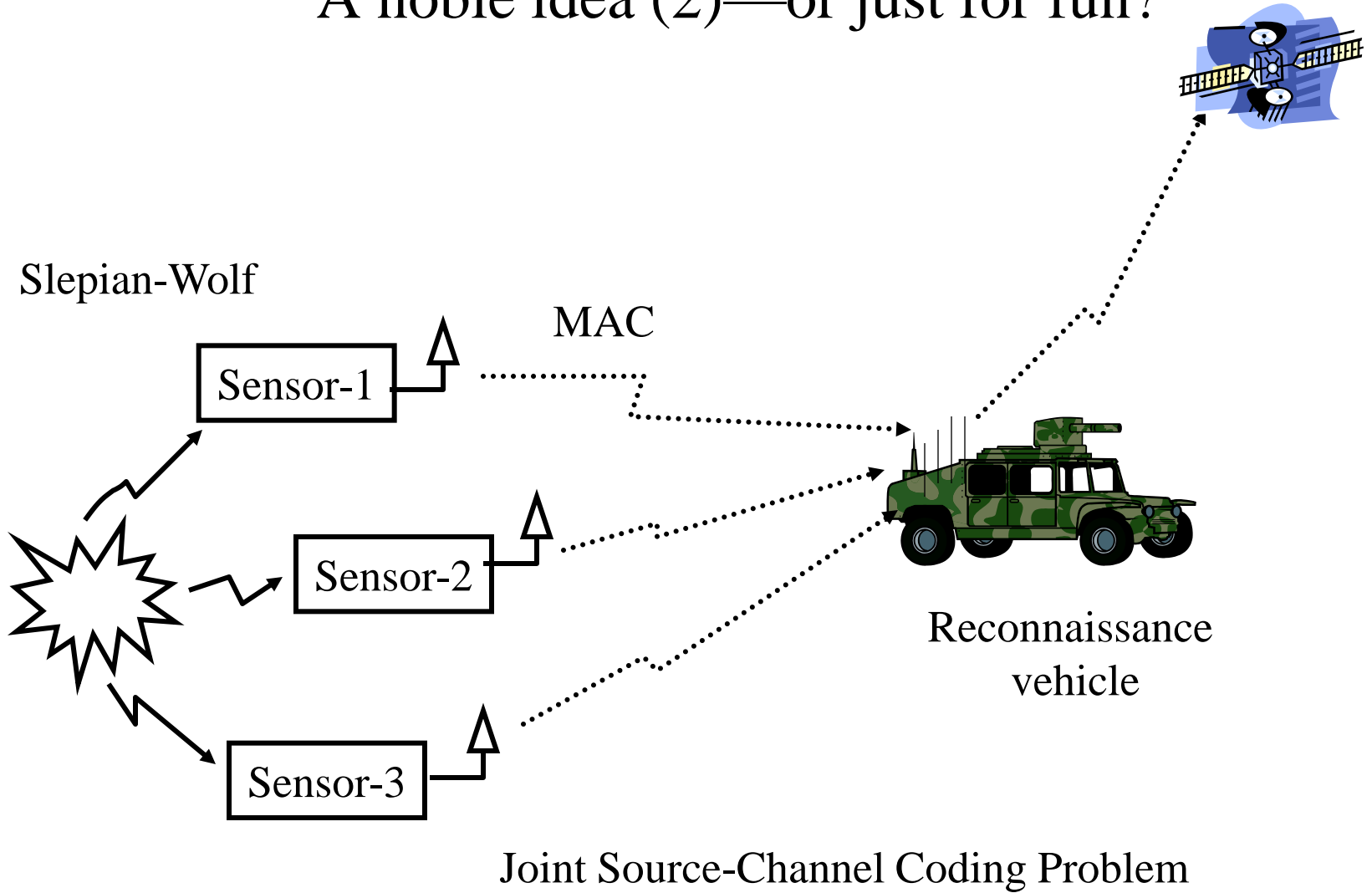
❖ Ex) 3 sources $(X_1, X_2, X_3) \sim p(x_1, x_2, x_3)$

- $R_1 > H(X_1|X_2, X_3) \ \& \ R_2 > H(X_2|X_1, X_3) \ \& \ R_3 > H(X_3|X_2, X_1)$
- $R_1+R_2 > H(X_1, X_2|X_3) \ \& \ R_1+R_3 > H(X_1, X_3|X_2) \ \& \ R_2+R_3 > H(X_2, X_3|X_1)$
- $R_1+R_2+R_3 > H(X_1, X_2, X_3)$

A noble idea—or just for fun?

- ❖ Use the powerful LDPC codes to encode the source data with a certain $H(X_1, X_2, X_3)$
 - Huffman code will not be a good idea—why?
- ❖ Choose a rate pair (R_1, R_2, R_3) according to Slepian-Wolf, and design three independent codes whose code rate pair is slightly larger, i.e., $(R_1+\epsilon, R_2+\epsilon, R_3+\epsilon)$
 - What's the ϵ for?
 - What should be the magic secret here?
- ❖ You want to consolidate the information collected by three sensors observing a common phenomenon (a tank goes by)

A noble idea (2)—or just for fun?



Appendix

- ❖ Convexity of the rate region
- ❖ Use an auxiliary random variable $Q = \{1, 2, 3, 4\}$

Convexity of MAC Capacity Regions

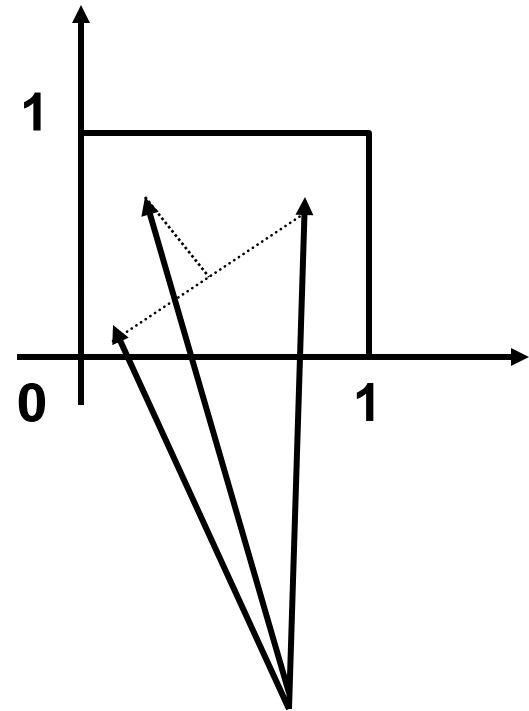
- ❖ The capacity region \mathcal{C} for MAC is convex, i.e.
- ❖ If $(R_1, R_2) \in \mathcal{C}$, and $(R_1', R_2') \in \mathcal{C}$, then $\lambda R_1 + (1-\lambda) R_2 \in \mathcal{C}$ for $0 \leq \lambda \leq 1$

A little bit of Math Definitions

- ❖ Given a set $S \subset \mathbb{R}^n$
- ❖ A **convex combination** of elements of S is a vector of the form $\sum_{i=1}^m \lambda_i s_i$, where $s_i \in S$, $\lambda_i \geq 0$ and $\sum_{i=1}^m \lambda_i = 1$
- ❖ The **convex hull** of S , $\text{conv}(S)$, is the intersection of all convex sets containing S . Also, it is the set of all convex combinations from S .
- ❖ Every s in $\text{conv}(S)$ can be represented as a convex combination of vectors s_1, \dots, s_m from S .
- ❖ Caratheodory Theorem: Any point $s \in \text{conv}(S)$ can be represented with $n+1$ or fewer convex comb. of points from S (see example next)

A little bit of Math Definitions

- ❖ $S = \{(0, 0), (1,0), (0,1), (1,1)\}$
- ❖ $0.5(1, 0) + 0.5(0, 1) = (0.5, 0.5)$
- ❖ Two dimensional: At most three-vectors are needed.



The MAC capacity using Time-Sharing R.V. Q

❖ Achievable rates of a MAC is given by the closure of the set of all (R_1, R_2) pairs satisfying

$$R_1 < I(X_1; Y|X_2, Q),$$

$$R_2 < I(X_2; Y|X_1, Q),$$

$$R_1 + R_2 < I(X_1, X_2; Y| Q),$$

for some choice of the joint distribution

$p(q)p(x_1|q)p(x_2|q)p(y|x_1, x_2)$ with $Q \in \{1,2,3,4\}$.

Information Theory

Broadcast Channel Capacity

Agenda

❖ Broadcast Channel

- General broadcast channel
- Degraded broadcast channel
- Broadcast channel capacity solved by the dirty paper coding technique

Broadcast Channel Capacity

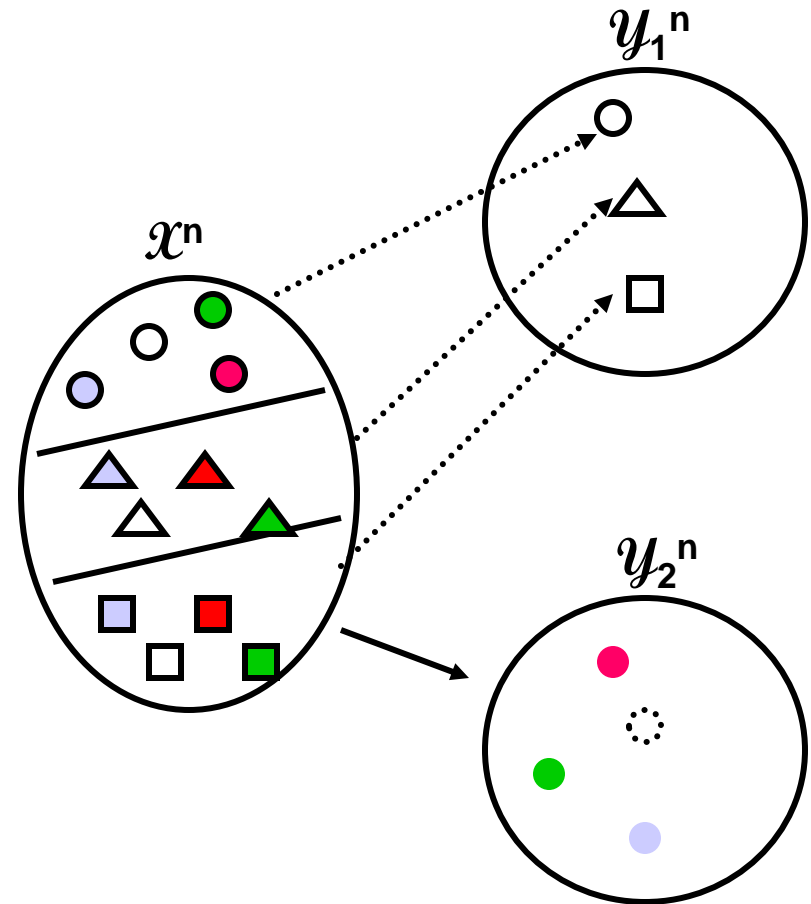
- ❖ One source, many receivers
- ❖ The channel $p(y_1, y_2|x)$ is memory-less:

$$p(y_1^n, y_2^n|x^n) = \prod_{i=1}^n p(y_1, y_2|x)$$

- ❖ Code: $((2^{nR_1}, 2^{nR_2}), n)$ code,
- ❖ $P_e^{(n)} = P(g_1(Y_1^n) \neq W_1, g_2(Y_2^n) \neq W_2)$

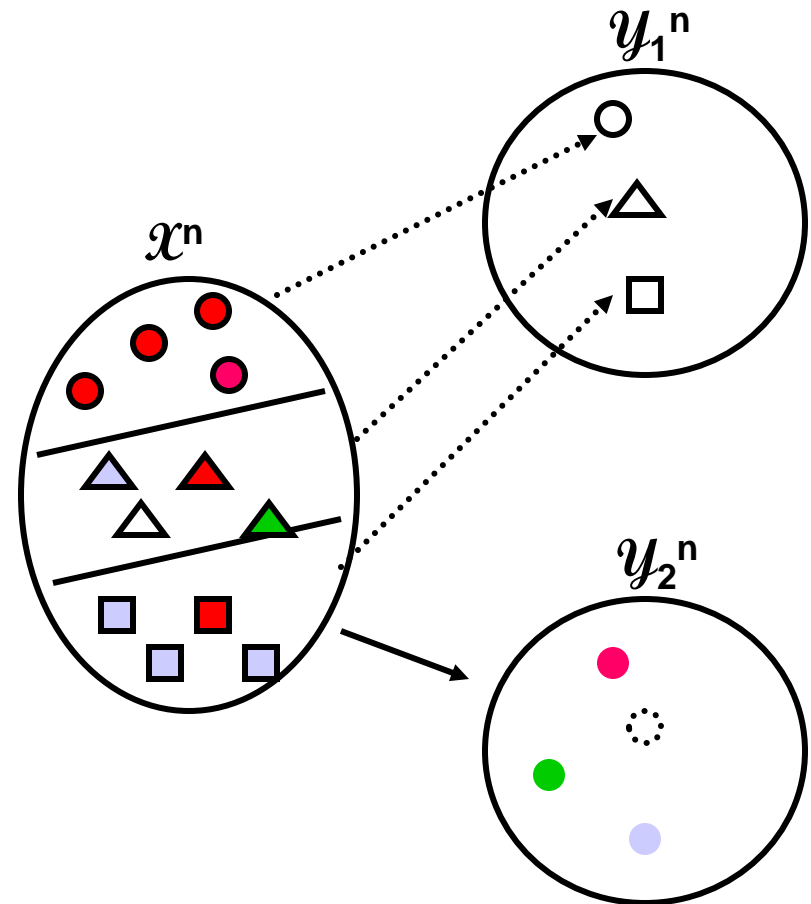
When Channel Introduce No Errors At all

- ❖ Assume $|X| > |Y_1|, |Y_2|, |Y_1| \times |Y_2|$
- ❖ $R_1 \leq H(Y_1)$
- ❖ $R_2 \leq H(Y_2)$
- ❖ $R_1 + R_2 \leq H(Y_1, Y_2)$
- ❖ No errors: $X \rightarrow (Y_1, Y_2)$
- ❖ Note that $(y_1, y_2|x) = 1$ or 0 .
- ❖ The coding scheme: For each and every typical pair (y_1^n, y_2^n) , we simply need to assign a single deterministic input x^n . As long as the rate is within the capacity region, the scheme should work
 - Ex) $R_1 + R_2 < H(Y_1, Y_2)$



When Colors and Shapes are correlated:

- ❖ When $p(\text{red}|\text{circle}) \approx 1$ and $p(\text{blue}|\text{square}) \approx 1$
- ❖ $I(Y_1; Y_2) > 0$
- ❖ $H(X, Y) < H(X) + H(Y)$ (strictly)
- ❖ We must make sure that the sum rate chosen is smaller than $H(X, Y)$.

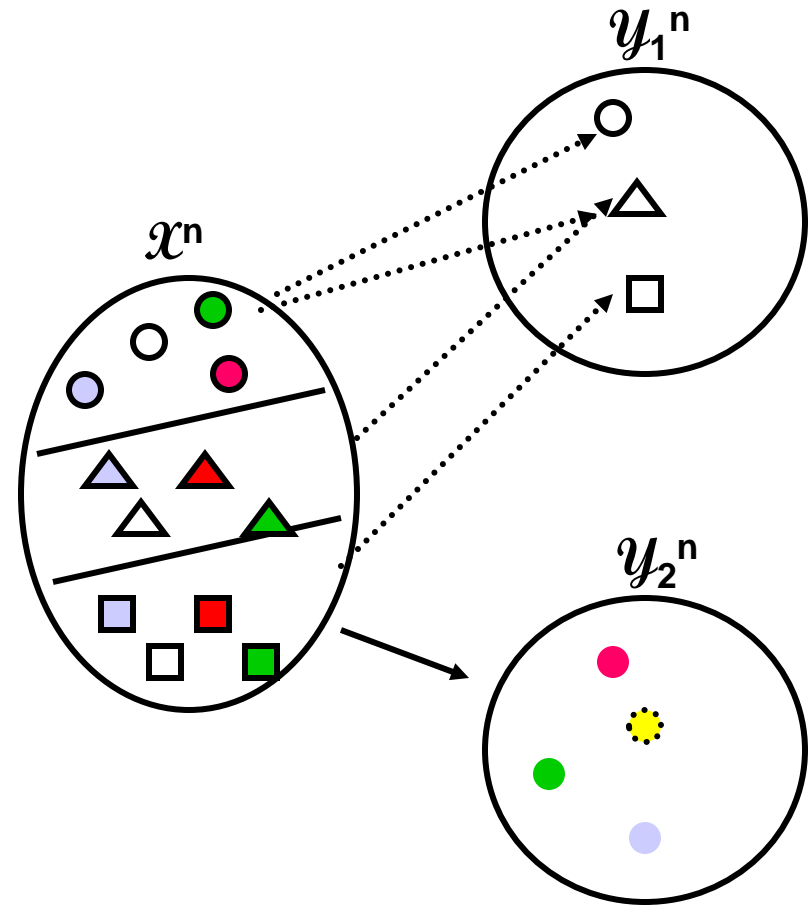


When Channel Introduces No Errors At all

- ❖ Auxiliary r.v.s U, V
 - U is the partition
 - V is the colors
- ❖ $R_1 \leq H(Y_1) - H(Y_1|U)$

$$= I(U; Y_1)$$
 - Ex) red circle becomes red triangle.
- ❖ $R_2 \leq H(Y_2) - H(Y_2|V)$

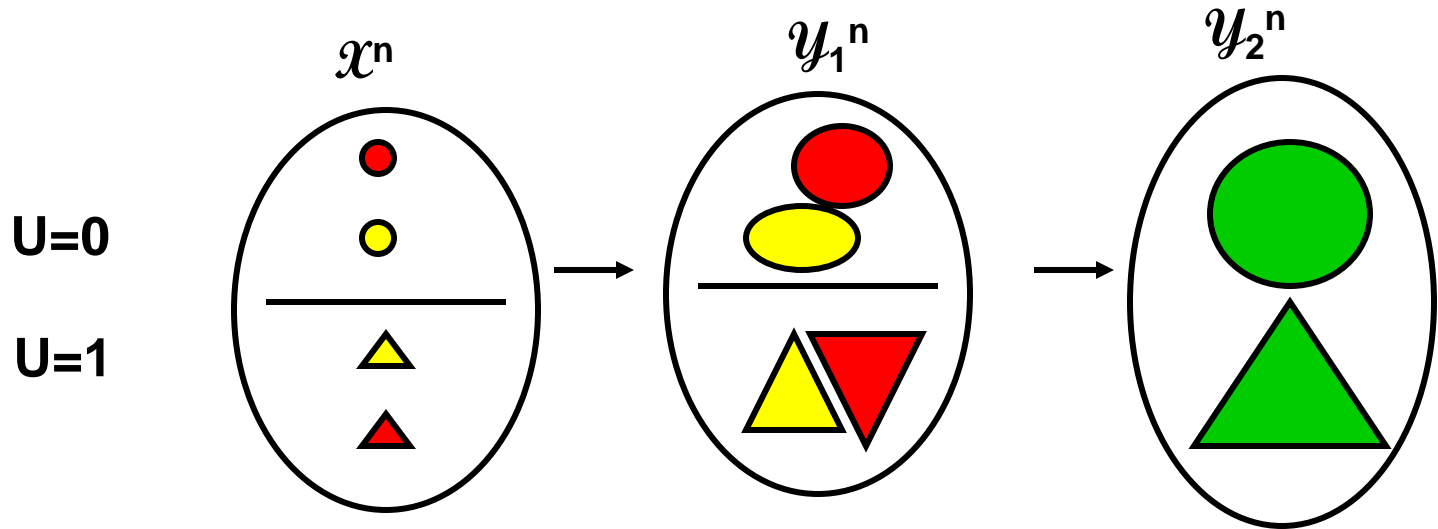
$$= I(V; Y_2)$$
 - Ex) white becomes yellow; green sometimes becomes yellow too.
- ❖ $R_1 + R_2 \leq I(U; Y_1) + I(V; Y_2) - I(U; V)$



Degraded Broadcast Channels

- ❖ Definition: if $p(y_1, y_2|x) = p(y_1|x) p(y_2|y_1)$, the channel is said to be physically degraded.
- ❖ Assumption is that one of the channel is better than the other channel.
- ❖ Degraded broadcast channel can be understood as: $X \rightarrow Y_1 \rightarrow Y_2$
- ❖ The capacity region of this channel is
$$R_2 \leq I(U; Y_2)$$
$$R_1 \leq I(X; Y_1|U)$$
for some utility random variable U whose cardinality $|U| \leq \min\{|X|, |Y_1|, |Y_2|\}$.
- ❖ NOTE: The rate R_1 and R_2 are rates of INDEPENDENT informations for user-1 and user-2 respectively.

Degraded Broadcast Channels (2)

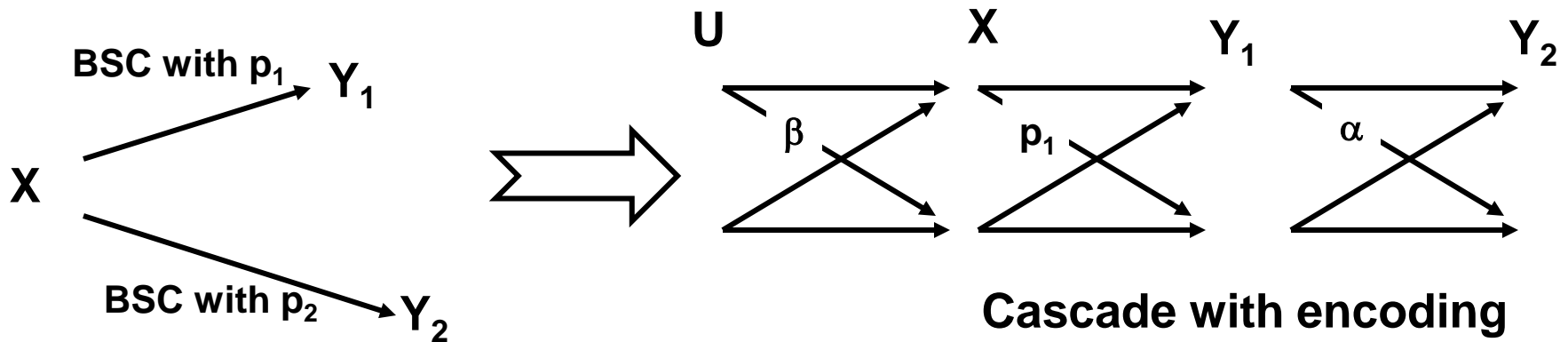


- ❖ As long as $R_2 \leq I(U; Y_2)$, and
- ❖ As long as $R_1 \leq I(X; Y_1|U)$, we can encode the data and obtain an error-free transmission
- ❖ Note that the code $\mathcal{C}_2 (2^{nR_2}, n)$ must be decodable to both receivers; the code $\mathcal{C}_1 (2^{nR_1}, n)$ is only decodable at the receiver 1.
- ❖ Proof: Use the typical set argument again (see textbook)

Common Broadcast Channel

- ❖ The independent rate pair (R_1, R_2) can be re-interpreted as $(R_0, R_1 - R_0, R_2 - R_0)$ with a common rate R_0 is achievable, provided that $R_0 \leq \min(R_1, R_2)$.
- ❖ TV broadcasting
 - HDTV: R_1
 - Color TV: R_2
 - $R_1 > R_2$

Binary Symmetric Broadcast Channels



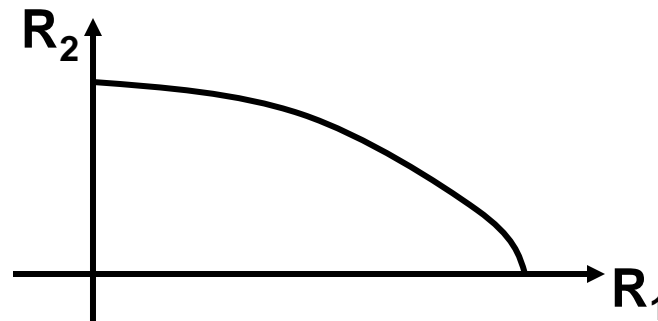
- ❖ Assumption: $p_1 < p_2 < \frac{1}{2}$
- ❖ U is the utility random variable:
 - $p(x) = \sum_u p(x, u) = \sum_u p(x|u) p(u)$
- ❖ The degradation from Y_1 to Y_2 : BSC with crossover α
 - $p_2 = p_1(1-\alpha) + (1-p_1)\alpha$
 - $\alpha = (p_2 - p_1)/(1-2p_1)$
- ❖ $R_2 \leq I(U; Y_2) = \max_p(u) H(Y_2) - H(Y_2|U) = 1 - H(\gamma)$
 - where $\gamma = \beta (1-p_2) + (1-\beta) p_2$

Binary Symmetric Broadcast Channels (2)

$$\begin{aligned} \diamond R_1 &\leq I(X; Y_1 | U) = H(Y_1 | U) - H(Y_1 | X, U) \\ &= H(Y_1 | U) - H(Y_1 | X) \\ &= H(\beta(1-p_1) + (1-\beta)p_1) - H(p_1) \end{aligned}$$

Example)

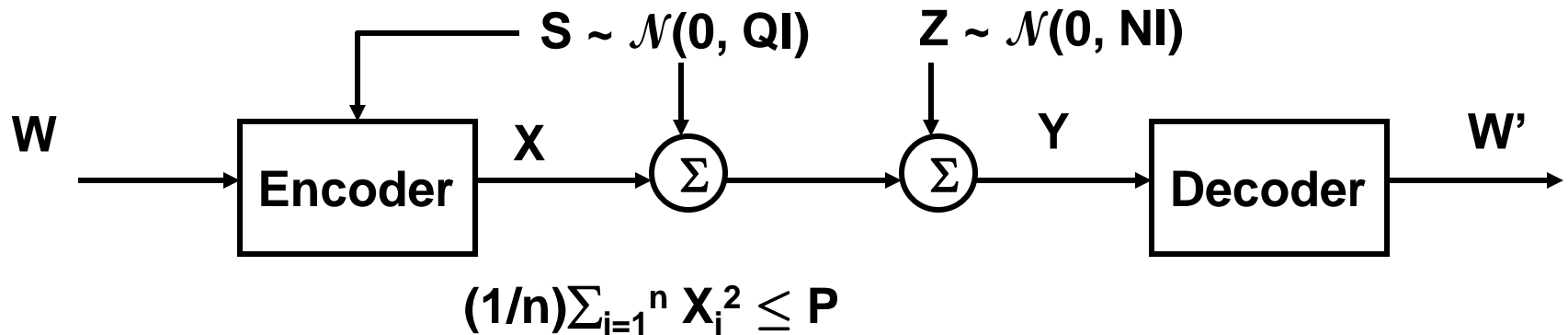
- When $\beta = 0$, $R_1 = 0$, $R_2 \leq 1 - H(p_2)$
- When $\beta = 1/2$, $R_1 \leq H(1/2) - H(p_1) = 1 - H(p_1)$, and $R_2 = 0$ since $\gamma = 1/2$.



Broadcasting Gaussian Channels

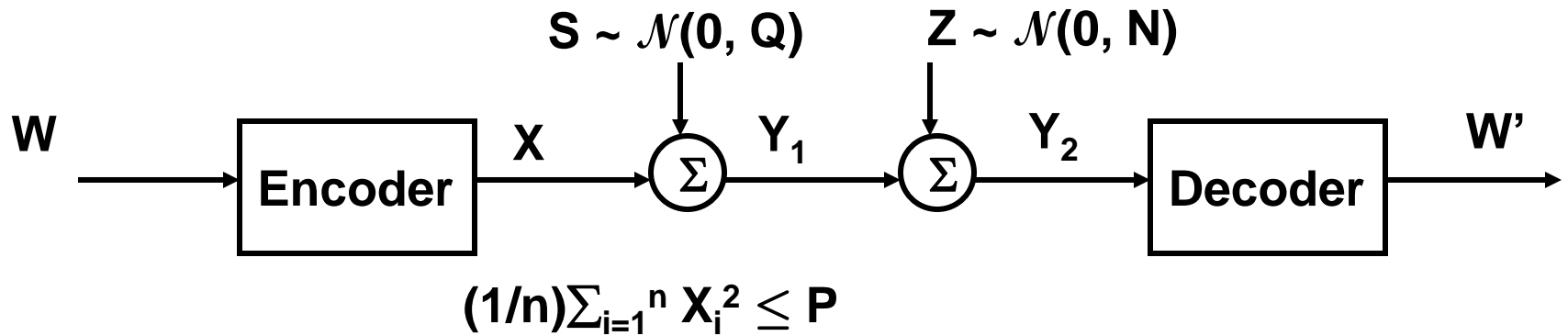
- ❖ $Y_1 = X + Z_1$ where $Z_1 \sim \mathcal{N}(0, N_1)$
- ❖ $Y_2 = X + Z_1 + Z_2 = X + Z_3$ where $Z_3 \sim \mathcal{N}(0, N_3 = N_1 + N_2)$, $N_3 > N_1$
- ❖ Achievable independent rate pair (R_1, R_2) is
 - $R_1 < C(\alpha P/N_1)$
 - $R_2 < C((1-\alpha)P)/(\alpha P + N_3)$
 - α is a parameter that the transmitter can choose
- ❖ Encoding: two codebooks
 - \mathcal{C}_1 with rate R_1 and with power αP ,
 - \mathcal{C}_2 with rate R_2 and with power $(1-\alpha)P$
 - $X = X_1 + X_2$,
- ❖ Decoding:
 - At Y_2 , search codebook \mathcal{C}_2 for match
 - At Y_1 , first decode Y_2 's codeword and then subtract; then decode for the first codebook

Dirty Paper Coding



- ❖ Max Costa, “Writing on Dirty Paper,” IT, 1983.
- ❖ $Y = X + S + Z$
- ❖ Results: When the encoder knows the interference S (non-causal), but the decoder does not, then the channel capacity is $C(P/N) = 0.5 \ln(1+P/N)$.
- ❖ Obvious/wrong option (when $P > Q$):
 - Encode $X = X' - S$; Thus, $Y = X' + Z$. Assuming all Gaussian, $P = \sigma_{x'}^2 + Q$;
 - Thus, variance of $X' = P - Q$;
 - Thus, we achieve $C((P-Q)/N)$ where $C(x) = 0.5 \ln(1+x)$.
 - Certainly, this is worse than what the DPC promises.
- ❖ Read the paper for further details.

Dirty Paper Coding (2)



- ❖ $Y = X + S + Z$
 - $Y_1 = X + S = X_1 + X_2 + S$ --- the first channel (want to have zero rate across)
 - $Y_2 = X+S+Z = X_1 + X_2 + S + Z$ --- the second channel (want to maximize the rate)
- ❖ Power distribution on αP and $(1-\alpha)P$ on C_1 and C_2
 - $C_1 < C(\alpha P/Q)$
 - $C_2 < C((1-\alpha)P/(\alpha P + Q + N))$
- ❖ By choosing $\alpha = Q/(P+N)$,
 - $C_1 < 0$
 - $C_2 < C(P/N_2)$ Q.E.D.

Dirty Paper Coding (3)

- ❖ The capacity of user-2 with poor channel is maximized.
- ❖ The decoder does not know the state S ; but the decoder knows the structure of the encoder.

Dirty Paper Coding (3)

- ❖ Generate $2^{nI(U^*; Y)}$ sequences of U^* drawn from $\mathcal{N}(0, P + \alpha^* 2Q)$

Information Theory

Course Overview

Agenda

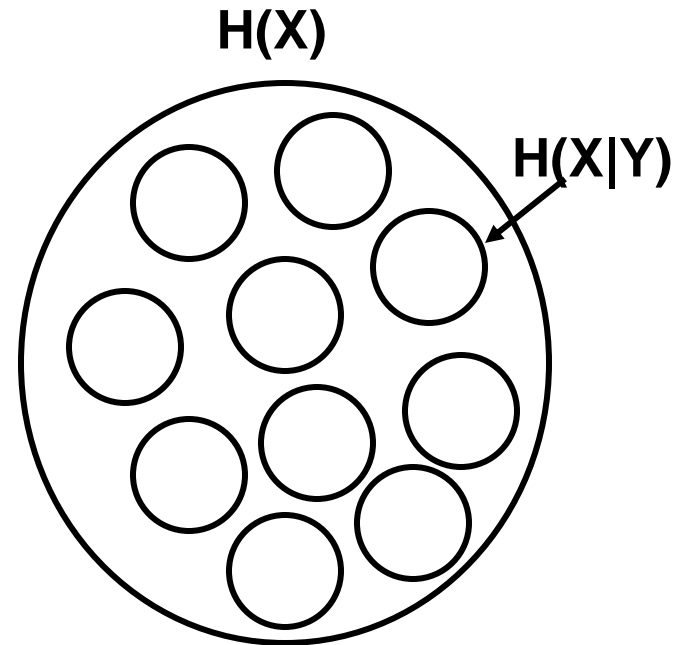
- ❖ Network Information Theory
- ❖ Current status
- ❖ Future direction
- ❖ Final
- ❖ Use of MIMO nodes in Networks

The Key Idea in IT is throwing balls at slots!!!!

❖ $2^{nI(X; Y)} = 2^{nH(X)}/2^{nH(X|Y)}$

❖ Suppose

- $2^{nI(X;Y)}$ number of **slots**.
- 2^{nR} distinct **balls**, from 1 to 2^{nR} .
- Choose small ε , $R = I(X; Y) - \varepsilon$.
- Choose large n , i.e.,
- $2^{nI(X;Y)} \gg 2^{nR}$, i.e.,
 $2^{nI(X;Y)}/2^{nR} = 2^{n\varepsilon} = 2^{10}$.
- # slots \gg # balls



❖ $\Pr\{\text{\#balls at any slot} > 1\} = \Pr\{\text{ambiguity}\}$

Key Idea (2)

- ❖ Let $S = 2^{nI(X; Y)}$ and $B = 2^{nR}$. Then, $S \gg B$.
- ❖ Let the index $b = 1, 2, \dots, B$; $s \in \{1, 2, \dots, S\}$.
- ❖ $\Pr\{\text{ambiguity}\} = 1 - \Pr\{\text{distinct slot}\}$
- ❖ $\Pr\{\text{distinct slot}\}$
 - $= \sum_{\text{first draw}} \Pr\{\text{distinct slot, first draw}\}$
 - $= \sum_{s_1=1}^S \Pr\{\text{distinct slot} \mid \text{first draw} = s_1\} \Pr\{\text{first draw} = s_1\}$
 - $= \sum_{s_1=1}^S (1/S) \Pr\{\text{distinct slot} \mid \text{first draw} = s_1\}$
 - $= \Pr\{\text{distinct slot} \mid s_1 = 1\}$
 - $= \sum_{s_2=1}^S \Pr\{\text{distinct slot, second draw} = s_2 \mid s_1 = 1\}$

Idea (3)

$$\begin{aligned} &= \sum_{s_2=1}^S \Pr\{\text{distinct slot, second draw}=s_2 \mid s_1 = 1\} \\ &= \Pr\{\text{distinct slot, } s_2=1 \mid s_1=1\} + \sum_{s_2=2}^S \Pr\{\text{distinct slot, } s_2 \mid \\ &\quad s_1=1\} \\ &= 0 + \sum_{s_2=2}^S \Pr\{\text{distinct slot} \mid s_2, s_1=1\} \Pr\{s_2 \mid s_1=1\} \\ &= 0 + (S-1)(1/S) \Pr\{\text{distinct slot} \mid s_1=1, s_2 = 2\} \\ &\dots \\ &= \frac{S-1}{S} \frac{S-2}{S} \dots \frac{S-B+1}{S} \end{aligned}$$

When $S \gg B$, the last term dominates

$$\sim (S - B) / S$$

Idea (4)

$$\begin{aligned} \diamond \Pr\{\text{Ambiguity}\} &= 1 - \Pr\{\text{distinct balls}\} \sim 1 - (S-B)/S = \\ &B/S = 2^{-n(I(X;Y) - R)} = 2^{-n\epsilon} \end{aligned}$$

◇ This shows that

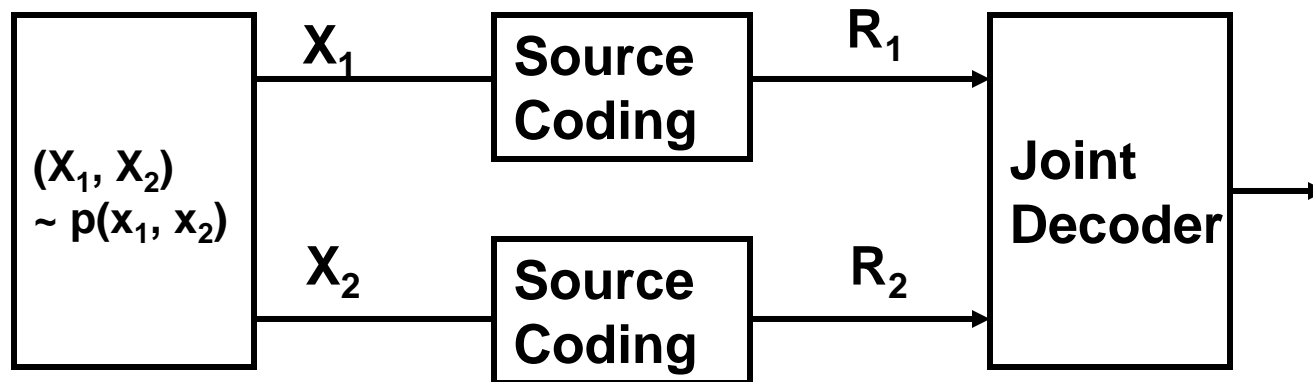
- with a random assignment of message indices to a vector in a slot, an arbitrary small error can be achieved.

◇ This may not be a rigorous proof but insightful argument.

The Slepian-Wolf *Source* Coding

- ❖ A *distributed* source coding theorem where two sources must be encoded separately, but decoded together at a common node

The Slepian-Wolf Source Coding (2)



- ❖ The source (X_1, X_2) drawn i.i.d. $\sim p(x_1, x_2)$.
- ❖ The achievable rate region is given by

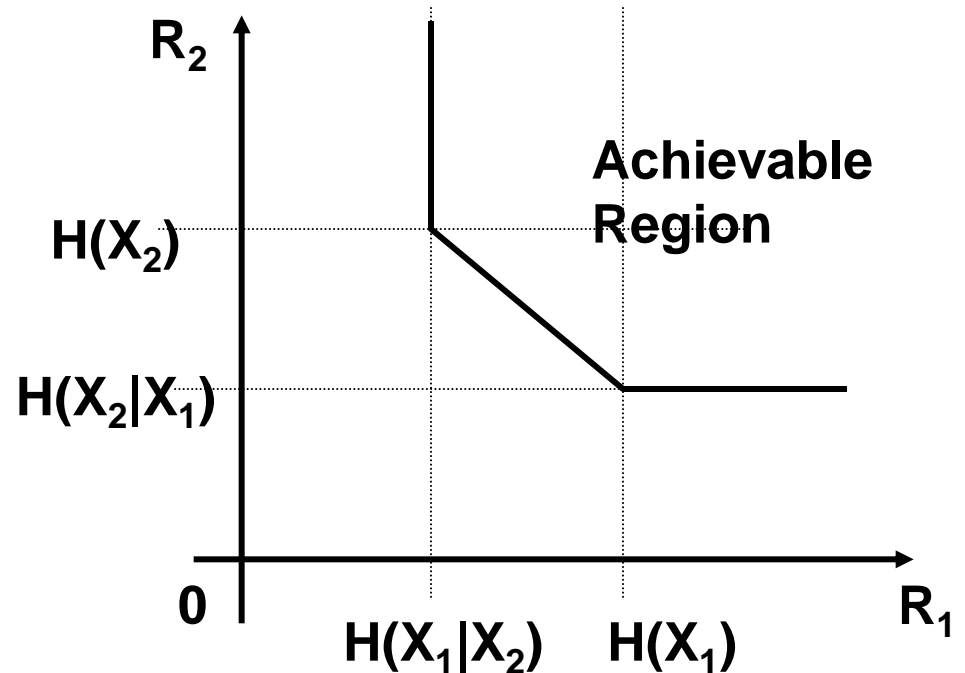
$$R_1 \geq H(X_1|X_2)$$

$$R_2 \geq H(X_2|X_1)$$

$$R_1 + R_2 \geq H(X_1, X_2)$$

Rate Regions for Slepian Wolf Encoding

- ❖ $H(X_1, X_2)$
 $= H(X_1) + H(X_2|X_1)$
 $= H(X_2) + H(X_1|X_2)$
 $\leq H(X_1) + H(X_2)$
- ❖ Equality, when?



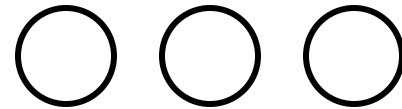
Example of Slepian-Wolf

$p(u, v)$	$u=0$	1
$v=0$	$1/3$	$1/3$
$v=1$	0	$1/3$

- ❖ $H(U, V) = \log_2(3) = 1.58$ bits
- ❖ $H(U) = -[(1/3) \log_2(1/3) + (2/3) \log_2(2/3)] = 0.9183$
- ❖ $H(V) = 0.9183$
- ❖ $H(U)+H(V) = 1.84$ bits
- ❖ You need only 1.58 bits, instead of 1.84 bits, with Slepian-Wolf encoding

Idea

- ❖ There are 3 balls and 7 different color paints—the rainbow
- ❖ Take one color at random and color the balls
- ❖ Find the probability that there are two or more balls with the same color = $1 - \Pr(\text{distinct colors})$.
- ❖ $\Pr(\text{distinct colors}) = \sum_7 (1/7)(6/7)(5/7) = (6/7)(5/7)$
- ❖ Thus, if the number of colors are sufficiently large, the chance that the probability of having an equivocation is small, especially when $n \rightarrow \infty$

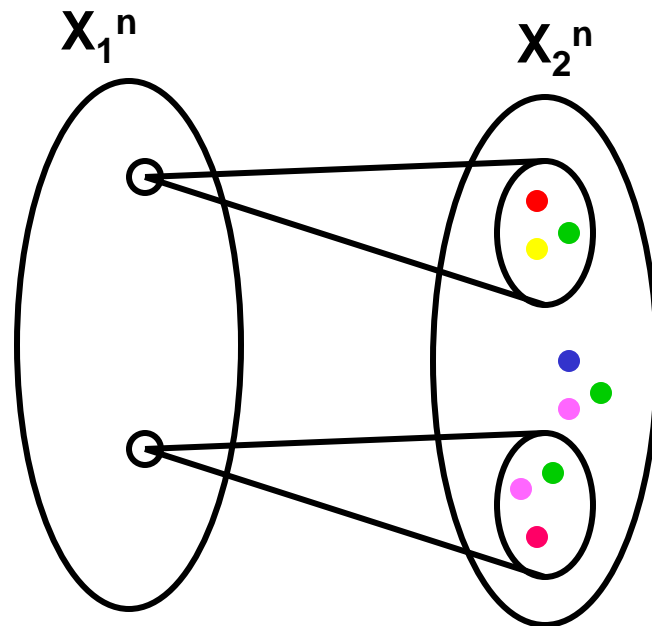


Idea (2)

- ❖ Number of balls: $B = 2^{nH(X_2|X_1)}$
- ❖ Number of colors: $C = 2^{nR}$
- ❖ $C \gg B$, $B/C = 2^{-n\epsilon}$
- ❖ The probability of non-distinctive colors in B balls = $2^{-n\epsilon}$

Idea (3)

- ❖ Encode X_1 with rate $R_1 \geq H(X_1)$ and then the decoding should be no problem.
- ❖ Now, how to encode/decode X_2 ?
- ❖ Encode X_2 with a rate $R_2 \geq H(X_2|X_1)$: randomly color every element in X_2^n with 2^{nR_2} colors.
- ❖ Then,
 - $\Pr(\text{color equivocation} \mid X_1 = i) = \frac{2^{nH(X_2|X_1)}}{2^{nR_2}} = 2^{-n\epsilon}$
 - $R_2 = H(X_2|X_1) + \epsilon$
 - Thus, we can choose a big enough n so that $n\epsilon$ be very big
- ❖ What happens then?



Slepian Wolf for Many Sources

❖ $(X_1, X_2, \dots, X_m) \sim p(x_1, \dots, x_m)$

❖ The rate region is

$$R(S) > H(X(S)|X(S^c))$$

where

- $S \subseteq \{1, 2, \dots, m\}$,
- $X(S) = \{X_j: j \in S\}$
- $R(S) = \sum_{i \in S} R_i$

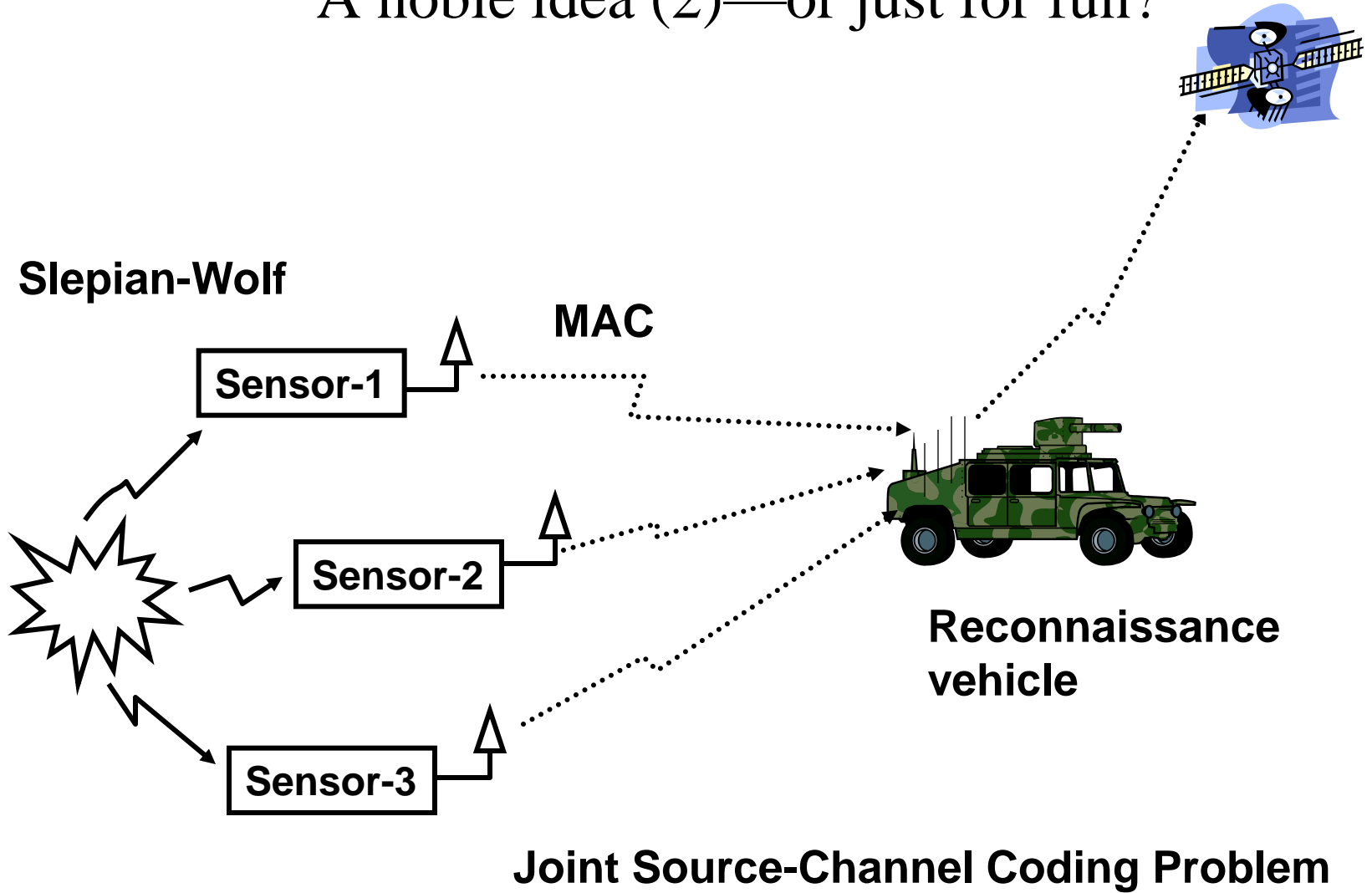
❖ Ex) 3 sources $(X_1, X_2, X_3) \sim p(x_1, x_2, x_3)$

- $R_1 > H(X_1|X_2, X_3)$ & $R_2 > H(X_2|X_1, X_3)$ & $R_3 > H(X_3|X_2, X_1)$
- $R_1+R_2 > H(X_1, X_2|X_3)$ & $R_1+R_3 > H(X_1, X_3|X_2)$ & $R_2+R_3 > H(X_2, X_3|X_1)$
- $R_1+R_2+R_3 > H(X_1, X_2, X_3)$

A noble idea—or just for fun?

- ❖ Use the powerful LDPC codes to encode the source data with a certain $H(X_1, X_2, X_3)$
 - Huffman code will not be a good idea—why?
- ❖ Choose a rate pair (R_1, R_2, R_3) according to Slepian-Wolf, and design three independent codes whose code rate pair is slightly larger, i.e., $(R_1+\epsilon, R_2+\epsilon, R_3+\epsilon)$
 - What's the ϵ for?
 - What should be the magic secret here?
- ❖ You want to consolidate the information collected by three sensors observing a common phenomenon (a tank passes by)

A noble idea (2)—or just for fun?



Broadcast Channel Capacity

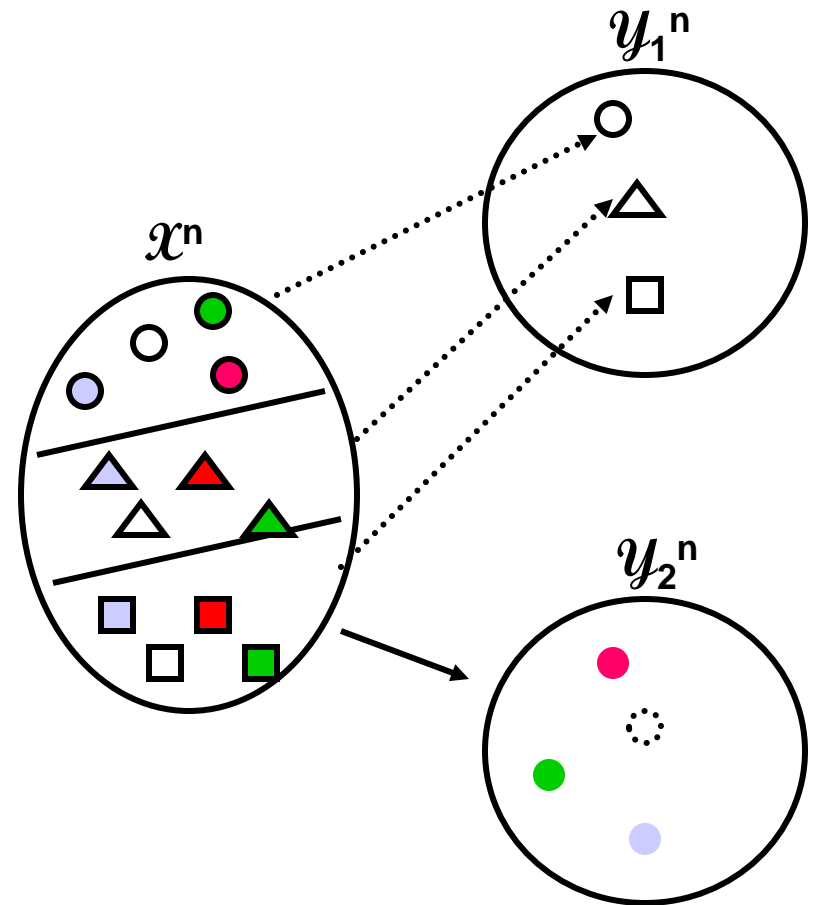
- ❖ One source, many receivers
- ❖ The channel $p(y_1, y_2|x)$ is memory-less:

$$p(y_1^n, y_2^n|x^n) = \prod_{i=1}^n p(y_1, y_2|x)$$

- ❖ Code: $((2^{nR_1}, 2^{nR_2}), n)$ code,
- ❖ $P_e^{(n)} = P(g_1(Y_1^n) \neq W_1, g_2(Y_2^n) \neq W_2)$

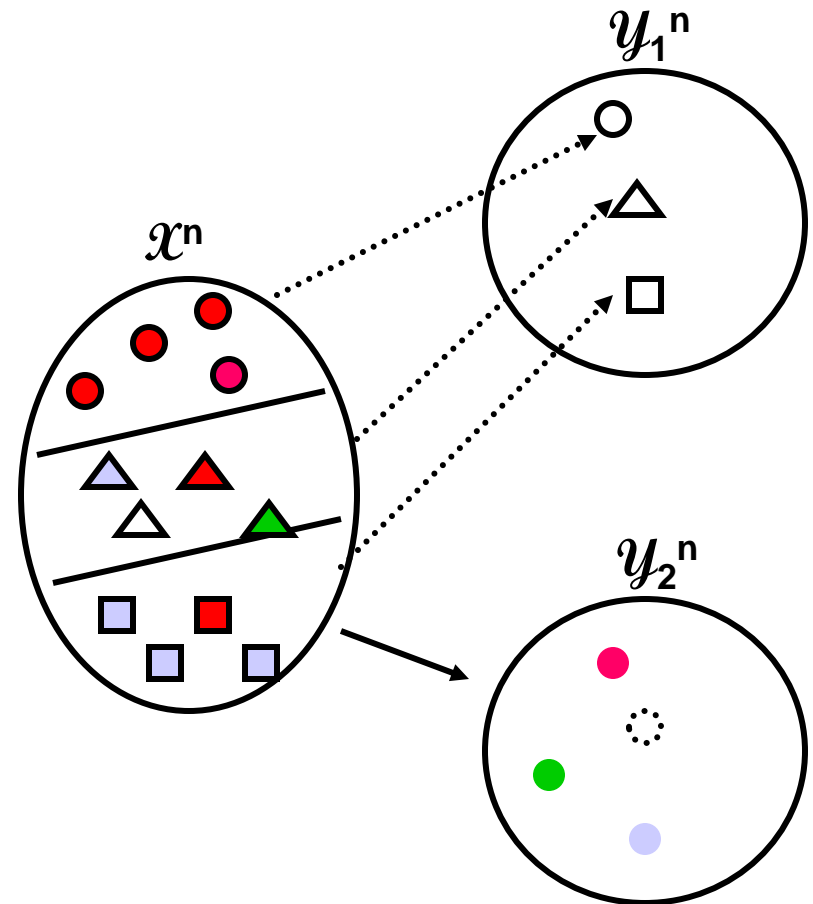
When Channel Introduces No Errors At all

- ❖ Assume $|X| > |Y_1|, |Y_2|, |Y_1| \times |Y_2|$
- ❖ $R_1 \leq H(Y_1)$
- ❖ $R_2 \leq H(Y_2)$
- ❖ $R_1 + R_2 \leq H(Y_1, Y_2)$
- ❖ No errors: $X \rightarrow (Y_1, Y_2)$
- ❖ Note that $(y_1, y_2|x) = 1$ or 0 .
- ❖ The coding scheme: For each and every typical pair (y_1^n, y_2^n) , we simply need to assign a single deterministic input x^n . As long as the rate is within the capacity region, the scheme should work
 - Ex) $R_1 + R_2 < H(Y_1, Y_2)$



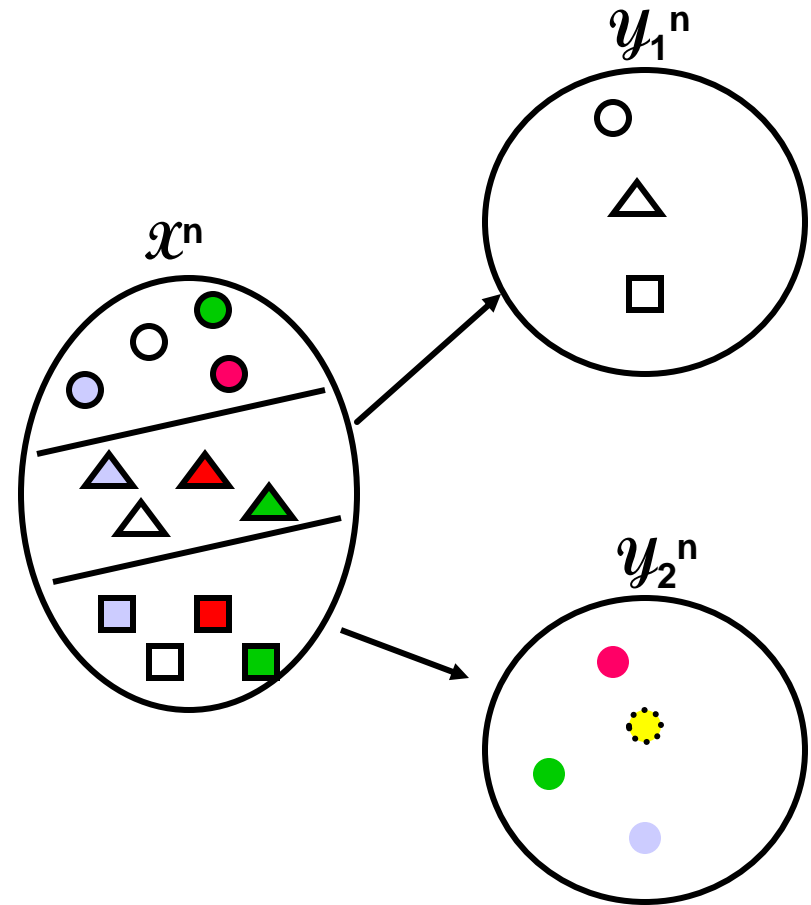
When Colors and Shapes are correlated:

- ❖ When $p(\text{red}|\text{circle}) \approx 1$ and $p(\text{blue}|\text{square}) \approx 1$
- ❖ Let U, V auxiliary r.v.s
 - U is the partition
 - V is the colors
- ❖ $I(U; V) > 0$.
- ❖ $H(U, V) = H(U) + H(V) - I(U; V) < H(U) + H(V)$ (strictly)
- ❖ We must make sure that the sum rate chosen is smaller than $H(U, V)$.



When Channel Introduces Errors

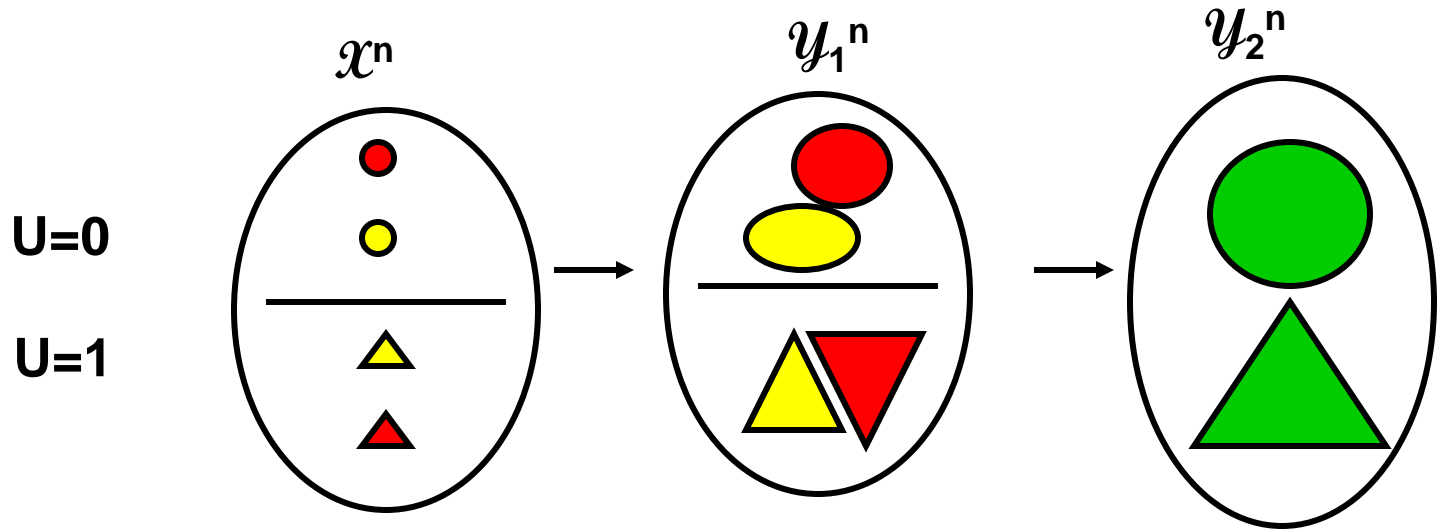
- ❖ $R_1 \leq H(U) - H(U|Y_1)$
 $= I(U; Y_1)$
 - Ex) red circle becomes red triangle.
- ❖ $R_2 \leq H(V) - H(V|Y_2)$
 $= I(V; Y_2)$
 - Ex) white becomes yellow; green sometimes becomes yellow too.
- ❖ $R_1 + R_2 \leq I(U; Y_1) + I(V; Y_2) - I(U; V)$



Degraded Broadcast Channels

- ❖ Definition: if $p(y_1, y_2|x) = p(y_1|x) p(y_2|y_1)$, the channel is said to be physically degraded.
- ❖ Assumption is that one of the channel is better than the other channel.
- ❖ Degraded broadcast channel can be understood as: $X \rightarrow Y_1 \rightarrow Y_2$
- ❖ The capacity region of this channel is
$$R_2 \leq I(U; Y_2)$$
$$R_1 \leq I(X; Y_1|U)$$
for some utility random variable U whose cardinality $|U| \leq \min\{|X|, |Y_1|, |Y_2|\}$.
- ❖ NOTE: The rate R_1 and R_2 are rates of INDEPENDENT information for user-1 and user-2 respectively.

Degraded Broadcast Channels (2)



- ❖ As long as $R_2 \leq I(U; Y_2)$, and
- ❖ As long as $R_1 \leq I(X; Y_1|U)$, we can encode the data and obtain an error-free transmission
- ❖ Note that the code $\mathcal{C}_2 (2^{nR_2}, n)$ must be decodable to both receivers; the code $\mathcal{C}_1 (2^{nR_1}, n)$ is only decodable at the receiver 1.
- ❖ Proof: Use the typical set argument again (see textbook)

Current Status of Information Theory

- ❖ Shannon's Theory has been successfully applied in practice for the past 60 years.
- ❖ The most problems Shannon posed in 1948 have been solved, especially in the point-to-point communications area.
- ❖ Tremendous amount of beautiful ideas have been accumulated which can be *applied* to various fields.
- ❖ **Channel Coding Theory (Ph.D. level, soon to be offered)**
 - Block codes, Cyclic codes, BCH codes, Reed-Solomon codes, Convolutional codes, Trellis codes, Turbo codes, Gallager codes
 - Group, Ring, Galois Field
 - Textbook: Error Correction Coding by Moon, Wiley 2005.

Networks, Networks, Networks!!!

- ❖ Today, the focus is to understand *network*.
- ❖ Application of information theory to networks
 - Network information theory
 - Network coding theory (2000 Ahlswede et. al.)
- ❖ Network is not a simple thing
 - Many sources, many relays, many sinks
 - Many traffic types (fractal traffic)
 - Delays
- ❖ Current trends: merge ideas from different fields
 - Information Theory, Convex Optimization, Queuing Theory, Control Theory, Reinforced Learning Theory, ...
 - Let's learn from nature, for example from human brain network!
- ❖ **Advanced Network Theory** (Ph.D. level, soon to be offered)

Final Exam

❖ Four problems

❖ Topics of importance

- Forward/Converse of the Coding Theorem
- Capacity for simple channels
- A simple MIMO problem
- Do you understand AEP/Jointly Typical Decoder?
- Product Distribution vs. Joint Distribution

❖ Final materials

- Every lecture note (not the materials in appendices)
- Cover & Thomas Ch1, 2, 3, 4, 5, 7, 8, 9, 12.1 , 15.3(Multiple access channel)
- All homework sets

The End