Unit 11: Information Theory and Capacity

EL-GY 6013: DIGITAL COMMUNICATIONS

PROF. SUNDEEP RANGAN

Learning Objectives

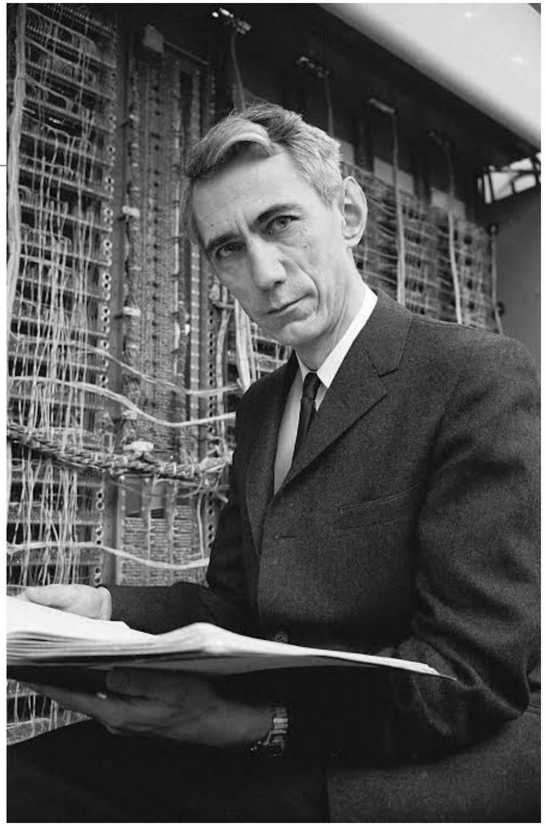
- ❑ Define and compute the Shannon capacity for simple memoryless channels
- ❑ Identify power-limited and bandwidth-limited regimes of operation
- ❑ Describe difficulties in achieving the Shannon capacity for practical systems
- ❑ Mathematically describe the performance of a system relative to the Shannon limit
- ❑ Define and compute the constellation-constrained capacity

Outline

- Information theory basics
- Shannon capacity
- Modeling capacity of practical systems
- Constellation constrained capacity
- Proof of the Shannon Theorem

What is Information Theory?

- There are many ways to design communication systems
- Two basic questions:
  - How do we measure the performance?
  - What is the best we can expect to do?
- Information theory provides:
  - Simple metrics to evaluate system performance
  - Fundamental bounds that can be achieved by *any* system
  - Apply to any communication system
  - No constraint in computation / delay
- Can be used as a benchmark for practical systems



Claude Shannon
Founder of IT

Entropy

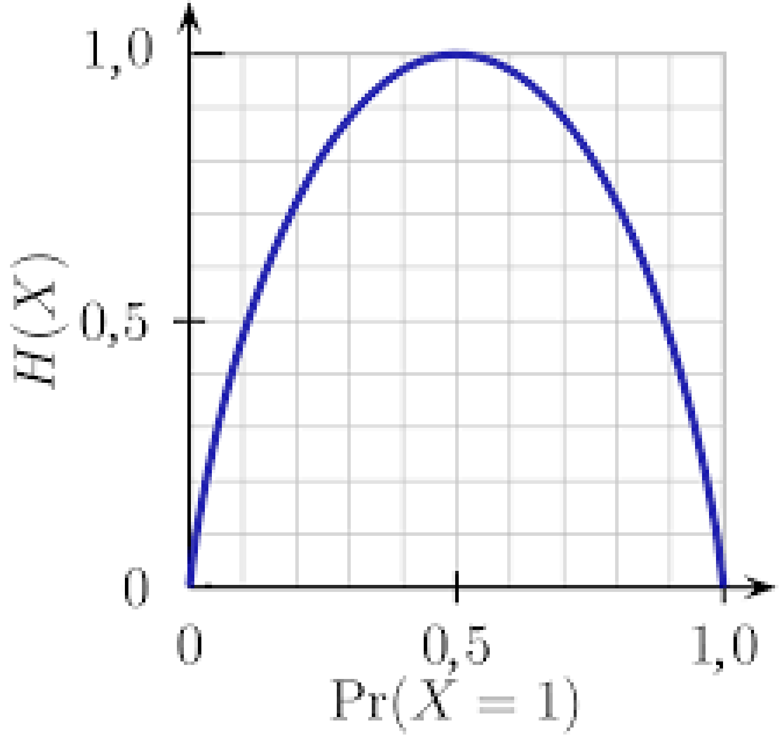
- Given a random variable X
- Entropy for a discrete X: $H(X) = -\sum p_i \log_2 p_i$
- Relative entropy for continuous X with PDF $p(x)$:

$$h(X) = -\int p(x)\log_2\big(p(x)\big)dx$$

- Measures amount of "variation" in X
  - But, unlike $var(X)$ does not depend on values of X
  - Just the number of values and their relative probability
- Sometimes measured in "nats"
  - Replace log base 2 with natural logarithm

Discrete Examples

- Ex 1: Binary
  - $P(X = 1) = 1 - P(X = 0) = p$
  - $H(X) = -p \log p - (1-p)\log(1-p)$
  - See figure to the right
  - Entropy maximized with most uncertainty, $p = 0.5$
- Ex 2: Discrete uniform
  - $X \in \{x_1, \ldots, x_N\}$ with $P(X = x_i) = \frac{1}{N}$
  - $H(X) = -\sum \frac{1}{N} \log\left(\frac{1}{N}\right) = \log(N)$
  - Entropy increases with number of values
  - Labels of the values do not matter



Continuous Examples

Distribution	Parameters	Relative Entropy in nats
Uniform	$X \sim U[a,b]$	$h(X) = \ln(b-a)$
Real Gaussian	$X \sim N(\mu, \sigma^2)$	$h(X) = \frac{1}{2}\ln(2\pi\, e\sigma^2)$
Complex Gaussian	$X \sim CN(\mu, \sigma^2)$	$h(X) = \ln(\pi\, e\sigma^2)$
Exponential	$E(X) = 1/\lambda$	$h(X) = 1 - \ln(\lambda)$

- ❑ Entropy increases with variance
- ❑ Entropy does not change with mean

Compression and Entropy

- Key interpretation of entropy

$$H(X) = \text{“number of bits to represent } X\text{”}$$

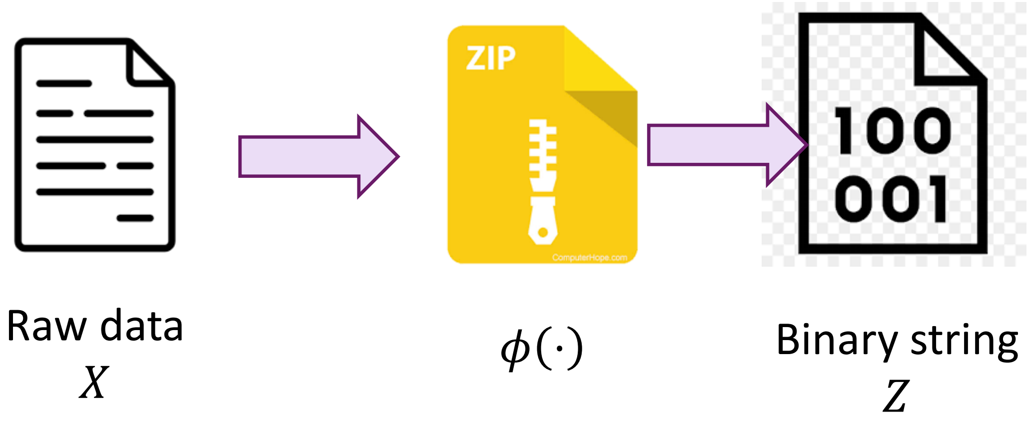
  - Related to the “compressibility” of X

- Specifically, consider variable length “encoder”:

$$Z = \phi(X)$$

  - Z is a binary string

- Want $\phi(X)$ is “prefix” free
  - $\phi(x_i)$ is not a prefix of $\phi(x_j)$ when $x_i \neq x_j$
  - Ensure mapping is invertible
  - Given sequence of outputs, we can always tell boundaries

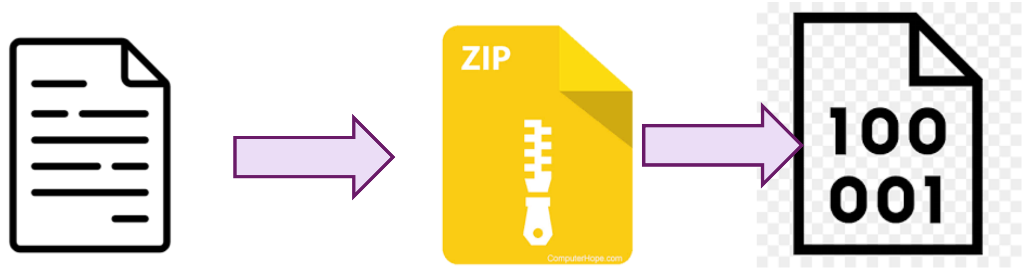


Length of an Encoder

- ❑ Given encoder $Z = \phi(X)$
- ❑ Define $L(\phi)$ =avg length of $\phi(X)$
- ❑ Ex to the right:

$$L(\phi) = 0.6(1) + 0.3(2) + 0.1(2) = 1.4 \text{ bits / sym}$$

- ❑ To minimize length:
  - Select short sequences for likely x
  - Reserve long sequences for unlikely x



Raw data X

$\phi(\cdot)$

Binary string Z

X	$P(X)$	$\phi(X)$
A	0.6	0
B	0.3	10
C	0.1	11

Compression and Entropy

- Theorem: If X is a discrete random variable, there exists a prefix free variable length code with

$$\text{Avg. length} \le H(X) + 1$$

- By encoding N symbols at a time, can achieve

$$\text{Avg. length} \le H(X) + \frac{1}{N} \to H(X)$$

- Proof uses a Huffman code
- Entropy shows how much information is in a random variable

Joint and Conditional Entropy

- Let (X, Y) be a pair of discrete random variables with a joint distribution
- Joint entropy: Entropy of the pair $Z = (X, Y)$

$$H(X, Y) = -\sum_y \sum_x P(x, y) \log P(x, y)$$

- Recall: For every y, $P(X|Y = y)$ is a distribution on X
- Conditional entropy for a given y: $H(X|Y = y) = -\sum_x P(x|y) \log P(x|y)$
  - Represents entropy in X after seeing $Y = y$
- Conditional entropy:

$$H(X|Y) \coloneqq \sum_y H(X|Y = y) = -\sum_y \sum_x P(x, y) \log P(x|y)$$

- Similar equations for continuous random variables

Properties

- Conditional: $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$
- Independence:
  - $H(X|Y) = H(X)$ if and only if X and Y are independent
  - In this case, $H(X,Y) = H(X) + H(Y)$
- For all X,Y: $H(X,Y) \leq H(X) + H(Y)$

Example

- Suppose X, Y are binary with joint PMF in table
- $H(X) = -0.5\log_2(0.5) - 0.5\log_2(0.5) = 1$
- For $Y = 0$:
  - $P(X|Y=0) = \left[\frac{2}{3}, \frac{1}{3}\right] \Rightarrow H(X|Y=0) = 0.91$
- For $Y = 1$:
  - $P(X|Y=1) = \left[\frac{1}{4}, \frac{3}{4}\right] \Rightarrow H(X|Y=1) = 0.81$
- Conditional entropy:

$$H(X|Y) = 0.6(0.91) + 0.4(0.81) \approx 0.86 \text{ bits}$$

	$Y = 0$	$Y = 1$	$P(X = x)$
$X = 0$	0.4	0.1	0.5
$X = 1$	0.2	0.3	0.5
$P(Y = y)$	0.6	0.4	

Mutual Information

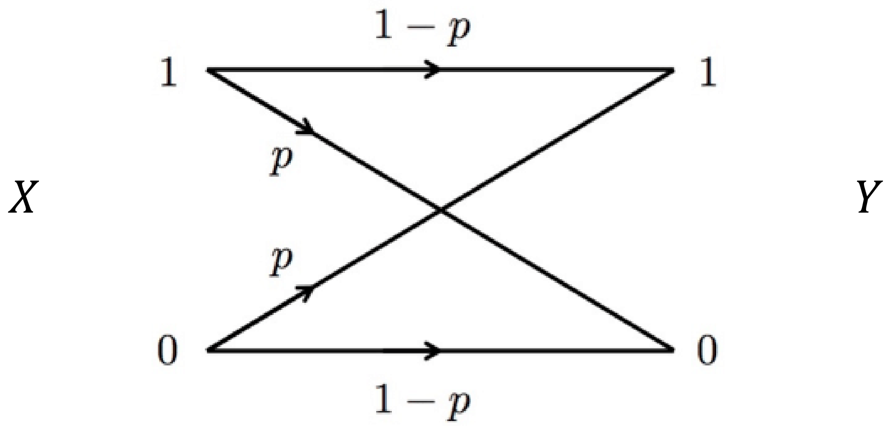
- How much are two random variables related?
- Mutual information:

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

- Represents decrease in entropy in X from knowing Y
- Can also define for differential entropy
- Special cases:
  - If X and Y are independent, $I(X;Y) = 0$
  - If $Y = f(X)$, then $I(X;Y) = H(X)$

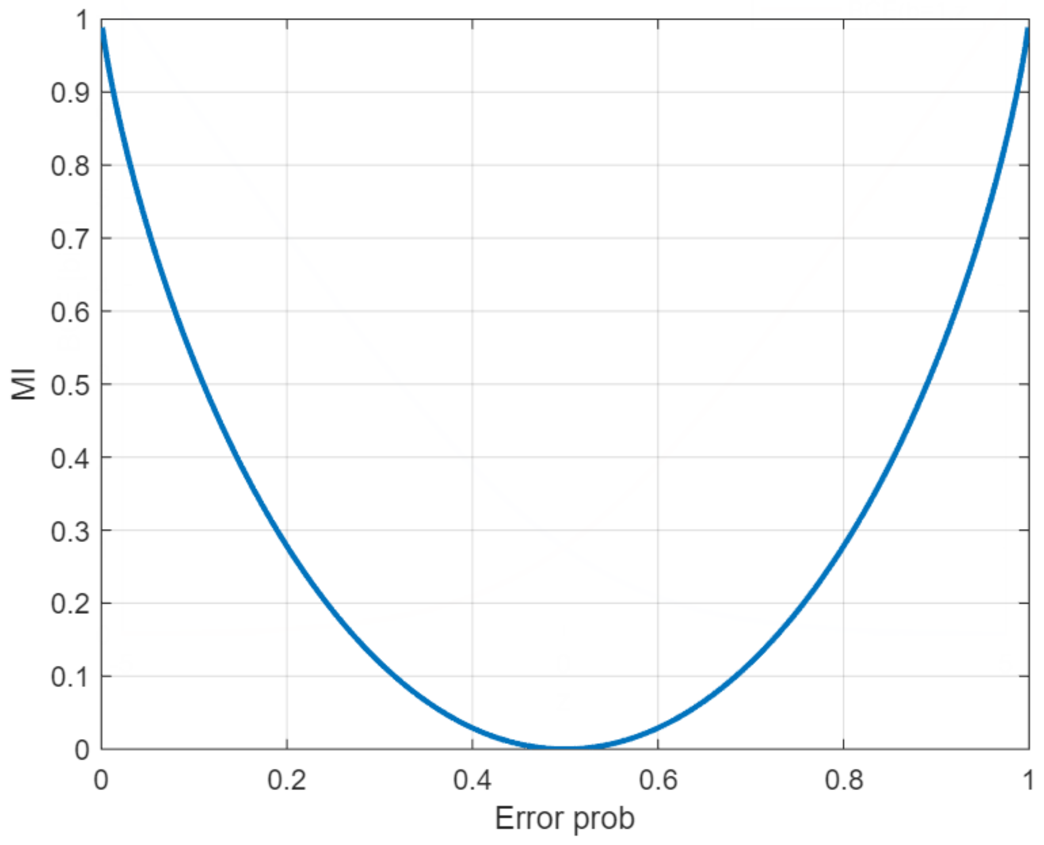
Example: BSC Channel

- For communications
  - X is the typ. the channel input and Y is the output
- Binary symmetric channel:
  - Input $X \in \{0,1\}$ equiprobable
  - Output $Y \in \{0,1\}$
  - $P(X \neq Y|X = x) = p$ = Probability of error
  - $P(X = Y|X = x) = 1 - p$ = Probability no error
- Mutual information
  - $H(X) = 1$ bit
  - $P(X|Y = 0) = [p, 1 - p]$
  - $H(X|Y = 0) = H(p) \coloneqq -p \log_2 p - (1 - p) \log_2(1 - p)$
  - Similarly, $H(X|Y = 1) = H(p)$
  - Hence: $I(X;Y) = 1 - H(p)$



BSC Channel Illustrated

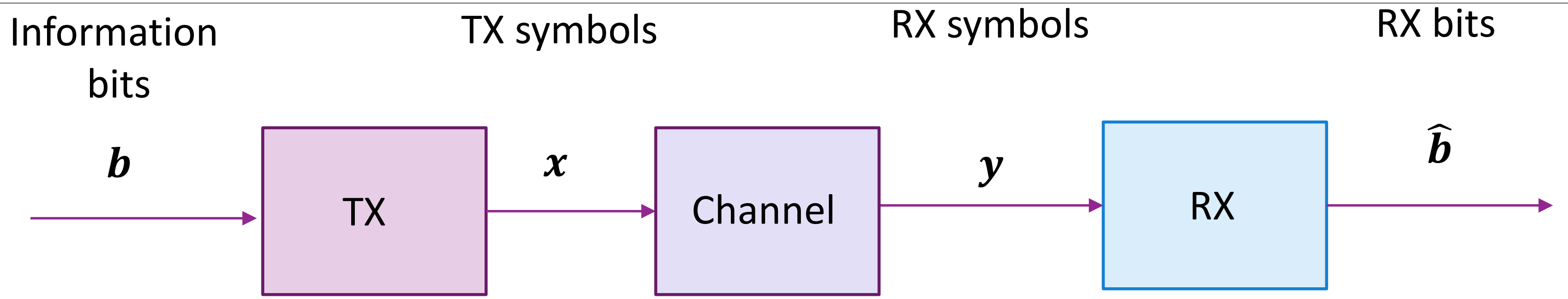
- From $I(X;Y) = 1 - H(p)$
  - $H(p) = -p\log(p) - (1-p)\log(1-p)$
- See $I(X;Y)$ vs. p on right
- When $p \to 0$ or $1 \Rightarrow I(X;Y) \to 1$
  - Y perfectly describes X
- When $p = \frac{1}{2}, I(X;Y) = 0$
  - X and Y are independent



Outline

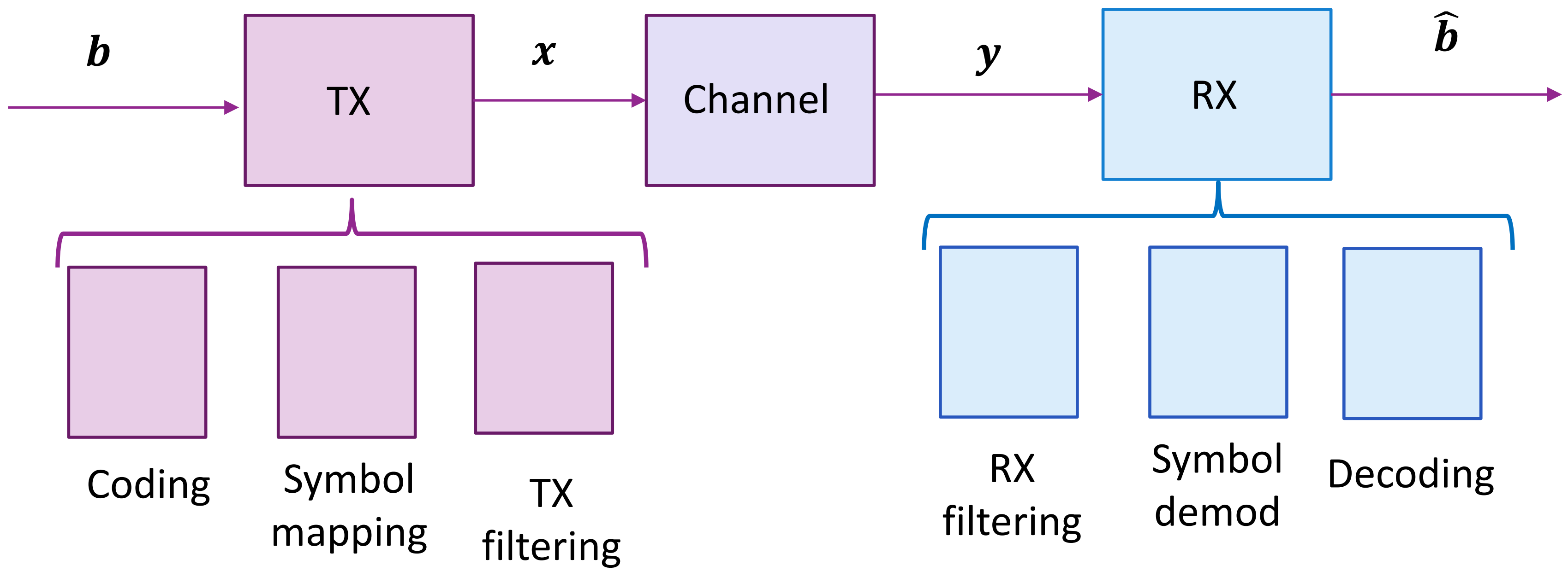
- Information theory basics
- Shannon capacity
- Modeling capacity of practical systems
- Constellation constrained capacity
- Proof of the Shannon Theorem

Abstract Communication System



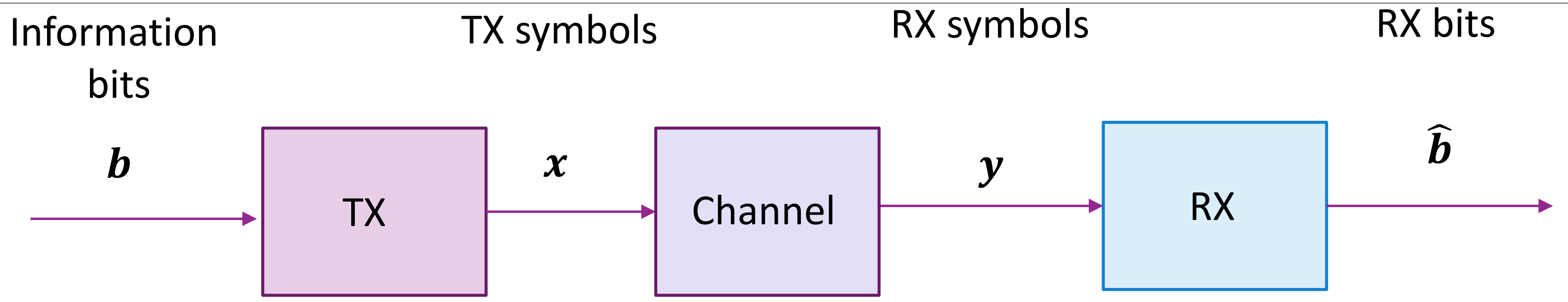
- TX k bits: $\boldsymbol{b} = (b_1, \ldots, b_k)$
- Maps bits to n symbols $\boldsymbol{x} = (x_1, \ldots, x_n)$ into "channel"
- Channel outputs n RX symbols $\boldsymbol{y} = (y_1, \ldots, y_n)$
- Channel is modeled probabilistically $P(\boldsymbol{y}|\boldsymbol{x})$
- RX attempts to estimate TX bits: $\hat{\boldsymbol{b}}$

Practical System is an Example



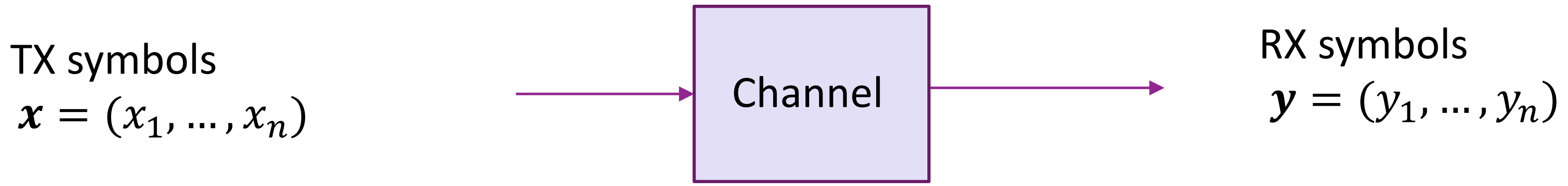
- In the abstract model, the TX and RX can include typical block we have studied up to now
- But they are not restricted to a particular structure

Key Parameters



- Block length: n = number of symbols
- Rate: $R = \frac{k}{n}$ = number of bits per symbol
- Block error rate: $P_e = P(\hat{\boldsymbol{b}} \neq \boldsymbol{b})$
  - Depends on randomness in channel
- Key goal in communication: maximize rate with a low BLER

Discrete Memoryless Channel (DMC)



- Model channel probabilistically via conditional distribution $P(\boldsymbol{y}|\boldsymbol{x})$
  - $P(\boldsymbol{y}|\boldsymbol{x}) =$ conditional distribution of the RX symbols given the TX symbols
- Say channel is memoryless if $P(\boldsymbol{x}|\boldsymbol{y}) = \prod_i P(y_i|x_i)$
  - Each RX symbol y_i depends only on x_i
- For simplicity, we restrict to the discrete case: $x_i \in \mathcal{X}, y_i \in \mathcal{Y}$
  - $\mathcal{X}, \mathcal{Y}$ are finite sets

Example Channels

TX symbols $\boldsymbol{x} = (x_1, \dots, x_n)$ → Channel → RX symbols $\boldsymbol{y} = (y_1, \dots, y_n)$

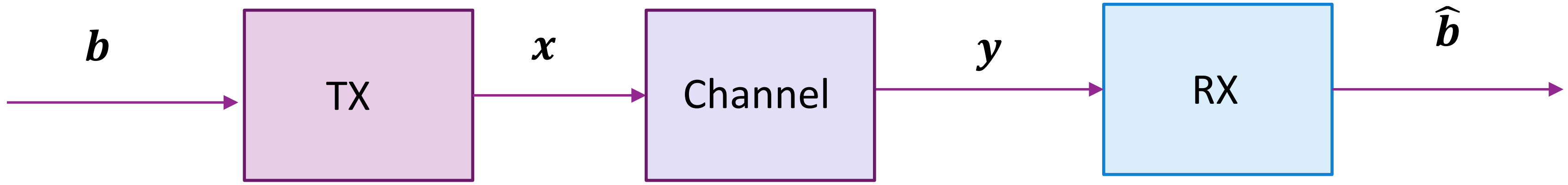
- Example 1: AWGN channel is memoryless

$$y_i = x_i + w_i, \qquad w_i \sim CN(0, N_0)$$

  - Assume w_i are independent

- Example 2: BSC channel is memoryless and discrete
  - TX and RX symbols are binary $y_i, x_i \in \{0,1\}$
  - BSC channel is independent on each symbol

Asymptotic Rate and Reliability



- To obtain sharp results, we often look at the case of long block lengths
- Formally, consider a sequence of TX-RX pairs as a function of the block length n
- For each n:
  - $k = k(n) =$ number of information bits
  - TX is some function: $(x_1, \ldots, x_n) = f_n(b_1, \ldots, b_k)$
  - RX is some function: $\left(\hat{b}_1, \ldots, \hat{b}_k\right) = g_n(y_1, \ldots, y_n)$
- Asymptotic rate: $R = \lim_{n\to\infty} \frac{k}{n}$
- Say it is asymptotically reliable if: $\lim_{n\to\infty} P_e = 0$

Achievable Rate and Capacity

- ❑ Achievable rate: We say a rate R is achievable if:
  - There exists a sequence of encoder-decoders indexed by block length n with rate R, and
  - The BLER vanishes: $\lim_{n\to\infty} P_e = 0$
- ❑ Capacity: Is the supremum over all achievable rates R
  - Optimized over all possible encoders & decoders
  - No regard to complexity or delay

Shannon's Capacity Theorem

- Theorem: Given a DMC with transition $P(y|x)$, the channel capacity is:

$$\boxed{C = \max_{p(x)} I(X;Y)}$$

  - We sketch the proof at the end of the lecture

- Maximization is performed over distributions $p(x)$
- With $p(x)$ and $p(y|x)$, we can compute $I(Y|X)$

Example: BSC

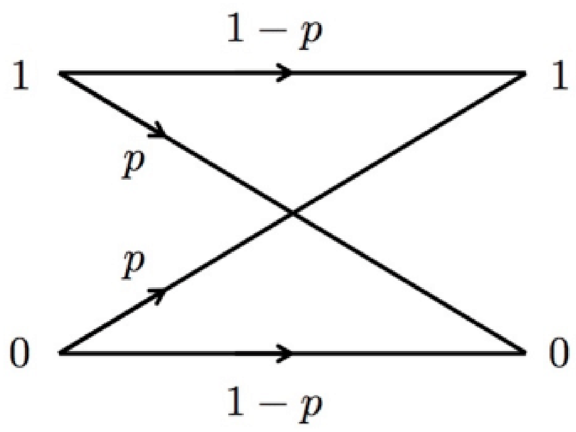
- Input $X \in \{0,1\}$, output $Y \in \{0,1\}$
- Probability of error: $p = P(X \neq Y)$
- Can show that maximizing distribution is:

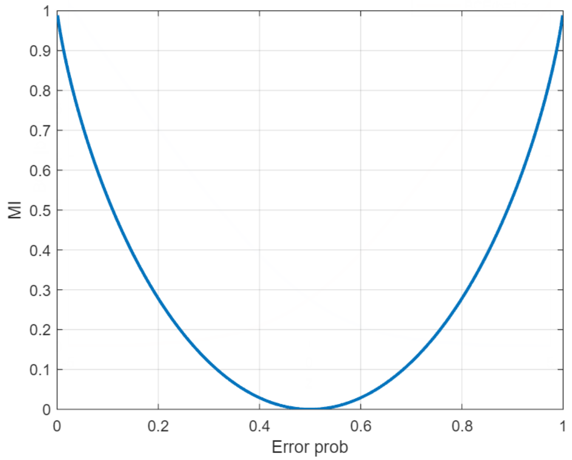
$$P(X = 0) = P(X = 1) = \frac{1}{2}$$

- In this case, the mutual information is computed as before

$$C = I(X;Y) = 1 - H(p)$$

  - Capacity $C \in [0,1]$ with higher capacity as $p \to 1$





AWGN Channel Capacity

- Now suppose that $y = x + w, \quad w \sim CN(0, N_0)$
- Although this channel is not discrete, similar theory applies using relative entropy
- Limit input distributions such that $E|x|^2 \leq E_x$ where E_x is a maximum energy per symbol
- Theorem: The capacity of the AWGN channel with energy limit E_x is:

$$\boxed{C = \log_2(1+\gamma), \qquad \gamma = \frac{E_x}{N_0}}$$

- Simple relation relating capacity to SNR γ

Proof of AWGN Channel Capacity

- AWGN channel: $y = x + w, \quad w \sim CN(0, N_0)$
- First suppose that $x \sim CN(0, E_x)$, a Gaussian input
- Entropy of complex Gaussian, $z \sim CN(\mu, \sigma^2)$ is $h(z) = \log(\pi e\, \sigma^2)$
- Therefore
  - $p(y) = CN(0, E_x + N_0) \Rightarrow h(y) = \log_2(\pi e(E_x + N_0))$
  - Given $x,\ p(y|x) = CN(x, N_0) \Rightarrow h(y|x) = \log_2(\pi e N_0)$
- Hence $I(x; y) = h(y) - h(y|x) = \log_2(\pi e(E_x + N_0)) - \log_2(\pi e N_0) = \log_2(1 + \frac{E_x}{N_0})$
  - Therefore, Gaussian input achieves the capacity
- Can also show that for any distribution with $E|x|^2 \le E_x,\ h(y) \le \log_2(\pi e(E_x + N_0))$
  - Hence, any other distribution has lower $I(x; y)$

Continuous Time Capacity

- Consider continuous-time system: $y(t) = x(t) + w(t)$
  - Assume $E|x(t)|^2 \leq P_x$ and $x(t)$ is bandlimited to bandwidth B
  - Noise $w(t)$ is AWGN with PSD N_0
- Theorem: The capacity of the continuous-time AWGN system is:

$$\boxed{C = B \log_2(1+\gamma), \qquad \gamma = \frac{P_x}{BN_0}}$$

  - Most important formula in IT!
  - Relates SNR, bandwidth and achievable rate
  - Proof sketch in next slide

Proof of Continuous-Time Capacity

- We convert the continuous-time channel to a discrete-time channel
- If $x(t)$ is band-limited to B, then there are B degrees of freedom per second
- So, we can find an orthonormal basis:

$$x(t) = \sum_k x_k \, \phi(t - nT), \qquad T = \frac{1}{B}$$

  - The energy per symbol will be: $E_x = \frac{P_x}{B}$
- We can similarly write the received signal as $y(t) = \sum_k y_k \, \phi(t - nT)$ where

$$y_k = x_k + w_k$$

- Noise energy per symbol is $E|w_k|^2 = N_0$
- Capacity per symbol is $C_0 = \log_2(1 + \frac{E_x}{N_0}) = \log_2(1 + \frac{P_x}{BN_0})$
- Since there are B symbols / sec, the continuous-time capacity is $C = B \log_2(1 + \frac{P_x}{BN_0})$

Example

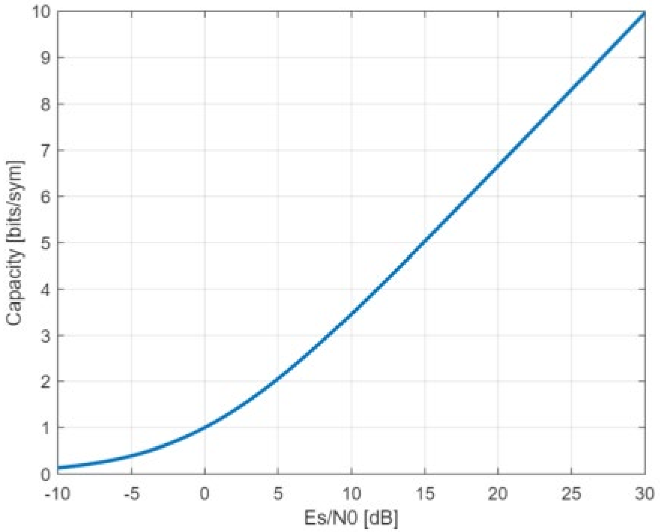
- Suppose:
  - TX power, $P_{tx} = 20$ dBm
  - Path loss, $L = 110$ dB
  - Bandwidth, $B = 20$ MHz
  - Noise density (with noise figure) is $N_0 = -170$ dBm/Hz
- Capacity:
  - RX power, $P_{rx} = 20 - 110 = -90$ dBm
  - SNR is $\gamma = P_{rx} - 10\log_{10}(B) - N_0 = -90 - 73 - (-170) = 7$ dB
  - In linear scale: $\gamma = 10^{0.7} \approx 5.0$
  - Spectral efficiency is $\rho = \log_2(1+\gamma) = 2.59$ bps/Hz
  - Capacity is $C = B\log_2(1+\gamma) = 20(2.59) \approx 51.7$ Mbps

Regimes

- Two regimes

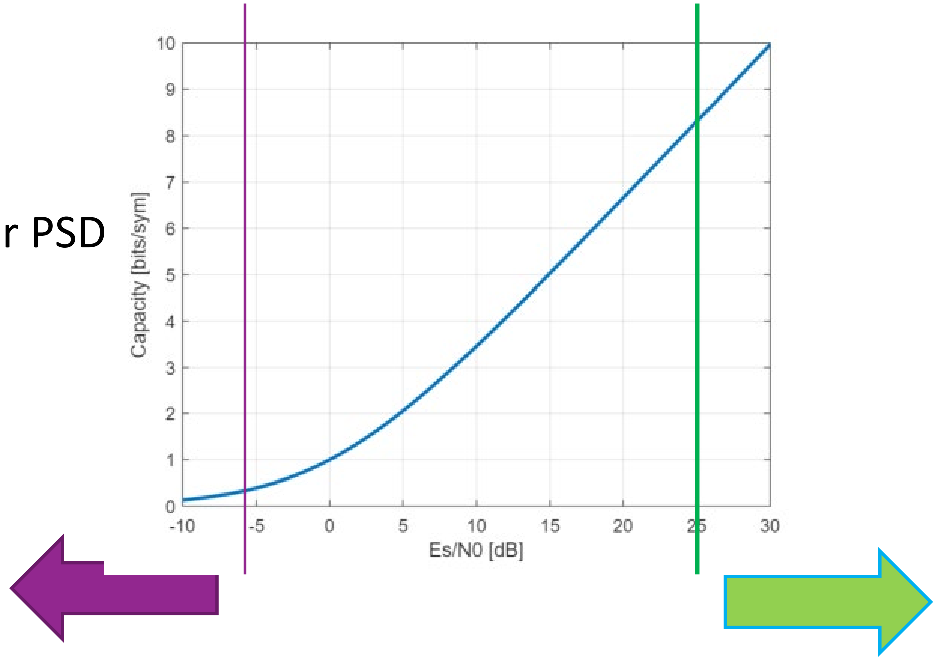
- Power limited regime
  - Suppose SNR $\gamma = \frac{P_x}{BN_0}$ is low
  - $C = B\log_2(1 + \frac{P_x}{BN_0}) \approx \frac{1}{\log(2)}\frac{P_x}{N_0}$
  - Capacity is linear in power
  - Bandwidth does not help

- Bandwidth limited regime
  - Suppose SNR is high
  - $C \approx B\log_2(\gamma)$
  - Capacity is only logarithmic in SNR. SNR does not help much
  - But grows much faster with bandwidth



Practical Design Guidelines

- Practical systems operate in a limited SNR range
- Avoid very power limited regime
  - Generally, keep $\gamma \geq -6$ dB
  - Below this SNR, better to use smaller bandwidth and higher PSD
  - Reduces overhead and computation
- Avoid highly bandwidth limited regime
  - Generally, keep $\gamma \leq 25$ to 30 dB
  - Gains are very low with higher SNR
  - Also, the gains are hard to achieve in practice
  - In these cases, use more bandwidth



SNR Per Bit and Spectral Efficiency

- Shannon formula: $C = B\log_2(1+\gamma_s)$, $\gamma_s = \frac{P_x}{BN_0}$
- Spectral efficiency: $\rho = \frac{C}{B} = \log_2(1+\gamma_s)$
  - Units are bits per second / Hz
  - Represents rate / bandwidth
- SNR per bit:

$$\gamma_b = \frac{P_{rx}}{N_0 C} = \frac{\gamma_s}{\rho}$$

  - Written as $\gamma_b = \frac{E_b}{N_0}$
  - Pronounced “Ebb-noh”

Outline

- ❑Information theory basics
- ❑Shannon capacity
- ❑Modeling capacity of practical systems
- ❑Constellation constrained capacity
- ❑Proof of the Shannon Theorem

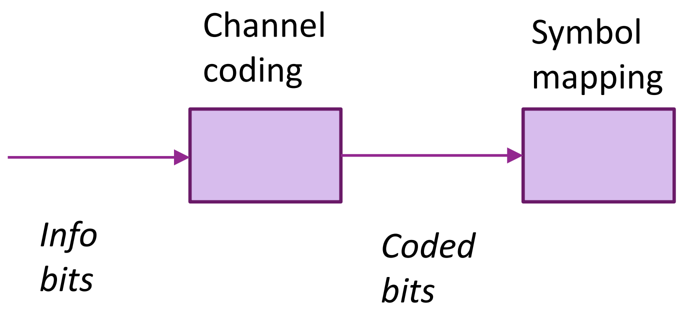
Problems Achieving Shannon Capacity

- Shannon's capacity formula is impossible to exactly achieve in practice
- Achieving the capacity requires generating a "random codebook":
- Codebook requires $M = 2^{Rn}$ entries
- Grows exponentially with block length $\Rightarrow$ Prohibitive computation and memory
- Also, $n \to \infty$ introduces infinite delay

How close can we get to Shannon capacity in practice?

Modulation and Coding Schemes

- Practical systems use a modulation and coding scheme (MCS)
- Coding:
  - Ex: Convolutional, Turbo, …
  - Defined by rate $R_{cod} < 1$
- Modulation via symbol mapping
  - Typically, M QAM
  - Defined by bits / sym, $R_{mod} = \log_2(M)$
- Spectral efficiency is: $\rho = R_{cod} R_{mod}$
- Ex: 16-QAM with a Rate ¾ code
  - $R_{mod} = 4,\ \mathrm{R}_{cod} = 0.75 \Rightarrow \rho = 0.75(4) = 3$ bps/Hz



Measuring Gap to Shannon Capacity

- Each MCS has a spectral efficiency (SE): $\rho = R_{cod} R_{mod}$
- By Shannon Theory, we should achieve this SE at an SNR $\rho = \log_2(1+\gamma_s)$
- Practical codes obtain a lower SE

$$\rho = \log_2(1+\beta\,\gamma_s), \qquad \beta < 1$$

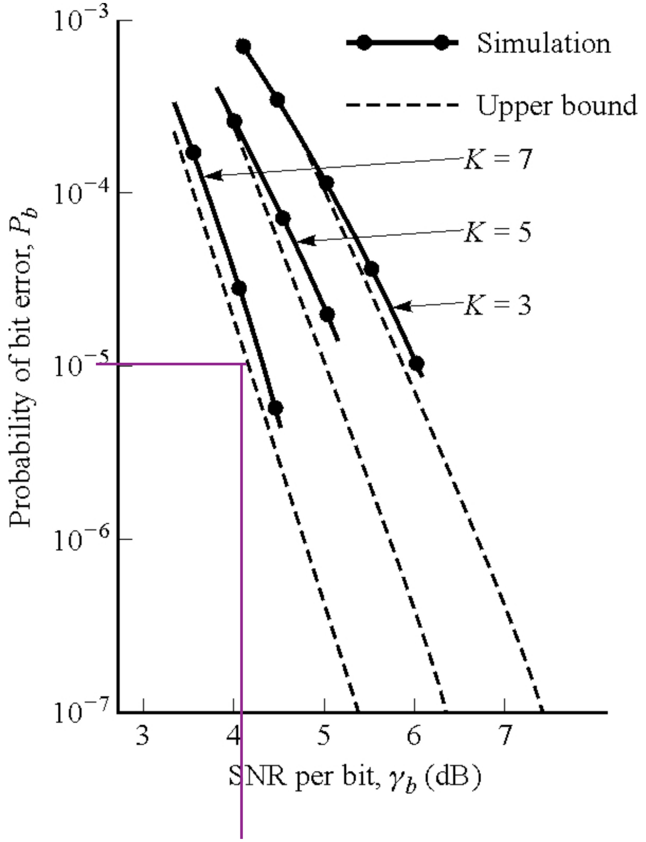
- We system operates β below Shannon capacity
  - Often quoted in dB: $10\log_{10}(\beta)$
- Gap depends on the level of reliability (e.g., BLER) and implementation

Example

- Rate $R_{cod} = \frac{1}{2}$ convolutional code with QPSK $R_{mod} = 2$
- Spectral efficiency achieved is:

$$\rho = R_{cod} R_{mod} = \frac{1}{2}(2) = 1$$

- SNR required for BER$= 10^{-5}$ is $\gamma_b \approx 4.1$ dB
  - See simulation to the right
- Shannon theory: $\rho = \log(1 + \gamma_s) \Rightarrow \gamma_s = 2^\rho - 1$
  - For $\rho = 1 \Rightarrow \gamma_s = 1$ in linear scale
  - SNR per bit is $\gamma_b = \frac{\gamma_s}{\rho} = 1$ in linear scale, $\gamma_b = 0$ dB
- Hence, we say this system operates 4.1 dB below Shannon



Capacity and Bandwidth Loss

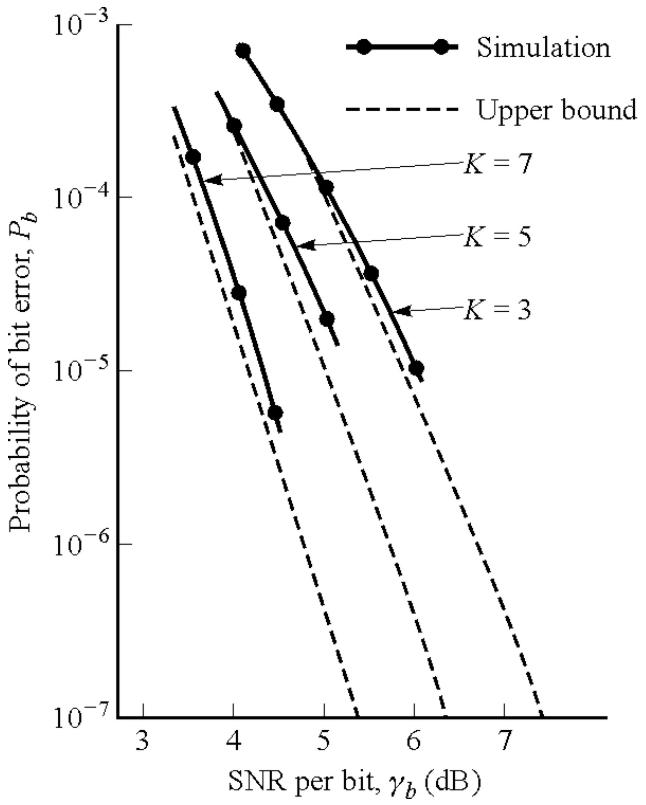
- Most systems have loss to imperfect codes and bandwidth overhead
- Simple model for achievable rate:

$$R = (1-\alpha)B \min\{\rho_{max}, \log_2(1+\beta\,\gamma)\}$$

  - α = fraction bandwidth overhead
  - β = power loss
  - ρ_{max} = maximum spectral efficiency (due to max MCS)
- Example:
  - System operates 6 dB below capacity with a 20% bandwidth overhead and $\rho_{max} = 5$ bps/Hz
  - Bandwidth B =20 MHz
  - Suppose $\gamma = 10$ dB. In linear scale, $\beta\gamma = 10^{0.1(10-6)} = 2.5$
  - Rate is: $R = (0.8)(20)\log_2(1+2.5) = 29$ Mbps
  - Shannon rate is $C = (20)\log_2(1+10^{0.1(10)}) \approx 69$ Mbps

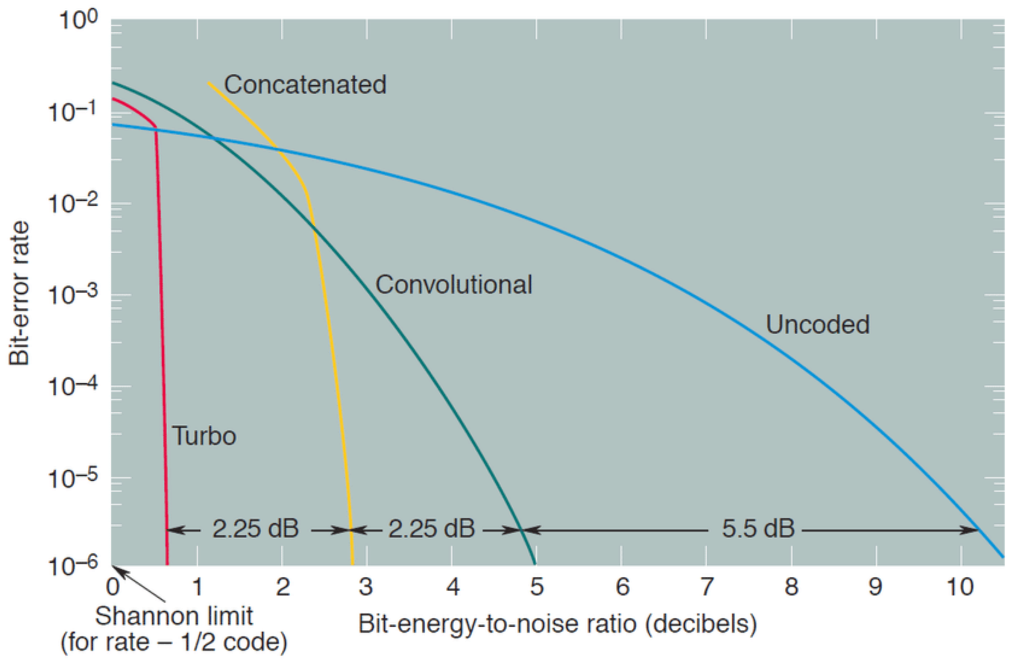
Gaps to Shannon Theory for Early Codes

- Shannon capacity formula and random codes, 1948.
  - Determines the capacity
  - But no practical code to achieve it.
- Hamming (7,4) code, 1950
- Reed-Solomon codes via polynomials over finite fields:
  - Invented in 1960 at MIT Lincoln Labs
  - Berlekamp-Massey decoding algorithm, 1969.
  - Used in Voyager program, 1977. CD players, 1982.
- Convolutional codes.
  - Viterbi algorithm, 1969. Widely used in cellular systems. (Viterbi later invents CDMA and founds Qualcomm)
  - Typically, within 4-5 dB of capacity



Improvements with Modern Codes

- 1990s: major breakthrough via graphical models
- Turbo codes (next class)
  - Berrou, Glavieux, Thitimajshima, 1993.
  - Able to achieve capacity within a fraction of dB.
  - Adopted as standard in all cellular systems by the late 1990s.
- LDPC codes
  - Similar iterative technique as turbo codes.
  - Re-discovered in 1996
  - Used in 5G today
  - Can provably hit Shannon capacity using graphs with coupling, Richardson & Urbanke, 2012



Outline

- Information theory basics
- Shannon capacity
- Modeling capacity of practical systems
- Constellation-constrained capacity
- Proof of the Shannon Theorem

Loss from Finite Constellations

- Consider AWGN channel: $y_i = x_i + w_i, \quad w_i \sim CN(0, N_0)$
- Theoretically optimal codebook is Gaussian
- But, in practice, we use M-QAM or some discrete constellation for ease
- Constellation-constrained capacity: Capacity given that x_i must be in some given constellation
- This section, we will show:
  - How to define a constellation-constrained capacity
  - How to compute a constellation-constrained capacity
  - How to account for loss for sub-optimal bitwise decoding

Capacity-Constrained Capacity Defined

- AWGN channel: $R = S + W,\ w \sim CN(0, N_0)$
- With only constraint that $E|S|^2 \leq E_s$, capacity is:

$$C = \max_{p(s)} I(S;R) = \log_2\left(1 + \frac{E_s}{N_0}\right)$$

  - Optimal distribution $p(s)$ is complex Gaussian
- Now consider fixed constellation: $S \in \mathcal{A} = \{s_1, \ldots, s_M\}$ with equiprobable symbols
  - $\mathcal{A}$ is the constellation (ex. M-QAM)
- Define constellation-constrained capacity:

$$C_{\mathcal{A}} = I(S;R)$$

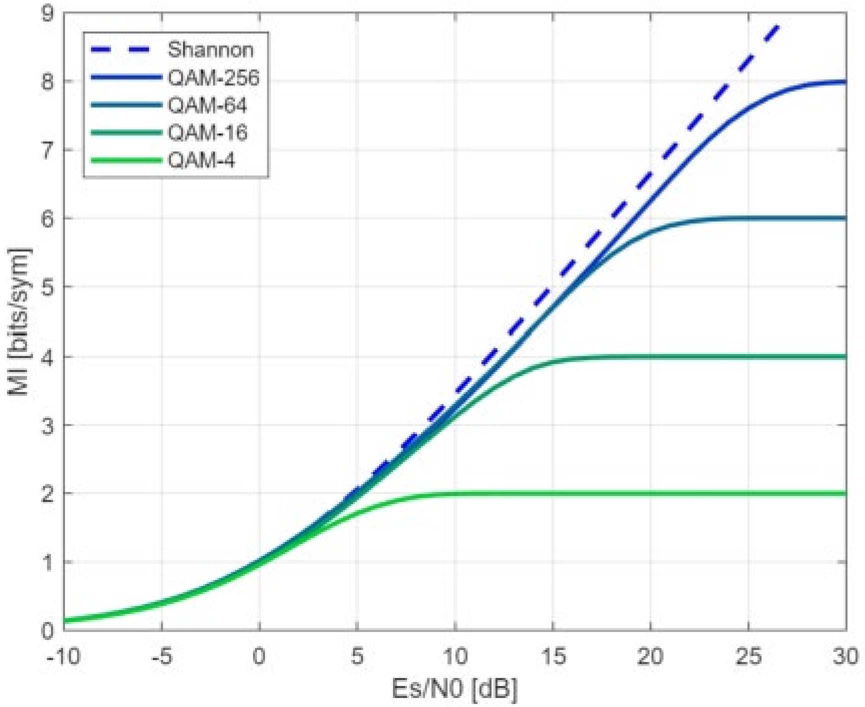
  - Capacity for a fixed constellation

Computing Capacity-Constrained Constellation

- AWGN channel: $R = S + W,\ w \sim CN(0, N_0)$ with $S \in \mathcal{A} = \{s_1, \dots, s_M\}$
- Mutual $I(R; S)$ can be computed numerically or via simulation easily:
  - $I(R; S) = H(S) - H(S|R)$
  - Since S is equiprobable, $H(S) = \log_2(M) =$ number of bits / symbol
  - Given $\mathrm{S} = s,\ r$ is Gaussian: $p(r|s) = Ce^{-|r-s|^2/N_0}$
  - Hence, by Bayes Rule: $P(S = s_i | R = r) = \frac{1}{Z(r)} e^{-|r-s_i|^2/N_0},\ Z(r) = \sum_j e^{-|r-s_j|^2/N_0}$
  - Therefore, $H(S|R = r) = -\sum_i P(s = s_i|r) \log_2 P(s = s_i|r)$
  - Find $I(R; S) = \log_2(M) - E[H(S|R = r)]$
  - Generate N random pairs $(r_n, s_n), n = 1, \dots, N$ and obtain estimate

$$\mathrm{C}_{\mathcal{A}} = I(R; S) = \log_2(M) - \frac{1}{N} \sum_n H(S|R = r_n)$$

Constellation-Constrained Capacity



Key insights:

- Capacity with M −QAM saturates
  - $C_{\mathcal{A}} \leq \log_2(M)$
- Hence, high SNR requires large M
- Relative to Shannon Capacity
  - Minimal loss at low SNRs (< 2 dB)
  - Loss of 1-2 dB at high SNRs

Bitwise LLRs

- AWGN channel: $r = s + w,\ s \in \{s_1, \dots, s_M\}$
- Up to now, we assume we decode each symbol
  - Requires we find a PMF $P(s = s_m | r),\ \ m = 1, \dots, M$
  - Finding this PMF is computationally expensive since $M = 2^K, K =$ number of bits / symbol
  - Also, most decoders requires probabilities on bits not symbols
- Practical systems decode each bit
  - Suppose $s = \phi(c_1, \dots, c_K)$ a mapping from K bits to the symbol
  - We then compute the bitwise LLR: $z_k = \log \frac{P(c_k=1|r)}{P(c_k=0|r)}$
  - This method is computationally simpler.
  - But it is not optimal
- What is the loss in capacity with bitwise LLRs?

Binary Cross Entropy

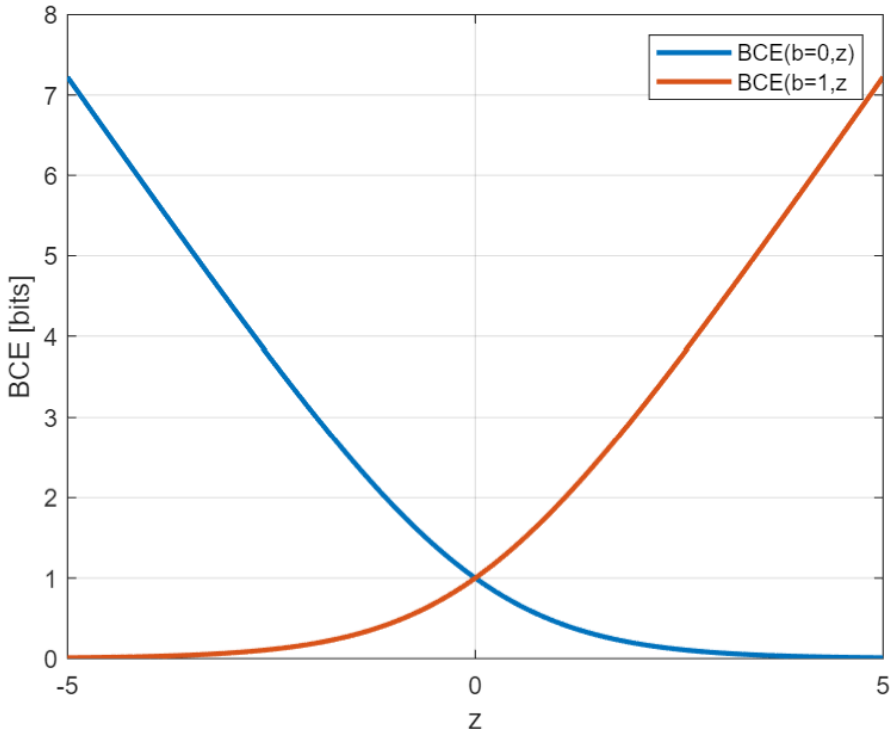
- Let $b \in \{0,1\}$: unknown binary variable
- Let $z \in \mathbb{R}$: Estimate of the LLR

$$z \approx \log \frac{P(b=1|z)}{P(b=0|z)}$$

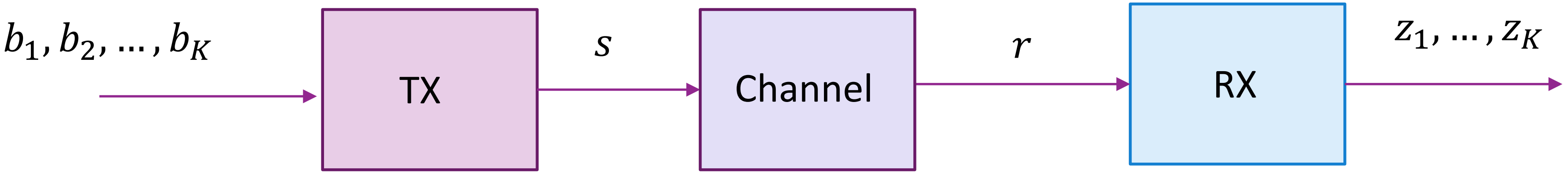
- Define binary cross entropy

$$BCE(b,z) := \frac{1}{\ln(2)}\left[\ln(1+e^{z}) - zb\right]$$

- Measure of error: Large when:
  - $b = 0$ and z large positive or
  - $b = 1$ and z large negative
- Commonly used in training binary classifiers
  - See ML class



BCE Mutual Information Bound

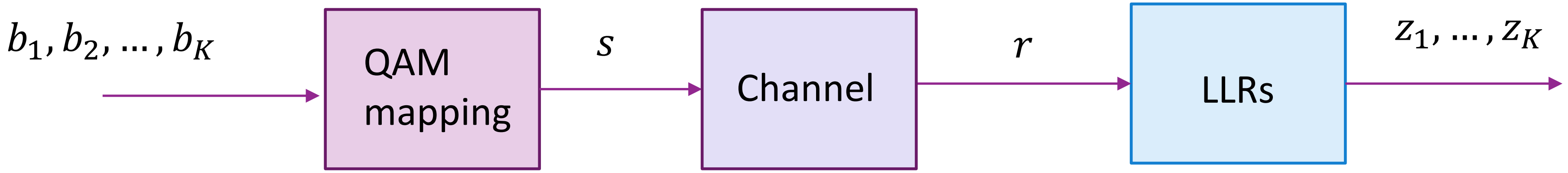


- We derive a bound for a general binary input channel
- TX: K bit binary input $\boldsymbol{b} = (b_1, \dots, b_K)$ bits and maps to a symbol vector $\boldsymbol{s}$
- RX: Obtains any output $\boldsymbol{r}$ and creates any vector $\boldsymbol{z} = (z_1, \dots, z_K)$
  - Values z_k can be the LLRs or any approximation of the LLRs of the bits
- Theorem: The mutual information is bounded as:

$$I(\boldsymbol{b};\boldsymbol{r}) \geq H(\boldsymbol{b}) - \sum_{k=1}^{K} E[BCE(b_k, z_k)] \quad [bits]$$

  - Proven at end of section

LLR Mutual Information Bound



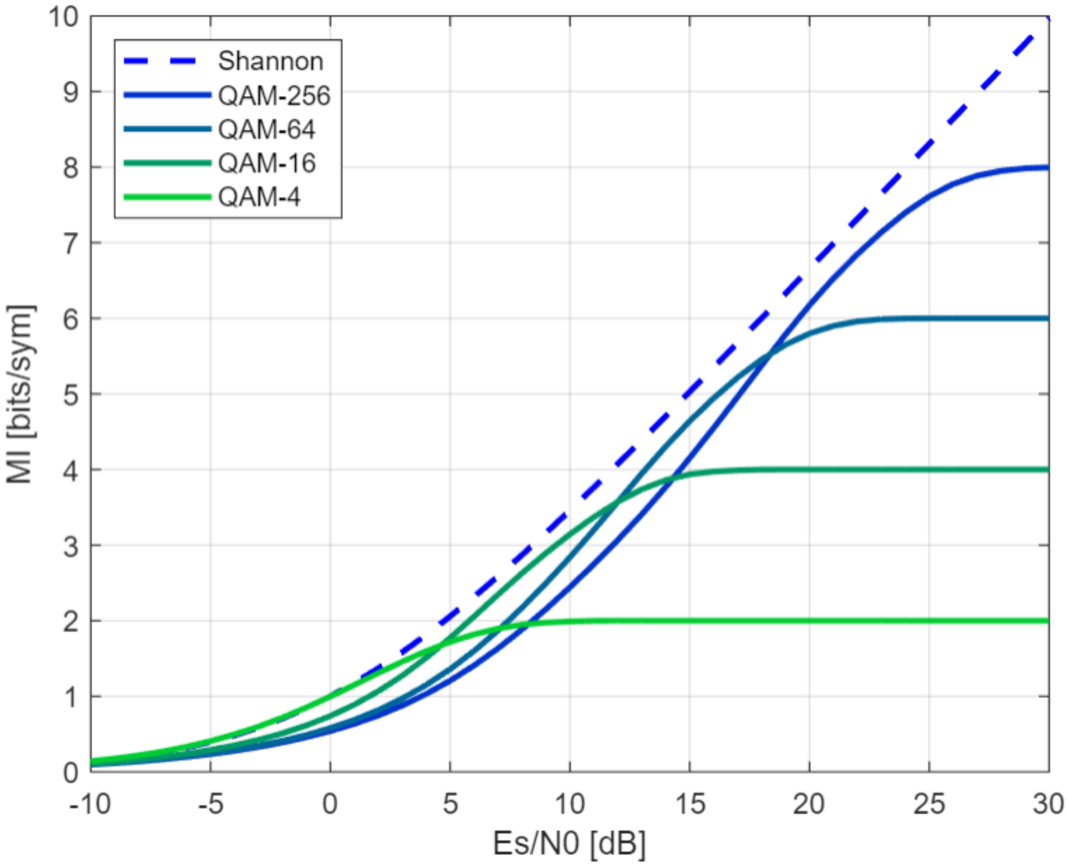
- BCE bound can be used to find capacity with practical symbol demodulation
- TX: Takes $b = (b_1, \ldots, b_K)$ bits and creates QAM symbol s with energy $E_s = E|s|^2$
- Channel is $r = s + w, \quad w \sim CN(0, N_0)$
- RX performs demodulation and creates LLRs $z = (z_1, \ldots, z_K)$

QAM Capacity with Bitwise LLRs

- Can compute the bound easily
- Generate N bits $b_1, \ldots, b_N$ over S symbols
- Modulate to $s_1, \ldots, s_P$ symbols
- Add noise and get $r_1, \ldots, r_P$ RX symbols
- Compute N LLRs $z_1, \ldots, z_N$
- Compute MI:

$$I(b; r) \geq \frac{1}{P} \sum_{i=1}^{N} [1 - BCE(b_i, z_i)]$$

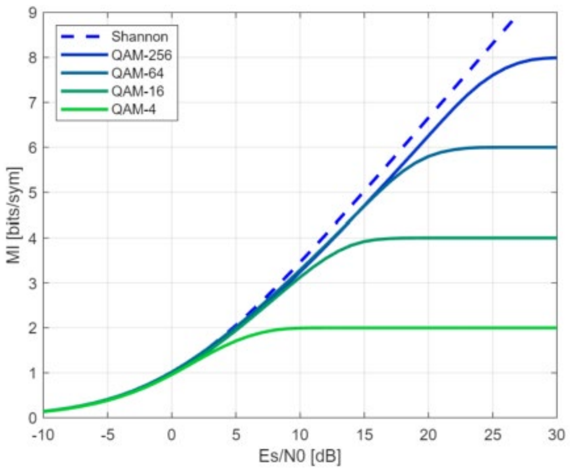
Bitwise Capacity



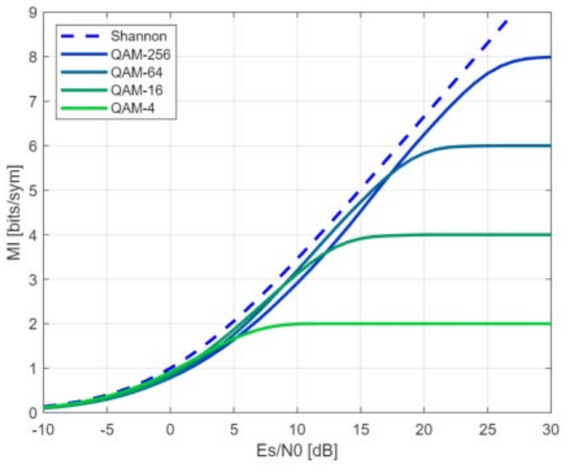
- Each modulation is optimal in a range
  - Select higher modulations at higher SNRs
- At high SNRs:
  - Need to select high modulation
- Relative to Shannon Capacity
  - Minimal loss at low SNRs (< 2 dB)
  - Loss of 1-2 dB at high SNRs

Bitwise vs Symbol-wise Decoding

Symbol-wise decoding

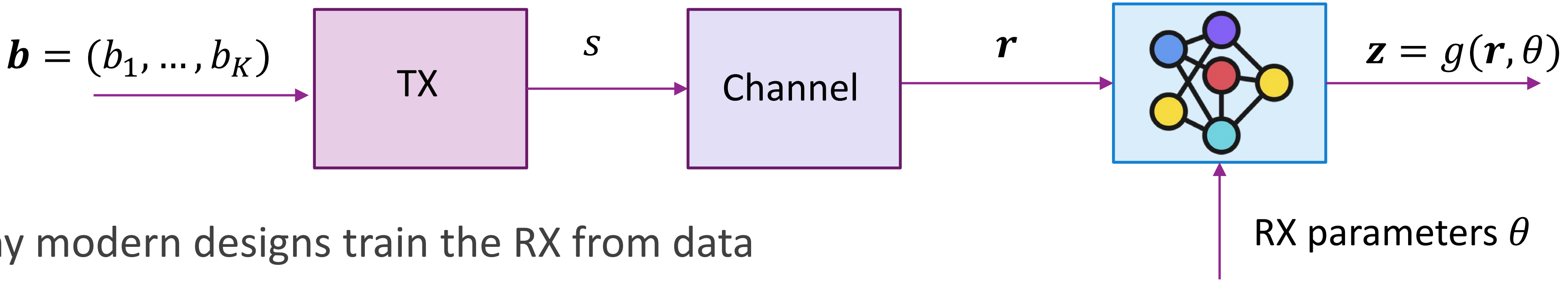


Bitwise decoding



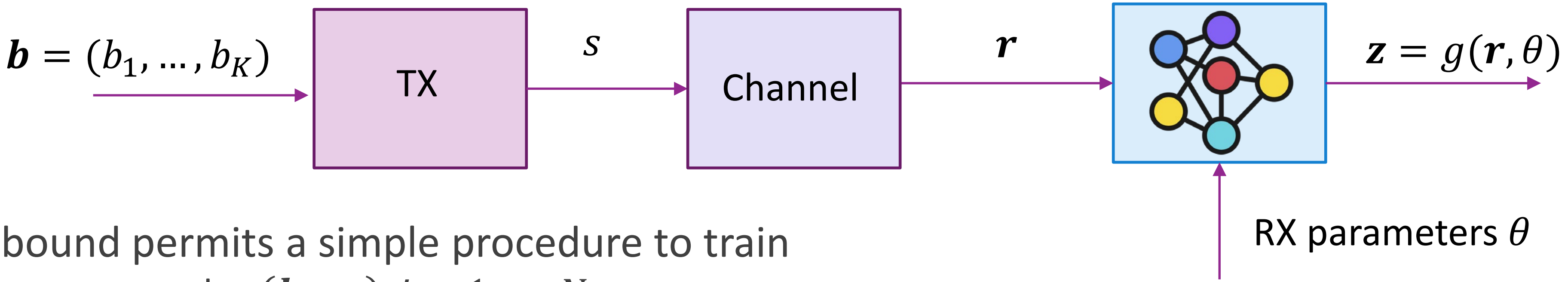
- Bitwise decoding has a small loss
- But if correct constellation is chosen:
  - Loss is small
  - In most regimes, loss < 0.5 dB

ML Perspective: Learn a RX from Data



- Many modern designs train the RX from data
- Represent RX as a function $\boldsymbol{z} = g(\boldsymbol{r}, \theta)$ where θ represents parameters to train
  - Ex: $g(\boldsymbol{r}, \theta)$ is a neural network and θ are the weights and biases
- Can be useful when optimal receiver is difficult to derive or implement
  - Non-coherent channel (when the channel must be estimated)
  - Joint equalization and decoding
  - Non-linearities
  - Computational constraints
  - Many possibilities…

Training a RX



- BCE bound permits a simple procedure to train
  - Generate samples $(\boldsymbol{b}_i, \boldsymbol{r}_i), i = 1, \dots, N$.
  - Each $\boldsymbol{b}_i = (b_{i1}, \dots, b_{iK})$ = true bits transmitted
  - RX will generate outputs: $\boldsymbol{z}_i = g(\boldsymbol{r}_i, \theta)$ with outputs $\boldsymbol{z}_i = (z_{i1}, \dots, z_{iK})$
  - Adjust parameters θ to minimize BCE loss:

$$J(\theta) = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} BCE(b_{ik}, z_{ik})$$

  - Then mutual information is bounded above by:

$$I(b; z) \geq K - J(\theta)$$

BCE Bound Proof: Entropy Bound

- To prove BCE bound, we need the following Lemma
- Lemma: Suppose X has some PMF $P(x)$ and Q is any other distribution. Then

$$H(X) \leq -\sum_x P(x) \log Q(x)$$

- Equality holds when $Q(x) = P(x)$
- Also applies to conditional distributions. If $Q(x|y)$ is any conditional distribution:

$$H(X|Y) \leq -\sum_x P(x,y) \log Q(x|y)$$

Proof of Lemma

- Let $J(P,Q) \coloneqq -\sum_x P(x) \log Q(x)$
- Find $Q(x)$ to minimize $J(P,Q)$ s.t. $\sum Q(x) = 1$
- Lagrangian: $L = -\sum_x P(x) \log Q(x) + \lambda \sum Q(x)$
- Take derivative: $\frac{\partial L}{\partial Q(x)} = -\frac{P(x)}{Q(x)} + \lambda = 0 \Rightarrow Q(x) = \lambda P(x)$
- Since $\sum Q(x) = 1 \Rightarrow Q(x) = P(x)$
- Hence the minimum is achieved at $Q(x) = P(x)$
- Therefore, for all $Q(x)$:

$$J(P,Q) \geq \min_{\mathrm{Q}} J(P,Q) = J(P,P) = H(X)$$

Proof BCE Bound

- Let $P(\boldsymbol{b})$ = true distribution on bits $\boldsymbol{b} = (b_1, \dots, b_K)$
- Given z_k define the conditional binary distribution:

$$\phi(b_k = 1|z_k) = \frac{e^{z_k}}{1+e^{z_k}}, \qquad \phi(b_k = 0|z_k) = \frac{1}{1+e^{z_k}},$$

- Given $\boldsymbol{z}$, define the distribution on the bits $\boldsymbol{b}$ as $Q(\boldsymbol{b}|\boldsymbol{r}) = \prod_k \phi(b_k|z_k)$
  - The bits are conditionally independent
- Can verify that $-\log \phi(b_k|z_k) = \log(1+e^{z_k}) + b_k z_k = BCE(b_k, z_k)$
- Also $\log Q(\boldsymbol{b}|\boldsymbol{r}) = \sum \log \phi(b_k|z_k)$
- By Lemma: $H(\boldsymbol{b};\boldsymbol{r}) \leq -\sum E\{\log \phi(b_k|z_k)\} = \sum E\,[BCE(b_k, z_k)]$
- Therefore: $I(\boldsymbol{b};\boldsymbol{r}) = H(\boldsymbol{b}) - H(\boldsymbol{b};\boldsymbol{r}) \geq H(\boldsymbol{b}) - \sum E\,[BCE(b_k, z_k)]$

Outline

- Information theory basics
- Shannon capacity
- Modeling capacity of practical systems
- Constellation constrained capacity
- Proof of the Shannon Theorem

Proof: Achievability

- First, we show that any $R < C$ is achievable
- Use a random codebook!
- Find a $P(x)$ to maximize $I(X;Y)$ and select any $R < I(X;Y)$
- For each n, generate $M = 2^{Rn}$ random messages or codewords:
  - $\boldsymbol{x}_m = (x_{m1}, \ldots, x_{mn}), \;\; x_{mi} \sim P(x)$ are iid
  - Set of $\boldsymbol{x}_m, \;\; m = 1, \ldots, M$ is called the message index
  - Encoder maps Rn bits to a message index m and transmits $\boldsymbol{x}_m$
- Each message $\boldsymbol{x}_m$ is called a codeword
- The set of messages is called the codebook:

$$\mathcal{C} = \{ \; \boldsymbol{x}_m, m = 1, \ldots, M \; \}$$

Joint Typicality

- For large n, we know (via the law of large numbers)
  - $(1/n) \log P(x_1, \dots, x_n) \to -H(X)$
  - $(1/n) \log P(y_1, \dots, y_n) \to -H(Y)$
  - $(1/n) \log P(x_1, y_1, \dots, x_n, y_n) \to -H(X, Y)$
- Say a vector $(\boldsymbol{x}, \boldsymbol{y})$ is jointly typical if it satisfies the asymptotic values within some $\epsilon > 0$
- Formally, we define the set A_ϵ^n of length n sequences $(\boldsymbol{x}, \boldsymbol{y})$ such that:
  - $|\,(1/n) \log P(x_1, \dots, x_n) \to -H(X)\,| \leq \epsilon$
  - $|\,(1/n) \log P(y_1, \dots, y_n) \to -H(Y)\,| \leq \epsilon$
  - $|(1/n) \log P(x_1, y_1, \dots, x_n, y_n) \to -H(X, Y)| \leq \epsilon$

Jointly Typical Decoder

- Let $\mathcal{C}$ be the set of codewords
- Given $\boldsymbol{y}$ receiver takes any $\boldsymbol{x} \in \mathcal{C}$ from codebook such that $(\boldsymbol{x}, \boldsymbol{y}) \in A_\epsilon^n$
  - That is, find $\boldsymbol{x} \in \mathcal{C}$ such that $(\boldsymbol{x}, \boldsymbol{y})$ is jointly typical
  - If no such $\boldsymbol{x}$ exists, or there is more than one, declare error
- To analyze, suppose we transmit a true sequence $\boldsymbol{x}$ and receive $\boldsymbol{y}$
- We bound two errors:
  - Type 1 Error: The correct codeword, $(\boldsymbol{x}, \boldsymbol{y})$, is not jointly typical
  - Type 2 Error: There is another codeword, $(\boldsymbol{x}', \boldsymbol{y}) \in A_\epsilon^n$ for some $\boldsymbol{x}' \neq \boldsymbol{x}$

Type 1 Error

- We use the following asymptotic equipartition property (AEP)
- AEP 1: Let $(\boldsymbol{x}, \boldsymbol{y}) = \{(x_i, y_i), i = 1, \ldots, n\}$ where $(x_i, y_i) \sim P(x, y)$ are i.i.d. Then

$$P\left((\boldsymbol{x}, \boldsymbol{y}) \in A_\epsilon^n\right) \to 1 \text{ as } n \to \infty$$

- Let $\boldsymbol{x}$ be the true transmitted codeword
- Then $(\boldsymbol{x}, \boldsymbol{y})$ has components $(x_i, y_i) \sim P(x, y)$
- Let P_1= Probability Type 1 error = Probability that $(\boldsymbol{x}, \boldsymbol{y})$ is not jointly typical
- By AEP 1, $\mathrm{P}_1 = 1 - P((\boldsymbol{x}, \boldsymbol{y}) \in A_\epsilon^n) \to 0$

Type 2 Error

- For this error, we use the following AEP property

- AEP 2: Let $(\boldsymbol{x}, \boldsymbol{y}) = \{(x_i, y_i), i = 1, \dots, n\}$ where $(x_i, y_i) \sim P(x)P(y)$ are i.i.d. Then :

$$P((\boldsymbol{x}, \boldsymbol{y}) \in A_\epsilon^n) \le 2^{-n(I(X;Y)-3\epsilon)}$$

  - In this case, for each i, x_i^n and y_i^n are drawn independent

- Property shows with very high probability they will not be jointly typical

Type 2 Error Continued

- For some n, let $\boldsymbol{x}$ be the true transmitted codeword and $\boldsymbol{y}$ the received symbols
- Let P_2= probability that there exists a codeword $\boldsymbol{x}' \neq \boldsymbol{x}$ where $(\boldsymbol{x}', \boldsymbol{y}) \in A_\epsilon^n$
- Since codewords are independent, $(\boldsymbol{x}', \boldsymbol{y})$ has components $(x_i', y_i) \sim P(x)P(y)$
- By AEP 2, $P((\boldsymbol{x}', \boldsymbol{y}) \in A_\epsilon^n) \leq 2^{-n(I(X;Y)-3\epsilon)}$
- Since there are 2^{nR} wrong codewords $\boldsymbol{x}'$ by union bound:

$$P_2 \leq 2^{nR} 2^{-n(I(X;Y)-3\epsilon)} = 2^{n(R-I(X;Y)-3\epsilon)}$$

- We know $R < I(X;Y) - 3\epsilon$ for some ϵ
- Therefore, we can select ϵ such that $\lim_{n\to\infty} P_2 = 0$

Converse Proof

- Must show that for any rate $R > C,\ P_e$ is bounded away from zero
- We will not cover this.
- This is proved via Fano's inequality

- Take information theory class for more!