CS3236
AY22/23 SEM 2
github/jovyntls

00. INTRODUCTION

data compression

- types of compression
  - **lossless compression** - can recover the contents
  - **lossy compression** - lose some quality - cannot convert back to the higher-quality version
- examples
  - sparse binary string - storing positions of 1s
  - equal number of 0/1s - $L \geq \log_2 \binom{64}{32} \approx 60.7$
  - english text - using relative frequency
  - morse code is NOT binary (contains spaces)
- info theory uses **probabilistic models** (letter frequency, sequence probabilities)
- 2 distinct approaches to compression:
  - **variable length** - map more probable sequences to shorter binary strings
  - **fixed length** - map most probable sequences to strings of a given length
    - insufficient strings for low-probability sequences
    - tradeoff between length/failure probability

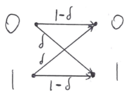
information theory concepts

- speed: **rate** $\rightarrow \frac{k}{n}$ (mapping k bits to n bits)
- reliability: $\mathbb{P}[error]$ = $\mathbb{P}[\text{estimated msg} \neq \text{true msg}]$
- **source coding theorem** $\rightarrow$ the fundamental compression limit is given by a source-dependent quantity known as the **(Shannon) entropy** H. The (average) storage length can be arbitrarily close to H, but can never be any lower than H.
  - H is a property of the *probability distribution*
- **channel coding theorem** $\rightarrow$ there exists a channel-dependent quantity called the **(Shannon) capacity** C such that arbitrarily small error probability can be achieved only for rates $< C$
  - can achieve $\mathbb{P}[error] \leq \epsilon \iff$ rate $< C$

data communication example

- a "transmitter" sends a sequence of 0s and 1s
- a "receiver" sends a sequence *with some corruptions*

channel transition diagram



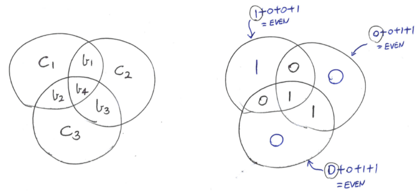
- each bit is flipped independently with probability $\delta \in (0, \frac{1}{2})$

naive

- **uncoded communication** - $\mathbb{P}[correct] = (1-\delta)^N$
- **repetition code** - transmit "000" for "0", "111" for "1"
  - $\mathbb{P}[correct] = [(1-\delta)^3 + 3\delta(1-\delta)^2]^N$
  - more reliable but 3x slower!

Hamming code

- able to correct one bit flip
- maps binary string of length 4 to binary string of length 7
- fill in $b_1 b_2 b_3 b_4$ and assign $c_1 c_2 c_3$ such that the sum of bits in each circle is even



- $\mathbb{P}[correct] \geq \mathbb{P}[\leq 1\text{bit flips}] = (1-\delta)^7 + 7\delta(1-\delta)^6$
- with $\delta = 1$: Shannon capacity $C \approx 0.531$

01. INFORMATION MEASURES

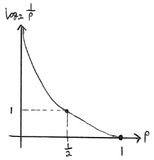
information of an event

- **entropy** $\rightarrow$ measure of "uncertainty" or "information" in a random variable
- given event A with some $\mathbb{P}[A] = p$, how much "information" learned by being told A occurred?
  - only $\mathbb{P}[A]$ matters
- if A occurs with probability p, then $Information(A) = \psi(p)$ for some function $\psi(\cdot)$

axioms for $\psi(\cdot)$

$$\psi(p) = \log_b \frac{1}{p} \text{ (for some base } b > 0)$$

we gain $\log_2 \frac{1}{p}$ "bits" of info if a probability-p event occurs.



- only $\psi(p) = \log_b \frac{1}{p}$ satisfies all axioms
- we focus on $b = 2$
  - information measured in bits
- all choices of b are equivalent up to scaling by a universal constant
  - e.g. # of nats $= \log_e 2 \times$ # of bits

1. $\psi(p) \geq 0$ (**non-negativity**)
2. $\psi(1) = 0$ (**zero for definite events**)
3. if $p \leq p'$, then $\psi(p) \geq \psi(p')$ (**monotonicity**)
   - the less likely an event is, the more information was learnt by the fact that it occurred
4. $\psi(p)$ in continuous in p (**continuity**)
   - small change in probability: no drastic change in info
5. $\psi(p_1 p_2) = \psi(p_1) + \psi(p_2)$
   - (**additivity under independence**) if A and B are independent events with probabilities p_1 and p_2, then $\mathbb{P}[A \cap B] = p_1 p_2$, and the information learnt from both A and B occurring is the sum of the two individual amounts of information (because they are independent)
   - $\psi(\mathbb{P}[A_1 \cap A_2]) = \psi(\mathbb{P}[A_1]) + \psi(\mathbb{P}[A_2])$

information of a random variable - entropy

- let X be a discrete r.v. with pmf P_X
- if we observe $X = x$ then we have learnt $\log_2 \frac{1}{P_X(x)}$ bits of information

(Shannon) entropy
is the average *information/uncertainty* in X wrt P_X:

$$H(X) = \mathbb{E}_{X \sim P_X}\left[\log_2 \frac{1}{P_X(X)}\right] = \sum_x P_X(x) \log_2 \frac{1}{P_X(x)}$$

- **binary entropy function** $\rightarrow$
  $$H_2(p) = p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{1-p}$$
- e.g.
  - binary source: $X \sim Bernoulli(p), \quad p \in (0,1)$
    $\Rightarrow H(X) = H_2(p)$
  - uniform source: X is uniform on a finite set $\mathcal{X}$
    - $P_X(x) = \frac{1}{|\mathcal{X}|}$
    $\Rightarrow H(X) = \mathbb{E}\left[\log_2 \frac{1}{1/|\mathcal{X}|}\right] = \log_2 |\mathcal{X}|$
- entropy $\neq$ variance
  - entropy depends *only* on the probability values

axiomatic view (Shannon)

X is a d.r.v. taking N values with $\mathbf{p} = (p_1, \dots, p_N)$. We consider a general information measure of the form
$$\Phi(\mathbf{p}) = \Phi(p_1, \dots, p_N)$$
only $\Phi(X) = constant \times H(X)$ satisfies all axioms.

1. $\Psi(\mathbf{p})$ is continuous on p (**continuity**)
2. if $p_i = \frac{1}{N}$, then $\Psi(\mathbf{p})$ is increasing in N (**uniform case**)
   - uniformity over a larger set of outcomes always means more uncertainty
3. (**successive decisions**) $\Psi(p_1, \dots, p_N) = \Psi(p_1+p_2, p_3, \dots, p_N) + (p_1+p_2)\Psi(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2})$

variations

- **joint entropy** of two random variables (X, Y) $\rightarrow$
  $$H(X,Y) = \mathbb{E}_{(X,Y)\sim P_{XY}}\left[\log_2 \frac{1}{P_{XY}(X,Y)}\right] = \sum_{x,y} P_{XY}(x,y) \log_2 \frac{1}{P_{XY}(x,y)}$$
- **conditional entropy** of Y given X $\rightarrow$
  $$H(Y|X) = \mathbb{E}_{(X,Y)\sim P_{XY}}\left[\log_2 \frac{1}{P_{Y|X}(Y|X)}\right] = \sum_{x,y} P_{XY}(x,y) \log_2 \frac{1}{P_{Y|X}(y|x)} = \sum_x P_X(x) H(Y|X=x)$$
  - on average, knowing X reduces uncertainty about Y ($H(Y|X) \leq H(Y)$), but seeing a *specific* outcome of X may increase uncertainty about Y ($H(Y|X=i) > H(Y)$ for some values of i)

properties of entropy

1. $H(X) \geq 0$ (**non-negativity**)
   - $H(X) = 0 \iff X$ if deterministic
   - *Proof.* information $\log_2 \frac{1}{p} \geq 0$ for $p \in [0,1]$, so entropy is the average of a non-negative quantity, and itself is non-negative
2. $H(X) \leq \log_2 |\mathcal{X}|$ (**upper bound**)
   if X takes values on a finite alphabet $\mathcal{X}$
   - $H(X) = \log_2 |\mathcal{X}| \iff X \sim Uniform(\mathcal{X})$
   - implies $H(X|Y) \leq \log_2 |\mathcal{X}|$
3. $H(X,Y) = H(X) + H(Y|X)$ (**chain rule**)
   - or $H(X,Y) = H(Y) + H(X|Y)$
   - overall information in (X,Y) is the information in X plus the remaining information in Y after observing X.
   - with conditioning:
     $H(X,Y|Z) = H(X|Z) + H(Y|X,Z)$
   - general chain rule:
     $H(X_1, \dots, X_n) = \sum_{i=1}^n H(X_i | X_1, \dots, X_{i-1})$
4. $H(X|Y) \leq H(X)$ (**conditioning reduces entropy**)
   - $H(X|Y) = H(X) \iff X$ and Y are independent
   - additional information Y can't increase uncertainty *on average* but *can* have $H(X|Y=y) > H(X)$
5. $H(X_1, \dots, X_n) \leq \sum_{i=1}^n H(X_i)$ (**sub-additivity**)
   - equality $\iff X$ and Y are independent

KL Divergence

for two pmfs P and Q on a finite alphabet $\mathcal{X}$, the **Kullback-Leibler (KL) divergence** or **relative entropy** is given by

$$D(P||Q) = \sum_x P(x) \log_2 \frac{P(x)}{Q(x)} = \mathbb{E}_{X\sim P}\left[\log_2 \frac{P(X)}{Q(X)}\right]$$

- $D(P||Q) \neq D(Q||P)$
- $D(P||Q) \geq 0$
  - *Proof.* $-D(P||Q) = -\sum_x P(x) \log_2 \frac{P(x)}{Q(x)}$
    $\leq \sum_x P(x)(\frac{Q(x)}{P(x)} - 1) = \sum_x Q(x) - \sum_x P(x) = 0$
    (using property that $\log \alpha \leq \alpha - 1$, equality iff $\alpha = 1$)
- $D(P||Q) = 0 \iff P = Q$
  - *Proof.* same as above, with $\ln \alpha = \alpha - 1 \iff \alpha = 1$ (then $\frac{P(x)}{Q(x)} = 1$)

Mutual Information

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(X) - H(X|Y) \\ &= H(X) + H(Y) - H(X,Y) \\ &= D(P_{XY} || P_X \times P_Y) \end{aligned}$$

- **mutual information**, $I(X;Y) \rightarrow$ the amount of information we learn about Y by observing X (on avg)
  - $H(Y)$ = uncertainty in Y
  - $H(Y|X)$ = (avg) uncertainty in Y after observing X
  - $D(P_{XY}||P_X P_Y)$ = how far X, Y are from being independent
- $I(X_1; X_2, X_3) \neq I(X_1, X_2; X_3)$
- **joint mutual information** $\rightarrow$
  $$I(X_1, X_2; Y_1, Y_2) = H(Y_1, Y_2) - H(Y_1, Y_2 | X_1, X_2)$$
- **conditional mutual information** $\rightarrow$
  $$I(X;Y|Z) = H(Y|Z) - H(Y|X,Z)$$
- if $X \perp Y$, then $I(X;Y) = 0$
  - *Proof.* $X \perp Y \Rightarrow P_{XY} = P_X \times P_Y \Rightarrow D(P_{XY}||P_X \times P_Y) = 0$
  - independent variables do not reveal any information about each other
- if $X = Y$, then $I(X;Y) = H(X) = H(Y)$
  - amt of information a r.v. reveals about itself is the entropy

properties of mutual information

1. $I(X;Y) = I(Y;X)$ **(symmetry)**
   - X and Y reveal an equal amount of information about each other
2. $I(X;Y) \geq 0$ **(non-negativity)**
   - equality $\iff X \perp Y$
3. $I(X;Y) \leq H(X) \leq \log_2 |\mathcal{X}|$ **(upper bounds)**
   $I(X;Y) \leq H(Y) \leq \log_2 |\mathcal{Y}|$
   - the information X reveals about Y is *at most* the prior information in X (entropy)
4. $I(X,Y;Z) = I(X;Z) + I(Y;Z|X)$ **(chain rule)**
   $$I(X_1,\dots,X_n;Y) = \sum_{i=1}^{n} I(X_i;Y|X_1,\dots,X_{i-1})$$
   $$= I(X_1;Y) + I(X_2;Y|X_1) + \dots$$
5. **(data-processing inequality)**
   $I(X;Z) \leq I(X;Y)$ if $X \to Y \to Z$
   variation: $I(X;Z) \leq I(Y;Z)$ if $X \to Y \to Z$
   $I(W;Z) \leq I(X;Y)$ if $W \to X \to Y \to Z$
   - holds if Z depends on (X,Y) only through Y (i.e. $X \to Y \to Z$ forms a **Markov chain**)
   - processing Y (to produce Z) cannot increase the information available regarding X
     - cannot do data processing to increase information
6. **(partial sub-additivity)**
   $$I(X_1,\dots,X_n;Y_1,\dots,Y_n) \leq \sum_{i=1}^{n} I(X_i;Y_i)$$
   if $(Y_1,\dots,Y_n)$ are conditionally independent given $(X_1,\dots,X_n)$, and Y_i depends on $(X_1,\dots,X_n)$ only through X_i

02. SYMBOL-WISE SOURCE CODING

X is a d.r.v. with pmf P_X over an alphabet $\mathcal{X}$ (set of symbols).

symbol-wise source coding maps each $x \in \mathcal{X}$ to some binary sequence $C(x)$ of length $\ell(x)$.

average length of a code $C(\cdot)$,
$$L(C) = \sum_{x \in \mathcal{X}} P_X(x)\ell(x)$$

decodability conditions

- **nonsingular property** $\to C(x) \neq C(x') \iff x \neq x'$
- $C(\cdot)$ is **uniquely decodable** $\to$ no 2 sequences (of equal or differing lengths) of symbols in $\mathcal{X}$ are coded to the same concatenated binary sequence.
  - $x_1,\dots,x_n$ can be always uniquely identified from the string $C(x_1)\dots C(x_n)$
- $C(\cdot)$ is **prefix-free** $\to$ no codeword is a prefix of another
  - aka **instantaneous code**

Kraft's Inequality and Entropy Bound

Kraft's inequality

if $C(\cdot)$ is *prefix-free*, then $\displaystyle\sum_{x \in \mathcal{X}} 2^{-\ell(x)} \leq 1$

- *Proof.* represent the codewords by a binary tree. If there is a codeword at some point in the tree, there are no codewords further down the tree. probability of branching to a codeword $= 2^{-\ell(x)}$ and sum of probabilities cannot exceed 1
- **existence property** $\to$ if a given set of integers $\{\ell(x)\}_{x \in \mathcal{X}}$ satisfies $\sum_{x \in \mathcal{X}} 2^{-\ell(x)} \leq 1$, then it is possible to construct a *prefix-free* code that maps each $x \in \mathcal{X}$ to a codeword of length $\ell(x)$.

entropy bound

entropy bound

expected length, $L(C) \geq H(X)$

with equality $\iff P_X(x) = 2^{-\ell(x)} \quad \forall x \in \mathcal{X}$

- entropy gives a *fundamental compression limit*
  - average length is at least equal to entropy
  - if all probabilities are negative powers of 2, we can match the entropy bound (optimal code)
- *Proof.* manipulate to get $L(C) - H(X) \geq D(P_X||Q) \geq 0$

Shannon-Fano Code

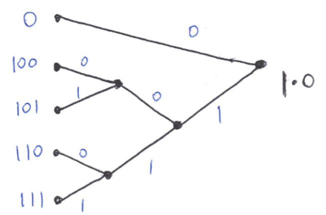
$$\ell(x) = \left\lceil \log_2 \frac{1}{P_X(x)} \right\rceil$$

- **average length**, $L(C)$ satisfies
  $$H(X) \leq L(C) < H(X) + 1$$
- **Kraft's inequality** holds -
  $$\sum_{x \in \mathcal{X}} 2^{-\ell(x)} \leq \sum_{x \in \mathcal{X}} 2^{-\log_2 \frac{1}{P_X(x)}} = \sum_{x \in \mathcal{X}} P_X(x) = 1$$
  - **Existence property** holds - we can construct a prefix-free code with these lengths
- 1 bit may be significant - e.g. if $H(X) = 0.5$
- **mismatched case** -

  if the true distribution is P_X but the lengths are chosen according to Q_X, then the Shannon-Fano code satisfies
  $$H(X)+D(P_X||Q_X) \leq L(C) \leq H(X)+D(P_X||Q_X)+1$$

Huffman Code

- no uniquely decodable symbol code can achieve a smaller length $L(C)$ than the Huffman code.
  - always prefix-free
  - satisfies average length bound (because it is at least as good as Shannon-Fano): $H(X) \leq L(C) < H(X) + 1$



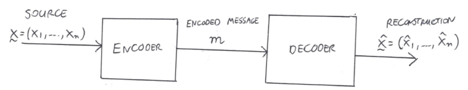
- extension: using blocks of n letters; Huffman coding with $\mathcal{X}^n$
  $nH(X) \leq L(C) < nH(X) + 1$
  $\Rightarrow H(X) \leq$ avg. length per symbol $\leq H(X) + \frac{1}{n}$
  - ✓ exploits *memory*, better guarantee (even independent)
  - × but it's harder to accurately know $P_{X_1 \dots X_n}$
  - × alphabet size increases to $|\mathcal{X}|^n \Rightarrow$ expensive to sort

other codes

- **arithmetic codes** - encodes a sequence $(x_1,\dots,x_n)$ to at most $\ell(x_1,\dots,x_n) \leq \log_2 \frac{1}{P_{X_1,\dots,X_n}(x_1,\dots,x_n)} + 2$
  - avg. length per letter $\leq H(X) + \frac{2}{n}$
- **Lempel-Ziv code** - does not require knowledge of the source distribution
  - near-optimal: $O(\frac{\log n}{n})$ instead of $O(\frac{1}{n})$

03. BLOCK-WISE SOURCE CODING

- aka **fixed-to-fixed** length source coding
- $\mathbb{P}[error] > 0$ (but small)
  - map likely source strings, fail on unlikely source strings
- instead of symbol-by-symbol, apply some encoding function to a length-n block $X_1,\dots,X_n$
  - map a string to some integer $m \in \{1,\dots,M\}$
- **discrete memoryless source** $(X_1,\dots,X_n)$
  - *discrete* - the alphabet $\mathcal{X}$ is finite
  - *memoryless* - $P_{\underset{\sim}{X}}(\underset{\sim}{x}) = \Pi_{i=1}^{n} P_X(x_i)$
    - every letter is independent (unrealistic)



- *decoder* maps m to an estimate $\hat{\underset{\sim}{X}} = g(m)$ (in $\mathcal{X}^n$)
- **error** $\to$ occurs if $\hat{\underset{\sim}{X}} \neq \underset{\sim}{X}$
  - $P_e = \mathbb{P}[\hat{\underset{\sim}{X}} \neq \underset{\sim}{X}] = \sum_{\underset{\sim}{x}:\text{DEC(ENC}(x))\neq x} P_{\underset{\sim}{X}}(\underset{\sim}{x})$
- **rate** $\to R = \frac{1}{n}\log_2 M$
  - ratio of compressed length ($\log_2 M$) to source length (n)
    - represents the number of bits per source symbol used to represent encoded value m
  - number of strings we can compress to, $M = 2^{nR}$
  - lower rate = more compression
  - $R \leq H(X) + \epsilon$
    - *Proof.* $R = \frac{1}{n}\log_2 M = \frac{1}{n}\log_2(|\mathcal{T}_n(\epsilon)| + 1)$
      $\simeq \frac{1}{n}\log_2 |\mathcal{T}_n(\epsilon)| \leq H(X) + \epsilon$ (using property 3)
- **fixed length source coding theorem** $\to$ for any discrete memoryless source with per-symbol distribution P_X,
  - (**achievability**) if $R > H(X)$, then for any $\epsilon > 0$, we can get $P_e \leq \epsilon$ for large enough n
  - (**converse**) if $R < H(X)$, then there exists $\epsilon > 0$ such that $P_e > \epsilon$ for all n

Typical Sequences

for i.i.d. sequence $\mathbf{X} = (X_1,\dots,X_n)$, let $P_{\mathbf{X}}(\mathbf{x}) = \Pi_{i=1}^{n} P_X(x_i)$ be the pmf of $\mathbf{X}$.

typical set, $\mathcal{T}_n(\epsilon) =$
$$\left\{ \mathbf{x} \in \mathcal{X}^n : 2^{-n(H(X)+\epsilon)} \leq P_{\mathbf{X}}(\mathbf{x}) \leq 2^{-n(H(X)-\epsilon)} \right\}$$
where $\epsilon > 0$ is a (small) fixed constant

i.e. $P_{\underset{\sim}{X}}(\underset{\sim}{x}) \simeq 2^{-nH(\underset{\sim}{X})}$

- we only assign a (unique) $m \in \{1,\dots,M\}$ to *some* $\underset{\sim}{x}$
  - choose $\underset{\sim}{x}$ such that $\mathbb{P}[\underset{\sim}{x} \in \mathcal{T}_n(\epsilon)] \simeq 1$

properties of a typical set

for any fixed $\epsilon > 0$,

1. **(equivalent definition)** $\mathbf{x} \in \mathcal{T}_n(\epsilon) \iff$
   $$H(X) - \epsilon \leq \frac{1}{n}\sum_{i=1}^{n}\log_2 \frac{1}{P_X(x_i)} \leq H(X) + \epsilon$$
   where x_i is the i-th entry of $\mathbf{x}$
   - $\mathbb{E}[\log P_X(x_i)] = H(X_i) = H(X)$
2. $\mathbb{P}[\mathbf{X} \in \mathcal{T}_n(\epsilon)] \to 1$ as $n \to \infty$ **(high probability)**
3. $|\mathcal{T}_n(\epsilon)| \leq 2^{n(H(X)+\epsilon)}$ **(cardinality upper bound)**
4. $|\mathcal{T}_n(\epsilon)| \geq (1 - o(1))2^{n(H(X)+\epsilon)}$ where $o(1) \to 0$ as $n \to \infty$ **(cardinality lower bound)**
   $\Rightarrow$ we can't improve much on property (3)

asymptotic equipartition property

asymptotic equipartition property

as $n \to \infty$, the distribution is roughly uniform over $\mathcal{T}_n(\epsilon)$

- with high probability (property 2), a randomly drawn i.i.d. sequence $\mathbf{X}$ will be one of roughly $2^{n(H(X))}$ sequences (property 3 + 4), each of which has probability of roughly $2^{-nH(X)}$ (definition of typical set)

Fano's Inequality

let X denote a *generic* r.v., and $\hat{X}$ is any estimate of X.

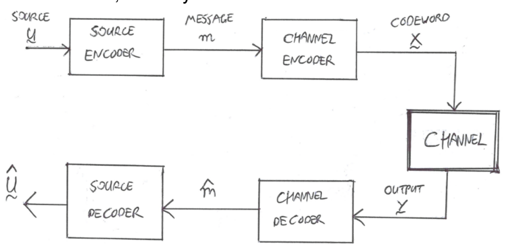
Fano's Inequality

$$H(X|\hat{X}) \leq H_2(P_e) + P_e \log_2(|\mathcal{X}| - 1)$$
$$\leq 1 + P_2 \log_2 |\mathcal{X}|$$

- intuition: if $H(X|\hat{X})$ is large, then $P_2 = \mathbb{P}[\hat{X} \neq X]$ should be large too
- uncertainty in X after observing $\hat{X} \leq$ uncertainty in "is $X = \hat{X}$?" + ($\mathbb{P}$[no]$= P_e$)(max uncertainty in the no case)
- implications for source coding: proves the **converse** clause of **fixed length source coding theorem**
  - if $R < H(X)$, then $P_e = \mathbb{P}[\hat{X} \neq X]$ cannot be made arbitrarily small as $n \to \infty$

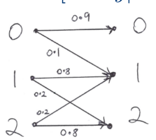
04. CHANNEL CODING

- transmit a message $m \in \{1,\dots,M\}$
  - using a fixed-length source code that outputs a length-k sequence, we can set $M = s^k$
- encoder: message $m \Rightarrow$ channel inputs $x_1,\dots,x_n$
- **codeword** $\to \mathbf{x}^{(m)} = (x_1^{(m)},\dots,x_n^{(m)})$
  - transmitted over the channel in n uses
- **codebook** $\to \mathcal{C} = \{\mathbf{x}^{(1)},\dots,\mathbf{x}^{(M)}\}$
  - collection of codewords known by both encoder and decoder, but only the encoder knows m



for input x, output y, input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$

- **channel** $\to$ medium over which we transmit information
  - **discrete** $\to$ input/output alphabets $\mathcal{X}$ and $\mathcal{Y}$ are finite
  - **memoryless** $\to$ outputs are (conditionally) independent:
    $\mathbb{P}[Y = y|X = x] = \Pi_{i=1}^{n} P_{Y|X}(y_i|x_i)$



probabilistic modelling approach $\to$ when the input is $x \in \mathcal{X}$, a given output $y \in \mathcal{Y}$ is produced with probability $P_{Y|X}(y|x)$
- see channel transition diagram

- **error probability** $\to P_e = \mathbb{P}[\hat{m} \neq m]$
  - assuming uniform distribution

- on non-uniform distribution: can use $P_{e_i \max}$
- **rate** $\to R = \frac{1}{n} \log_2 M$ for block length n
  - higher rate = sending faster (opposite of source coding where lower is better)
  - $= \frac{k}{n}$ for sending k bits
  - $R \leq 1$ for binary channels

Channel Capacity

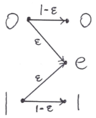
- **channel capacity**, $C \to$ maximum of all rates R such that, for any target error probability $\epsilon > 0$, there exists a block length n and codebook $\mathcal{C} = \{x^{(1)}, \ldots, x^{(M)}\}$ with $M = 2^{nR}$ codewords such that $P_e \leq \epsilon$

channel coding theorem

for any discrete memoryless channel $C(P_{Y|X})$, we have

$$C = \max_{P_x} I(X;Y)$$

- *capacity-achieving input distribution*: input distribution P_X that maximises the mutual information
  - we can maximise P_X, but cannot control $I(X;Y)$
  - usually (but not always) uniform for "symmetric" channels
- (**achievability**) for any $R < C$, there exists a code of rate $\geq R$ with arbitrarily small P_e
- (**converse**) for any $R > C$, any code rate $\geq R$ cannot have arbitrarily small P_e (for any codebook)
- examples
  - noiseless channel ($\mathcal{X} = \mathcal{Y} = \{0,1\}$) (deterministic):
    $C = \max_{P_X} I(X;Y) = \max_{P_X} H(X) = 1$
  - binary symmetric channel ($\mathcal{X} = \mathcal{Y} = \{0,1\}$):
    $$P_{Y|X}(y|x) = \begin{cases} 1-\delta & y = x \\ \delta & y = 1-x \end{cases}$$
    $$C = \max_{P_X} I(X;Y) = \max_{P_X}(H(Y) - H_2(\delta))$$
    $$= \max_{P_X}(H_2(\mathbb{P}[Y=1]) - H_2(\delta)) = 1 - H_2(\delta)$$
    - we can't maximise $\mathbb{P}[Y=1]$ directly but we can let P_X be uniform to get $P_Y(1) = \frac{1}{2}$
  - binary erasure channel ($\mathcal{X} = \{0,1\}, \mathcal{Y} = \{0,1,e\}$):
    - for erasure probability ϵ



$$P_{Y|X}(y|x) = \begin{cases} 1-\epsilon & y = x \\ \epsilon & y = e \\ 0 & y = 1-x \end{cases}$$
$$C = \max_{P_X} I(X;Y) = \max_{P_X}(H(X) - H(X|Y))$$
$$= \max_{P_X}(H(X) - \epsilon H(X)) = 1 - \epsilon$$

    - maximising $H(Y)$ doesn't work here - you can't get an arbitrary $P(Y)$ distribution

Jointly Typical Sequences

a pair of $(\mathbf{x}, \mathbf{y})$ of length-n input and output sequences is **jointly typical** wrt a joint distribution P_{XY} if

$$2^{-n(H(X)+\epsilon)} \leq P_X(\mathbf{x}) \leq 2^{-n(H(X)-\epsilon)}$$
$$2^{-n(H(Y)+\epsilon)} \leq P_Y(\mathbf{y}) \leq 2^{-n(H(Y)-\epsilon)}$$
$$2^{-n(H(X,Y)+\epsilon)} \leq P_{XY}(\mathbf{x},\mathbf{y}) \leq 2^{-n(H(X,Y)-\epsilon)}$$

- aka: the X sequence, Y sequence, and joint (X,Y) sequence are all typical
- **jointly typical set**, $\mathcal{T}_n(\epsilon) \to$ the set of all jointly typical sequences
- a joint distribution on sequences:
  $P_{\mathbf{x},\mathbf{y}}(\mathbf{x},\mathbf{y}) = \Pi_{i=1}^n P_{XY}(x_i, y_i)$ - independent product

properties

1. (**equivalent definition**) $(\mathbf{x},\mathbf{y}) \in \mathcal{T}_n(\epsilon) \iff$
   $H(X) - \epsilon \leq \frac{1}{n}\sum_{i=1}^n \log_2 \frac{1}{P_X(x_i)} \leq H(X) + \epsilon$
   $H(Y) - \epsilon \leq \frac{1}{n}\sum_{i=1}^n \log_2 \frac{1}{P_Y(y_i)} \leq H(Y) + \epsilon$
   $$H(X,Y) - \epsilon \leq \frac{1}{n}\sum_{i=1}^n \log_2 \frac{1}{P_Y(x_i,y_i)} \leq H(X,Y) + \epsilon$$
2. (**high probability**)
   $\mathbb{P}[(\mathbf{X},\mathbf{Y}) \in \mathcal{T}_n(\epsilon)] \to 1$ as $n \to \infty$
   - because law of large numbers on the above 3
3. (**cardinality upper bound**)
   $|\mathcal{T}_n(\epsilon)| \leq 2^{n(H(X,Y)+\epsilon)}$
4. (**probability for independent sequences**)
   if $(\mathbf{X}', \mathbf{Y}') \sim P_X(\mathbf{x}')P_Y(\mathbf{y}')$ are independent copies of $(\mathbf{X},\mathbf{Y})$, then the probability of joint typicality is $\mathbb{P}[(\mathbf{X}',\mathbf{Y}') \in \mathcal{T}_n(\epsilon)] \leq 2^{-n(I(X;Y)-3\epsilon)}$
   - intuition: for an independent draw from X and an independent draw from Y (instead of joint distribution), the probability of being typical is much lower
   - mutual information (computed from joint distribution): how far X,Y are from being independent

Achievability via Random Coding

for codebook $\mathcal{C} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)}\}$, where m is encoded into length-n sequence $\mathbf{x}^{(m)} = (x_1^{(m)}, \ldots, x_n^{(m)})$

- idea: prove the existence of a good codebook without explicitly constructing it
  - for some random $\mathcal{C}$, show $\mathbb{E}[P_e(\mathcal{C})] \leq \epsilon$ (thus $\exists$ some $\mathcal{C}$ with $P_e \leq \epsilon$)
  - let each codeword be i.i.d. according to P_X
- **random coding** $\to$ generate each symbol $X_i^{(m)}$ of each codeword randomly and independently according to some distribution P_X.
  - *encoder*: maps m to $\mathbf{X}^{(m)} = (X_1^{(m)}, \ldots, X_n^{(m)})$
  - *decoder*: form estimate $\hat{m}$ from output sequence $\mathbf{Y} = (Y_1, \ldots, Y_n)$
    - if $!\exists m'$ s.t. $(\mathbf{X}^{(m')}, \mathbf{Y}) \in \mathcal{T}_n(\epsilon)$, set $\hat{m} = m'$
      - if there is a single index where the codeword and received sequence are jointly typical
    - else give up (treat as error)
- for $\mathbf{X}^{(m)}$ transmitted (i.e. correct m)
  - $(\mathbf{X}^{(m)}, \mathbf{Y})$ is i.i.d. on $P_{XY} = P_X \times P_{Y|X}$
  - since $P_{\mathbf{Y}|\mathbf{X}}$ is i.i.d. according to $P_{Y|X}$, $\mathbf{X}^{(m)}$ is i.i.d. according to P_X (by construction)
- for $\mathbf{X}^{(\hat{m})}$ not transmitted (i.e. incorrect $\hat{m}$),
  - $(\mathbf{X}^{(m')}, \mathbf{Y}) \sim P_{\mathbf{X}}(\mathbf{x}')P_{\mathbf{Y}}(\mathbf{y}')$
  - joint distribution is an independent product - $\mathbf{Y}$ only depends on $\mathbf{X}^{(m)}$, and $P_{\mathbf{X}}$ is i.i.d.

error probability

- we have $\hat{m} = m$ if:
  1. $(\mathbf{X}^{(m)}, \mathbf{Y})$ is jointly typical
  2. none other $(\mathbf{X}^{(\hat{m})}, \mathbf{Y})$ is jointly typical (with $\hat{m} \neq m$)
- $\mathbb{P}$[success] $\geq \mathbb{P}$[① and ②] $\Rightarrow \mathbb{P}$[failure] $\leq \mathbb{P}$[not ① $\cup$ not ②]

$$P_e \leq \mathbb{P}[(\mathbf{X}^{(m)}, \mathbf{Y}) \notin \mathcal{T}_n(\epsilon) \cup \bigcup_{m' \neq m} \{(\mathbf{X}^{(m')}, \mathbf{Y}) \in \mathcal{T}_n(\epsilon)\}]$$
$$\leq \mathbb{P}[(\mathbf{X}^{(m)}, \mathbf{Y}) \notin \mathcal{T}_n(\epsilon)] + \sum_{\hat{m} \neq m} \mathbb{P}[(\mathbf{X}^{(\hat{m})}, \mathbf{Y}) \notin \mathcal{T}_n(\epsilon)]$$
$$\leq \delta_n + \sum_{\hat{m} \neq m} 2^{-n(I(X;Y)-3\epsilon)} \text{ where } \delta \to 0 \text{ as } n \to \infty$$
$$\leq \delta_n + M \times 2^{-n(I(X;Y)-3\epsilon)}$$

- $R < I(X;Y) - 3\epsilon$ since $M = 2^{nR} \Rightarrow$ thus P_e can be arbitrarily small for any rate R arbitrarily close to $I(X;Y)$
- choose P_X to achieve $C = \max_{P_x} I(X;Y)$
- then we can get vanishing error probability rates for rates arbitrarily close to capacity C

Converse via Fano's Inequality

relates $P_e = \mathbb{P}[\hat{m} \neq m]$ to $H(m|\hat{m})$ and thus to $I(m;\hat{m})$

Proof.

- Fano's inequality:
  $H(m|\hat{m}) \leq H_2(P_e) + P_2 \log_2(M-1) \leq 1 + P_e \log_2 M$
  - H(are they equal?) + remaining uncertainty if they're not
- mutual information: $I(m|\hat{m}) = H(m) - H(m|\hat{m})$
  $= \log_2 M - H(m|\hat{m})$ since m is uniform on $\{1, \ldots, M\}$
  $\geq (1 - P_e)\log_2 M - 1 \Rightarrow P_e \geq 1 - \frac{I(m;\hat{m})+1}{\log_2 M}$
- data processing inequality: $I(m;\hat{m}) \leq I(\mathbf{X};\mathbf{Y})$
  - $\mathbf{X} = \mathbf{X}^{(m)}$ is the transmitted codeword; $\mathbf{Y}$ is the channel output; markov chain $m \to \mathbf{X} \to \mathbf{Y} \to \hat{m}$
- manipulate: $I(m;\hat{m}) \leq I(\mathbf{X};\mathbf{Y}) = H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X})$
  $$\leq \sum_{i=1}^n H(Y_i) - \sum_{i=1}^n H(Y_i|\mathbf{X}) = \sum_{i=1}^n I(X_i;Y_i) \leq nC$$

result

combine with $\log_2 M = nR$ to get $P_e \geq 1 - \dfrac{nC+1}{nR}$

thus if $R > C$, we can't get $P_e \to 0$ as $n \to \infty$ (for any x)

05. CONTINUOUS-ALPHABET CHANNELS

- so far X and Y have been discrete/finite
- for continuous, we use *pdf* instead of *pmf*

Differential Entropy

- not directly interpretable as a measure of uncertainty

differential entropy of a continuous r.v. X with pdf f_X

$$h(X) = \mathbb{E}_{f_X}\left[\log_2 \frac{1}{f_X(X)}\right]$$
$$= \int_{\mathbb{R}} f_X(x) \log_2 \frac{1}{f_X(x)}\, dx$$

joint version, $h(X,Y) = \mathbb{E}\left[\log_2 \dfrac{1}{f_{XY}(x,y)}\right]$

conditional version,

$$h(Y|X) = \mathbb{E}_{(X,Y)\sim f_{XY}}\left[\log_2 \frac{1}{f_{Y|X}(Y|X)}\right]$$
$$= \int_{\mathbb{R}} f_X(x) H(Y|X=x)\, dx$$

properties

properties of entropy that still hold:

- (**chain rule**)
  $h(X_1, \ldots, X_n) = \sum_{i=1}^n h(X_i|X_1, \ldots, X_{i-1})$
- (**conditioning reduces entropy**) $h(X|Y) \leq h(X)$
- (**sub-additivity**) $h(X_1, \ldots, X_n) \leq \sum_{i=1}^n h(X_i)$
- $h(X) = h(X + c)$ for a constant c

properties of entropy that *do not* hold:

- non-negativity: we can have $h(X) < 0$
- invariance under one-to-one transformations: we can have $h(X) \neq h(\psi(X))$ even if ψ is invertible
- *counterexample:* let $Y = cX$
  - then $f_Y(y) = \frac{1}{|c|} f_X(\frac{y}{c})$, which gives
    $$h(Y) = \mathbb{E}[\log_2 \tfrac{1}{f_Y(y)}] = \mathbb{E}[\log_2 \tfrac{|c|}{f_X(Y/c)}]$$
    $$= \log_2 |c| + h(X) \quad \neq h(\psi(X))$$
  - violation of non-negativity: $\log_2 |c| \to \infty$ as $c \to 0$

examples

- **uniform** r.v. $X \sim Uniform(a,b)$ for $a < b$
  - $h(X) = \mathbb{E}[\log_2 \frac{1}{f_X(x)}] = \log_2(b-a)$
- **gaussian** $X \sim N(\mu, \sigma^2)$
  - $h(X) = \frac{1}{2}\log_2(2\pi e \sigma^2)$
  - *Proof.* pdf: $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$
    $\Rightarrow \log_2 \frac{1}{f_X(x)} = \log_2(\sqrt{2\pi\sigma^2}) + \frac{(x-\mu)^2}{2\sigma^2}$
    - $h(X) = \mathbb{E}[\log_2(\sqrt{2\pi\sigma^2}) + \frac{(x-\mu)^2}{2\sigma^2}]$
      $= \log_2(\sqrt{2\pi\sigma^2}) + \frac{1}{2\sigma^2}\mathbb{E}[(x-\mu)^2]$
      $= \frac{1}{2}(\log_2(\sqrt{2\pi\sigma^2}) + 1)$ since variance=1
      $= \frac{1}{2}(\log_2(2\pi\sigma^2) + 1)$
    - $h(X)$ in nats $= \frac{1}{2}(\ln(2\pi\sigma^2) + \ln e)$
      $= \frac{1}{2}\ln(2\pi e \sigma^2)$

Mutual information & KL Divergence

mutual information

$$I(X;Y) = h(Y) - h(Y|X)$$
$$= H(X) - h(X|Y)$$
$$= D(f_{XY} || f_X \times f_Y)$$
$$= \mathbb{E}_{f_{XY}}\left[\log_2 \frac{f_{XY}(x,y)}{f_X(x) f_Y(y)}\right]$$

KL divergence

$$D(f||g) = \int_{\mathbb{R}} f(x) \log_2 \frac{f(x)}{g(x)}\, dx$$

properties

- all key properties are retained, including non-negativity
- $D(f||g) \geq 0$, equality $\iff f = g$
- $I(X;Y) \geq 0$, equality $\iff X \perp Y$
- if $\psi(\cdot)$ and $\phi(\cdot)$ are invertible then $I(X;Y) = I(\psi(X);\phi(Y))$
- $h(\cdot)$ is invariant to shifting by a constant: $h(X+k) = H(X)$, $H(X+Y|X) = H(Y)$

Gaussian Random Variables

if $X \sim N(\mu, \sigma^2)$, then $h(X) = \frac{1}{2}\log_2(2\pi e \sigma^2)$

maximum entropy property $\rightarrow$ for any r.v. X with density f_X and variance $Var[X]$, we have

$$h(X) \leq \frac{1}{2}\log_2(2\pi e Var[X])$$

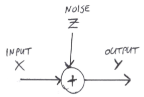
with equality $\iff$ X is Gaussian

- for a given variance, gaussian r.v. has highest entropy $h(\cdot)$
  - no constraint on values, just a constraint on variance
  - discrete: for a given alphabet, uniform maximises $H(\cdot)$
- if $X \in [a, b]$, then uniform maximises $h(\cdot)$
  - (constraint on values)

Gaussian Channel

a continuous channel can be described by conditional pdf $f_{Y|X}$

additive noise channels



- **additive noise channels** $\rightarrow Y = X + Z$
  - Z is a noise term independent of X
  - $f_{Y|X}(y|x) = f_Z(y - x)$
- **additive white Gaussian noise (AWGN) channel** $\rightarrow$ $Z \sim N(0, \sigma^2)$ for some noise variance $\sigma^2 > 0$
  - white = memoryless (independent noise each time)
- **power constraint:** $\mathbb{E}[X^2] \leq P$
  - energy consumed by transmitting X is $\propto X^2$
  - (all lead to the same capacity) average over
    - symbols for each codeword: $\frac{1}{n}\sum_{i=1}^{n} x_i^2 \leq P$ for codewords $\mathbf{x}^{(m)} = (x_1^{(m)}, \ldots, x_n^{(m)})$
    - all codewords: $\frac{1}{M}\sum_{m=1}^{M}(\ldots)$
    - random codebook
  - (not feasible) if X is unconstrained, we can just send different messages using inputs $0, \pm\Delta, \pm 2\Delta, \ldots$ for a huge value of Δ (e.g. 1 million times of variance)

Channel Capacity

AWGN capacity

$$C(P) = \frac{1}{2}\log_2(1 + \frac{P}{\sigma^2})$$

general (non-gaussian)

$$C(P) = \max_{f_X : \mathbb{E}_{f_X}[X^2] \leq P} I(X; Y)$$

- channel capacity $C(P)$ is same as discrete memoryless channels, but codebooks are constrained to satisfy average power constraint

properties of Gaussian channel capacity

- depends on P, σ^2 only through *signal-to-noise ratio* $\frac{P}{\sigma^2}$
- $P = 0 \Rightarrow SNR = 0 \Rightarrow C = 0$
- as $\sigma^2 \rightarrow 0$ for fixed P, then $SNR \rightarrow \infty, C \rightarrow \infty$
- diminishing returns of increasing P
  - for small $\frac{P}{\sigma^2}$, we have $C(P) \approx \frac{P}{2\sigma^2}$ $\Rightarrow$ almost proportional to P
  - for large $\frac{P}{\sigma^2}$, we have $C(P) \approx \frac{1}{2}\log_2\frac{P}{\sigma^2}$ $\Rightarrow$ diminishing returns, doubling P adds $\frac{1}{2}$ to capacity