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

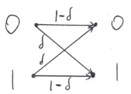
information theory concepts

- speed: **rate** $\rightarrow \frac{k}{n}$ (mapping k bits to n bits)
- reliability: $\mathbb{P}[error]$ = $\mathbb{P}$[estimated msg $\neq$ 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



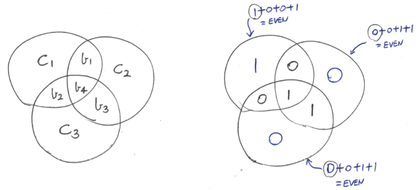
- each bit is flipped independently with probability $\delta \in (0, \frac{1}{2})$

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 $b_1b_2b_3b_4$ and assign $c_1c_2c_3$ 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

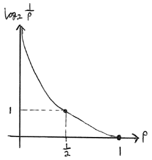
information of an event

- **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. $\psi(p_1p_2) = \psi(p_1) + \psi(p_2)$
   - (**additivity under independence**) if A and B are independent events with probabilities p_1 and p_2, then $\mathbb{P}[A \cap B] = p_1p_2$, 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

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

- **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
  - $D(P_{XY}||P_XP_Y)$ = 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**)
   $$I(X_1,\ldots,X_n;Y) = \sum_{i=1}^{n} I(X_i;Y|X_1,\ldots,X_{i-1})$$
   $$= I(X_1;Y) + I(X_2;Y|X_1) + \ldots$$
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**)
   $$I(X_1,\ldots,X_n;Y_1,\ldots,Y_n) \leq \sum_{i=1}^{n} I(X_i;Y_i)$$
   if $(Y_1,\ldots,Y_n)$ are conditionally independent given $(X_1,\ldots,X_n)$, and Y_i depends on $(X_1,\ldots,X_n)$ 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.
  - $x_1,\ldots,x_n$ can be always uniquely identified from the string $C(x_1)\ldots C(x_n)$
- $C(\cdot)$ 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

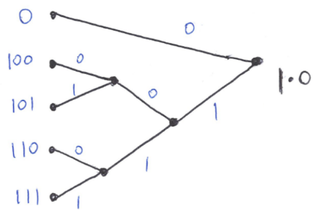
$$\ell(x) = \left\lceil \log_2 \frac{1}{P_X(x)} \right\rceil$$

- **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): $H(X) \leq L(C) < H(X) + 1$



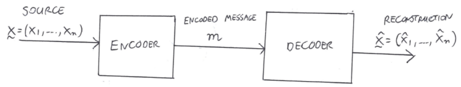
- extension: using blocks of n letters; Huffman coding with $\mathcal{X}^n$
  $nH(X) \leq L(C) < nH(X) + 1$
  $\Rightarrow H(X) \leq$ avg. length per symbol $\leq H(X) + \frac{1}{n}$
  - ✓ exploits *memory*, better guarantee (even independent)
  - × but it's harder to accurately know $P_{X_1\ldots X_n}$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

other codes

- **arithmetic codes** - encodes a sequence $(x_1,\ldots,x_n)$ to at most $\ell(x_1,\ldots,x_n) \leq \log_2 \frac{1}{P_{X_1,\ldots,X_n}(x_1,\ldots,x_n)} + 2$
  - 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 $X_1,\ldots,X_n$
  - map a string to some integer $m \in \{1,\ldots,M\}$
- **discrete memoryless source** $(X_1,\ldots,X_n)$
  - *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 $\underset{\sim}{\hat{X}} = g(m)$ (in $\mathcal{X}^n$)
- **error** $\to$ occurs if $\hat{X} \neq X$
  - $P_e = \mathbb{P}[\underset{\sim}{\hat{X}} \neq \underset{\sim}{X}] = \sum_{\underset{\sim}{x}:\text{DEC(ENC}(x))\neq x} P_{\underset{\sim}{X}}(\underset{\sim}{x})$
- **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 $\mathbf{X} = (X-1,\ldots,X_n)$, let $P_X(x) = \Pi_{i=1}^{n} P_X(x_i)$ be the pmf of X.

typical set, $\mathcal{T}_n(\epsilon) =$
$$\left\{x \in \mathcal{X}^n : 2^{-n(H(X)+\epsilon)} \leq P_X(x) \leq 2^{-n(H(X)-\epsilon}\right\}$$
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) $m \in \{1,\ldots,M\}$ 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**) $x \in \mathcal{T}_n(\epsilon) \iff$
   $$H(X) - \epsilon \leq \frac{1}{n}\sum_{i=1}^{n}\log_2\frac{1}{P_X(x_i)} \leq H(X) + \epsilon$$
   where x_i is the i-th entry of x
   - $\mathbb{E}[\log P_X(x_i)] = H(X_i) = H(X)$
2. $\mathbb{P}[X \in \mathcal{T}_n(\epsilon)] \to 1$ 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 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 $n \to \infty$