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

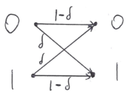
information theory concepts

- 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



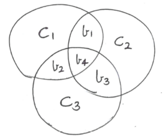
- each bit is flipped independently with probability $\delta \in (0, \frac{1}{2})$

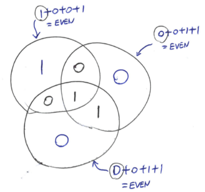
naive

- **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 $b_1 b_2 b_3 b_4$ 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

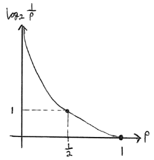
information of an event

- **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 $\mathbf{p} = (p_1, \ldots, p_N)$. We consider a general information measure of the form
$$\Phi(\mathbf{p}) = \Phi(p_1, \ldots, p_N)$$
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**) $\Psi(p_1, \ldots, p_N) = \Psi(p_1+p_2, p_3, \ldots, p_N) + (p_1+p_2)\Psi(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2})$

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.
   - general chain rule:
     $H(X_1, \ldots, X_n) = \sum_{i=1}^n H(X_i | X_1, \ldots, X_{i-1})$
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. $H(X_1, \ldots, X_n) \leq \sum_{i=1}^n H(X_i)$ (**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 $x - 1 > \ln x$)
- $D(P||Q) = 0 \iff P = Q$
  - *Proof.* same as above, using $\ln a = a - 1 \iff a = 1$ (then $\frac{P(x)}{Q(x)} = 1$)

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** $\rightarrow C(x) \neq C(x') \iff x \neq x'$
- a code $C(\cdot)$ is **uniquely decodable** $\rightarrow$ 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** $\rightarrow$ 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 $\sum_{x \in \mathcal{X}} 2^{-\ell(x)} \leq 1$

- *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** $\rightarrow$ if a 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 $\iff P_X(x) = 2^{-\ell(x)}$

- 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 $L(C) - H(X) \geq D(P_X || Q) \geq 0$