Classical Information Theory / Von Neumann Entropy / Entropy Inequalities

References:
1. "Elements of Information Theory" by Thomas Cover
2. Lecture notes from
   - Stanford University, COMP-761 "Quantum Information Theory" by Patrick Hayden
   - Caltech, Ph219, "Quantum Computation" by John Preskill

Classical Information Theory

Shannon (1948):

- Data compression (source coding)
- Transmission over noisy medium (channel coding)

Information theory is the study of these two questions in varying setups.

! We can put theoretical bounds on how good we can do.

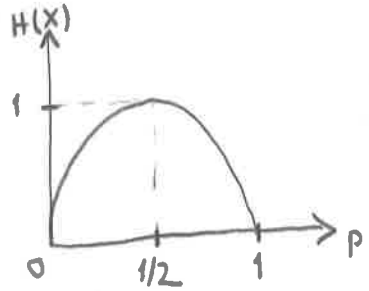
A key insight: Entropy is the appropriate tool to describe the uncertainty in some data.

Defn: Let X be a r.v taking values in a (finite) set $\mathcal{X}$ (alphabet) with prob. distribution $p(x)$. Then (Shannon) entropy of X is:

$$H(X) = -\sum_x p(x) \log p(x)$$

e.g. take a classical bit with $p(X=1) = p$, $p(X=0) = 1-p$.

$H(X) = -p \log p - (1-p)\log(1-p)$ -- take log in base 2.



Information ~ "that which reduces uncertainty"
or in this context average # of questions (w/ binary ans.) one needs to ask to determine a message

$H(X)$ answers how much information one needs to characterize a message $X_1 \dots X_n$ with iid $X_i \sim p(X)$ as $n \to \infty$

- $p=0$ or $p=1 \rightsquigarrow H(X)=0$: we don't need to store any data to characterize messages drawn from this prob. dist.
- $p=1/2 \rightsquigarrow H(X)=1$: 'You need to be told every next bit'

Claim: $H(X)$ characterizes the optimal rate of bits for other p as well.

- Consider the message $X_1 \dots X_n$ $\frac{1}{n}\sum_{i=1}^{n} X_i$ tends to a Gaussian around p with variance tending to zero as $n \to \infty$ (law of large numbers)
- With 'high probability' a message consists of np 1's and $n(1-p)$ 0's. ~ such messages are called 'typical'

# of such messages

$$\binom{n}{np} = 2^{\log\binom{n}{np}} \simeq 2^{n\log n - n - [np\log(np) - np + n(1-p)\log(n(1-p)) - n(1-p)]}$$
$$= 2^{nH(X)} \quad \left(\text{and each of these messages are equally likely with prob. } \sim 2^{-nH(X)}\right)$$

- Take $nH(X)$ bit long messages and encode each of the high prob. sequences $X_1 \dots X_n$ with these new messages. As $n \to \infty$ that is enough.

→ any rate $R > H(X)$ still works (can still encode all typical messages)

→ for any $R < H(X)$ we will have to leave some typical messages out and the prob. of successful lossless compression $\to 0$.

$\boxed{H(X) \geq 0}$ and $\boxed{H(X) \leq \log|\mathcal{X}|}$

↳ with eql. iff X is deterministic ↳ with eql. iff X is uniform over $\mathcal{X}$.

Conditional entropy and mutual information

Conditional entropy: "Uncertainty in X given we know the outcome of Y"

$$H(X|Y) = -\sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(y)} \xrightarrow{\text{multiply \& divide by } p(y)} \sum_y p(y)\, H(X|Y=y) \geq 0$$

$$\Rightarrow \boxed{H(Y) \leq H(X,Y)}$$

$$H(X,Y) - H(Y)$$

- Uncertainty increases with more random variables
- This is one important difference with quantum information where $H(X|Y) < 0$ is possible.

$\boxed{H(X|X) = 0}$

Mutual information (a measure of how correlated two sources are)

$$\boxed{I(X;Y) \equiv H(X) - H(X|Y)} \quad \text{--- how much uncertainty in } X \text{ is reduced when we know } Y.$$

$$= H(X) + H(Y) - H(X,Y) \quad \text{--- implies } I(X,Y) = I(Y;X)$$

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Seeing it as the remaining uncertainty it is intuitively clear that it should never be negative. In fact, using the convexity of log one can show $\boxed{I(X;Y) \geq 0}$

- If X and Y are independent knowing Y shouldn't reduce uncertainty in X and in fact $p(x,y) = p(x)p(y) \Rightarrow I(X;Y) = 0$.

Relative Entropy (A measure of closeness btw. two prob distributions)

$$D(p\|q) \equiv \sum_x p(x) \log \frac{p(x)}{q(x)}$$

Thm: $D(p\|q) \geq 0$ with eql iff $p = q$.

pf: $D(p\|q) = -\sum_x p(x) \log \frac{q(x)}{p(x)} \geq -\sum_x p(x)\left(\frac{q(x)}{p(x)} - 1\right) = 0$

($\log x \leq x-1$ with eql iff $x=1$)

Take log as ln in the pf.

- $D(p\|q)$ is not sym. in p & q. What does it mean?

  Let us have a prob. distribution $q(x)$ and draw m events iid $\sim q(x)$.

  let p be defined by # of times x appear $/m$. The prob. of some particular p to appear as such at m^{th} trial is $\sim e^{-m D(p\|q)}$.

  ex: Let q be a all heads distribution $\{1,0\}$ and p be a fair coin distribution $\{1/2, 1/2\}$

  $\rightsquigarrow D(p\|q) = \infty$ ... we can never draw a half heads half tails trial from an all heads distribution

  $\dashrightarrow D(q\|p) = \log 2$ ... there is a finite (but vanishing as $m \to \infty$) prob. that we obtain all heads trials from a fair coin

  This also illustrates the meaning of asymmetry under $p \leftrightarrow q$

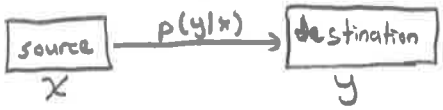
- For $u(x)$ the uniform distribution over $\mathcal{X}$

$$D(p\|u) = \sum_x p(x) \log \frac{p(x)}{1/|\mathcal{X}|} = \log|\mathcal{X}| - H(X) \geq 0$$

$$D(p(x,y)\,\|\,p(x)p(y)) = I(X;Y) \geq 0$$

Discrete memoryless channel capacity (Shannon noisy channel coding thm).

A discrete memoryless channel consists of $(\mathcal{X}, p(y|x), \mathcal{Y})$

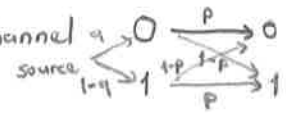


Achievable rate $R \equiv$ a rate of communication through this channel for which we can asymptotically achieve zero error probability transmission.

Capacity $C \equiv \sup_{\text{achievable rates}} R$ $\dashrightarrow$ that it can be nonzero is nontrivial.

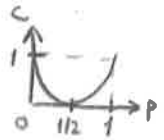
$$C = \sup_{p(x)} I(X;Y)$$

ex: binary sym. channel



$p(x=0, y=0) = pq$.

Optimize over q: $C = 1 - H(p)$



Von Neumann Entropy

Let ρ be a density matrix on some Hilbert space $\mathcal{H}$. Then $\boxed{S(\rho) \equiv -\mathrm{tr}(\rho \log \rho)}$

Diagonalizing ρ with an orthonormal basis $\rho = \sum_i p_i |e_i\rangle\langle e_i|$ (where necessarily $\sum_i p_i = 1$)

Then $S(\rho) = -\sum_i p_i \log p_i = H(p_i)$

↳ Shannon entropy corresponding to p_i

- If ρ is pure $\rho = |\psi\rangle\langle\psi|$ then $S(\rho) = 0$
- U is unitary $\Rightarrow S(\rho) = S(U\rho U^\dagger)$.
- $\boxed{S(\rho) \geq 0}$, $\boxed{S(\rho) \leq \log \dim \mathcal{H}}$
- Suppose we draw pure states from an ensemble $\{|\psi_i\rangle, p_i\}$, then $H(p_i) \geq S(\rho)$ with eql. iff $|\psi_i\rangle$'s are orthogonal.... distinguishability is lost

Quantum generalization of relative entropy: given two density matrices ρ, σ defined on $\mathcal{H}$. β (free energy)

$\boxed{D(\rho\|\sigma) \equiv \mathrm{tr}(\rho(\log\rho - \log\sigma))}$ ... when $\sigma_\beta = e^{-\beta H}$ we have $D(\rho\|\sigma_\beta) = \beta\langle E\rangle_\rho - S(\rho)$

- If $[\rho,\sigma]=0$ $\longrightarrow$ $\rho = \sum_i p_i |e_i\rangle\langle e_i|$, $\sigma = \sum_i q_i |e_i\rangle\langle e_i|$ $\longrightarrow$ $D(\rho\|\sigma) = D(p\|q)$ (classical version)

Thm (Klein's ineql.) $\boxed{D(\rho\|\sigma) \geq 0}$ with eql. iff $\rho = \sigma$. (works when $[\rho,\sigma] \neq 0$ as well)

! Difference with classical information mainly comes when we consider systems with several subsystems (remember entanglement)

Consider a density matrix ρ_{AB} (→ generalizing $p(A,B)$) on the tensor product Hilbert space $H_A \otimes H_B$.

Reduced density matrix $\boxed{\rho_A = \mathrm{tr}_{\mathcal{H}_B} \rho_{AB}}$ ↳ generalizing $p(A)$

Define mutual information $I(A:B) = S(\rho_A) + S(\rho_B) - S(\rho_{AB})$.

- Subadditivity $\boxed{S(\rho_A) + S(\rho_B) \geq S(\rho_{AB})}$ or $I(A:B) \geq 0$ (just as in the classical case) $I(A,B) \geq 0$ or $H(A) \geq H(A|B)$

  Pf: $I(A:B) = D(\rho_{AB} \| \rho_A \otimes \rho_B)$, then use Klein's ineql.

- Monotonicity of Relative Entropy: $\boxed{D(\rho_{AB}\|\sigma_{AB}) \geq D(\rho_A\|\sigma_B)}$ ---- pf. from convexity of $D(\cdot\|\cdot)$, Lieb's concavity,...

- Strong subadditivity: $\boxed{S(\rho_B) + S(\rho_{ABC}) \leq S(\rho_{AB}) + S(\rho_{BC})}$

  Pf: from monotonicity: $D(\rho_{ABC} \| \rho_A \otimes \rho_{BC}) \geq D(\rho_{AB} \| \rho_A \otimes \rho_B)$

- Araki-Lieb ineql. (Triangle ineql.) $\boxed{S(\rho_{AB}) \geq |S(\rho_A) - S(\rho_B)|}$

This is significantly different than the lower bound we found in the classical context: $H(X,Y) \geq \max(H(X), H(Y))$ ... following from $H(X|Y), H(Y|X) \geq 0$.

In fact consider $\rho_{AB} = |\Psi\rangle\langle\Psi| \rightsquigarrow S(\rho_{AB}) = 0$

$\longrightarrow$ if $|\Psi\rangle$ is entangled $S(\rho_A) = S(\rho_B) > 0$.

Use Schmidt decomp. to show this

- Can prove with a trick called 'purification'. Diagonalize ρ_{AB} as $\rho_{AB} = \sum p_m |e_m\rangle\langle e_m|$

Then using an auxiliary Hilbert space C form the pure state

$$|\Psi\rangle = \sum_m \sqrt{p_m}\, |e_m\rangle \otimes |\psi_m\rangle$$

$\hookrightarrow \{|\psi_m\rangle\}$ are orthonormal

and consider $\rho_{ABC} = |\Psi\rangle\langle\Psi|$ ... ρ_{AB} obtained from partial tracing is the same ρ_{AB} we had at the beginning.

$S(\rho_{ABC}) = 0$ and $S(\rho_{AB}) = S(\rho_C)$, $S(\rho_{BC}) = S(\rho_A)$, $S(\rho_{AC}) = S(\rho_B)$

Now use subadditivity $S(\rho_B) + \underbrace{S(\rho_C)}_{=S(\rho_{AB})} \geq \underbrace{S(\rho_{BC})}_{=S(\rho_A)}$ ... similarly for $A \leftrightarrow B$.