Classical Information Theory / Von Neumann Entropy /

Entropy Inequalities

References:
1. "Elements of Information Theory" by Thomas Cover
2. Lecture notes from Stanford University, COMP-261 "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.

Definition: Let $X$ be a r.v. taking values in a (finite) set $X$ (alphabet) with probability 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 $X_1 \ldots X_n$ with iid $X_i \sim p(X)$ as $n \to \infty$.

- $p = 0$ or $p = 1 \implies H(X) = 0$: we don't need to store any data to characterize messages drawn from this prob. dist.
- $p = 1/2 \implies H(X) = 1$: 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 $X_1 \ldots X_n$, $\frac{1}{n} \sum_{i=1}^{n} X_i$ 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'.

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

- Take $nH(X)$ bit long messages and encode each of the high prob. sequences $X_1 \ldots X_n$ 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 $\to 0$.

\[
H(X) \geq 0 \quad \text{and} \quad H(X) \leq \log |X|
\]

- with eq. iff $X$ is deterministic

- with eq. iff $X$ is uniform over $X$.

Conditional entropy and mutual information

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

\[
H(X|Y) = - \sum_{x,y} p(x,y) \log \frac{p(x|y)}{p(y)} = \underbrace{\sum_{y} p(y) H(X|Y=y)}_{\geq 0} \implies H(Y) \leq H(X,Y)
\]

- Uncertainty increases with more random variables

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

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

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

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

- If $X$ and $Y$ are independent knowing $Y$ shouldn't reduce uncertainty in $X$ and in fact $p(x,y) = p(x)p(y) \implies I(X;Y) = 0$.
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 eqn. iff \( p=q \) \( \log_x \) x \( = 1 \) with eqn. if \( x \neq 1 \)

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

Take log as \( \ln \) in the pt.

- \( D(p \| q) \) 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 \( \sim q(x) \). The prob. of some particular \( p \) to appear as such at \( m \)th trial is \( \sim e^{-mD(p \| q)} \).

**Ex:** Let \( q \) be a all heads distribution \( \{1,0\} \) and \( p \) be a fair coin distribution \( \{1/2, 1/2\} \)

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

\( \rightarrow D(q \| p) = \log 2 \ldots \) 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 \( p \to q \).

- For \( u(x) \) the uniform distribution over \( X \)

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

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

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

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

\[ \text{source} \xrightarrow{\text{p}(y|x)} \text{destination} \]

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

Capacity \( C \equiv \sup_{P} R \rightarrow \) that it can be nonzero at all rates.

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

**Ex:** binary sym. channel \( \rightarrow 0 \xrightarrow{p} 0 \)

\[ p(x=0, y=0) = pq. \]

Optimize over \( q : C = 1 - H(p) \)
Von Neumann Entropy

Let $\rho$ be a density matrix. Then
$$S(\rho) \equiv -\text{tr}(\rho \log \rho)$$

Diagonalizing $\rho$ with an orthonormal basis $\rho = \sum P_i |\psi_i\rangle \langle \psi_i|$,\ (where necessarily $\sum P_i = 1$)

Then $S(\rho) = -\sum P_i \log P_i = H(P_i)$

- If $\rho$ is pure $\rho = \ket{\psi}\bra{\psi}$ then $S(\rho) = 0$
- $U$ is unitary $\Rightarrow S(\rho) = S(U\rho U^\dagger)$
- $S(\rho) \geq 0$, $S(\rho) \leq \log \text{dim} \mathcal{H}$
- Suppose we draw pure states from an ensemble $\{\ket{\psi_i}, P_i\}$, then $H(P_i) \geq S(\rho)$ with eqn. if $\psi_i$'s are orthogonal... distinguishability is lost.
- Quantum generalization of relative entropy:

$$D(\rho \|\sigma) \equiv -\text{tr}(\rho \log \sigma)$$

when $\sigma = e^{-\beta H}$ we have $D(\rho \|\sigma) = \beta \langle E \rangle - S(\rho)$

- If $[\rho, \sigma] = 0 \Rightarrow \rho = \sum P_i |\psi_i\rangle \langle \psi_i|, \sigma = \sum Q_i |\phi_i\rangle \langle \phi_i|$

Then (Klein's ineq.) $D(\rho \|\sigma) \geq 0$ with eqn. if $\rho = \sigma$. (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 $\rho_{AB}$ on the tensor product Hilbert space $\mathcal{H}_A \otimes \mathcal{H}_B$.

Reduced density matrix $\rho_A = \text{tr}_{\mathcal{H}_B} \rho_{AB}$

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

- Subadditivity: $S(\rho_A) + S(\rho_B) \geq S(\rho_{AB})$ or $I(A;B) \geq 0$
  - Pf: $I(A;B) = D(\rho_{AB} \| \rho_A \otimes \rho_B)$, then use Klein's ineq.

- Monotonicity of Relative Entropy: $D(\rho_{AB} \| \rho_{AB}) \geq D(\rho_A \| \rho_B)$

- Strong subadditivity: $S(\rho_B) + S(\rho_{ABC}) \leq S(\rho_A) + S(\rho_{BC})$
  - Pf: from monotonicity $D(\rho_{ABC} \| \rho_A \otimes \rho_{BC}) \leq D(\rho_{AB} \| \rho_A \otimes \rho_B)$
Araki-Lieb ineq. (Triangle ineq.) \[ S(g_{AB}) \geq |S(g_A) - S(g_B)| \]

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 \( g_{AB} = |\psi \rangle \langle \psi| \rightarrow S(g_{AB}) = 0 \)

\[ \rightarrow \text{if } |\psi \rangle \text{ is entangled } S(g_A) = S(g_B) > 0. \]

- Can prove with a trick called 'purification'. Diagonalize \( g_{AB} \) as \( g_{AB} = \sum P_m |e_m \rangle \langle e_m| \)

Then using an auxiliary Hilbert space \( C \) form the pure state
\[ |\psi \rangle = \sum_m \sqrt{P_m} |e_m \rangle \otimes |\psi_m \rangle \]

\( |e_m \rangle \) are orthonormal

and consider \( g_{ABC} = |\psi \rangle \langle \psi| \) \( g_{AB} \) obtained from partial tracing is the same \( g_{AB} \) we had at the beginning.

\[ S(g_{ABC}) = 0 \text{ and } S(g_{AB}) = S(g_C), S(g_{BC}) = S(g_A), S(g_{AC}) = S(g_B) \]

Now use subadditivity \( S(g_B) + S(g_C) \geq S(g_{BC}) \) ... similarly for \( A \leftrightarrow B \).

\[ = S(g_{AB}) = S(g_A) \]