Information Theory and Disjointness Lower Bounds

Dean Hirsch and Oliver Korten
Columbia University

March 30, 2022

Intro

- Cover basics of information theory:
  1. Entropy
  2. Mutual Information
  3. Important Properties
- Use information theory to prove an $\Omega(n)$ lower bound on the randomized complexity of disjointness.

Entropy

The fundamental concept in information theory is "entropy." For a random variable X, the entropy of X (denoted $H(X)$) is a measure of the "information content" of a typical sample from X.

How Should We Define Entropy?

It's not immediately clear what the proper definition of such a concept is, but it should satisfy some key properties, for example:

1. Should be nonnegative (reading a message, or viewing a sample, should never decrease your total knowledge)
2. For independent random variables X, Y, we should have $H(X) + H(Y) = H(XY)$, i.e. the information gained from a pair of independent sources should equal the sum of the information gained from each.
3. Should only depend on the distribution of X and not the specific values it takes on.

Entropy Definition

It was discovered by Shannon that there is a unique measure satisfying these (and a few other) requirements, which he called entropy:

$$H(X) = \sum_x p(x) \log \frac{1}{p(x)}$$

Here, we take $p \log \frac{1}{p} = 0$ when $p = 0$, which agrees with the limit as $p \to 0$.

Basic Properties of Entropy

1. Entropy is oblivious to the specific universe of elements and only depends on the distribution
2. Let X be a distribution over a finite set U.
   1. $H(X) \geq 0$, with equality if $p(x) = 1$ for some $x \in U$.
   2. $H(X) \leq \log |U|$, with equality if X is the uniform distribution over U.

Jensen's Inequality

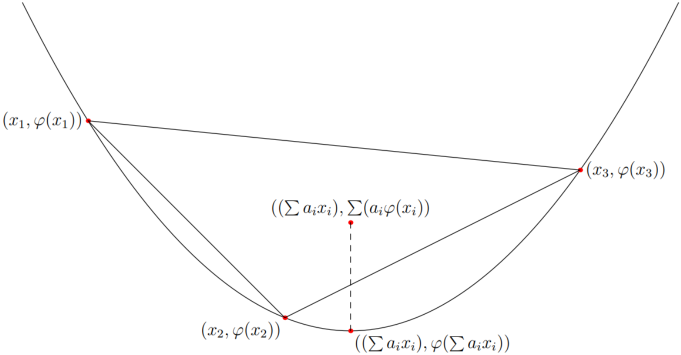
A basic lemma used again and again in information theory is Jensen's inequality:

Theorem

If $\phi : \mathbb{R} \to \mathbb{R}$ *is a concave function and* X *is a real-valued random variable, then* $\phi(\mathbb{E}[X]) \geq \mathbb{E}[\phi(X)]$*.*

By symmetry the same inequality holds in the other direction if ϕ is instead assumed to be convex. The above is often applied in information theory in the setting $\phi(x) = \log x$, which can be readily seen to be concave.

Proof of Jensen's Inequality


$(x_1, \varphi(x_1))$
$(x_3, \varphi(x_3))$
$((\sum a_i x_i), \sum(a_i \varphi(x_i))$
$(x_2, \varphi(x_2))$
$((\sum a_i x_i), \varphi(\sum a_i x_i))$

Entropy is at most log of Support

Again let X be a distribution over some finite set U. Then:

$$\begin{aligned} H(X) &= \sum_{x \in U} p(x) \log \frac{1}{p(x)} \\ &= \mathbb{E}_{x \sim X} \left(\log \frac{1}{p(x)} \right) \\ &\leq \log \left(\mathbb{E}_{x \sim X} \frac{1}{p(x)} \right) \\ &= \log \left(\sum_{x \in U} p(x) \frac{1}{p(x)} \right) = \log |U| \end{aligned}$$

So if X is a distribution over n-bit strings, then $H(X) \leq n$.

Operational Interpretation of Entropy

As it turns out, the entropy of a random variable X can be defined (up to a ± 1 error) in purely operational terms as follows:

Theorem

Let X be a random variable taking on values in some universe U, and let $Q(X)$ be the minimum, over all uniquely decodable encodings $U \to \{0,1\}^*$, of the expected code-length of of a sample from X.
Then we have $H(X) \leq Q(X) \leq H(X) + 1$.

Conditional Entropy

When we have two jointly distributed random variable X, Y, we would like some way of quantifing the information entropy that remains in X when we know Y. We call this the conditional entropy, denoted $H(X|Y)$, defined as:

$$H(X|Y) = \mathbb{E}_{y \sim Y} H(X|Y = y)$$

In other words, its the expected entropy of X conditioned on a particular value of y, when y is sampled from Y.

Mutual Information

In many situations we will have multiple jointly distributed random variables and it will be useful to measure the amount of information that is shared between them. For random variables X, Y, we will use $I(X : Y)$ ("mutual information between X, Y") to quantify this.

1. $I(X : Y)$ should capture the amount of information gained on X by knowing Y.
2. When X, Y are independent, we should have $I(X : Y) = 0$. If $X = Y$, we should have $I(X : Y) = H(X) = H(Y)$.

Defining Mutual Information

We will define mutual information as follows for jointly distributed X, Y:

$$I(X:Y) = \mathbb{E}_{x,y\sim XY}\left(\log \frac{p(x,y)}{p(x)p(y)}\right)$$

We now show the following alternative characterization:

$$I(X:Y) = H(X) + H(Y) - H(XY)$$

Note that mutual information is symmetric.

Mutual Information

$$
\begin{aligned}
I(X:Y) &= \mathbb{E}_{x,y\sim XY}\left(\log\frac{p(x,y)}{p(x)p(y)}\right)\\
&= \mathbb{E}_{x,y\sim XY}\left(\log\frac{1}{p(x)}+\log\frac{1}{p(y)}-\log\frac{1}{p(x,y)}\right)\\
&= H(X)+H(Y)-H(XY)
\end{aligned}
$$

Mutual Information is Nonnegative

$$\mathbb{E}_{x,y\sim XY} \log \frac{p(x,y)}{p(x)p(y)} = -\mathbb{E}_{x,y\sim XY} \log \frac{p(x)p(y)}{p(x,y)}$$

$$\geq -\log\left(\mathbb{E}_{x,y\sim XY} \frac{p(x)p(y)}{p(x,y)}\right) = -\log\left(\sum_{x,y} p(x)p(y)\right)$$

$$= -\log 1 = 0$$

Conditional Mutual Information

Analogously to entropy, we will define $I(X : Y|Z)$ to be the expectation over $z \sim Z$ of $I(X : Y|Z = z)$:

$$I(X : Y|Z) = \mathbb{E}_{z \sim Z} \left(I(X|Z = z : Y|Z = z) \right)$$

Chain Rules

Chain rules allow us to break down a joint distribution into a marginal and a conditional part. The basic chain rule of probabilities says:

$$P(x, y) = P(x)P(y|x)$$

This simple fact will allow us to similarly break down the various information measures into such a nice form.

Chain Rule for Entropy

For entropy of joint distributions we have the following:

$$H(XY) = H(X|Y) + H(Y)$$

Intuitively, this says that the entropy of XY is equal to the entropy in X plus the entropy that remains in Y once you know X.

Chain Rule for Entropy

$$
\begin{aligned}
H(X|Y) &= \sum_y p(y) H(X|y) \\
&= \sum_y p(y) \left(\sum_x p(x|y) \log \frac{1}{p(x|y)} \right) \\
&= \sum_{x,y} p(x,y) \log \frac{1}{p(x|y)} \\
&= \sum_{x,y} p(x,y) \log \frac{p(y)}{p(x,y)} \\
&= \sum_{x,y} p(x,y) \log \frac{1}{p(x,y)} - \sum_{x,y} p(x,y) \log \frac{1}{p(y)} \\
&= H(XY) - H(Y)
\end{aligned}
$$

Chain Rule for Entropy

Recall that we showed before that:

$$I(X : Y) = H(X) + H(Y) - H(XY)$$

Combining this with the chain rule for entropy we get an alternative expression:

$$I(X : Y) = H(X) - H(X|Y)$$

So the information between X and Y is the amount of uncertainty about X which you eliminate by knowing Y (and vice versa).

Chain Rule for Mutual Information

For mutual information we have the following chain rule:

$$I(XY : Z) = I(X : Z) + I(Y : Z|X)$$

Subadditivity

As we have hinted at, the various information quantities satisfy several nice inequalities, whereby conditioning on some additional information can only increase or decrease some partiticular measure. We will refer to such inequalities as "subaditivity" inequalities.

Subadditivity of Entropy

For entropy we have the following natural inequality:

$$H(AB) \leq H(A) + H(B)$$

This follows from the identity $I(A : B) = H(A) + H(B) - H(AB)$, and the fact from before that mutual information is always nonnegative.

Conditioning Never Increases Entropy

Combining the chain rule and subadditivity we have:

$$H(Y) + H(X|Y) = H(XY) \leq H(Y) + H(X)$$

So in particular, $H(X|Y) \leq H(X)$.

Disjointness Problem Refresher

Alice has a vector $u \in \{0,1\}^n$ and Bob has a vector $v \in \{0,1\}^n$. They want to compute the DISJ function, defined by

$$DISJ(u,v) = \begin{cases} 0, & \text{If } \exists i : u_i = v_i = 1 \\ 1, & \text{otherwise} \end{cases}$$

We will also refer to the equivalent task of INT instead of DISJ.

Disjointness Lower Bound

Theorem

Any randomized protocol that computes the disjointness function with error $\leq \frac{1}{2} - \varepsilon$ *must have communication* $\Omega(\varepsilon^2 n)$*.*

By repeating $\Theta(\frac{1}{\varepsilon^2})$ times, this is equivalent to

Theorem

Any randomized protocol that computes the disjointness function with error $\leq \frac{1}{100}$ *must have communication* $\Omega(n)$*.*

Disjointness Lower Bound - Obstacles

- Hard distribution?
- Uniform is not - disjointness probability $(3/4)^n$.
- Neither does any product distribution. But we can get $\Omega(\sqrt{n})$ lower bound with it, but also have $O(\sqrt{n}\log n)$ upper bound.
- Easy upper bound of $\tilde{O}(n^{2/3})$ for product distributions:
    1. Fix $\varepsilon = n^{-1/3}$.
    2. Alice sends coordinates with $H(A_i) \leq \varepsilon$, needs $\leq \varepsilon n$ bits.
    3. Bob sends coordinates with $H(B_i) \leq \varepsilon$, needs $\leq \varepsilon n$ bits.
    4. Exchange $\frac{1}{\varepsilon^2}$ coordinates where $H(A_i), H(B_i) > \varepsilon$.

Disjointness Lower Bound - Proof

- Considering distribution ζ over $\{0,1\}^2 \times \{a,b\}$.
    - Let D be uniform on $\{a,b\}$ and Z be uniform on $\{0,1\}$.
    - If $D = a$ then $(A,B) \sim (0,Z)$.
    - If $D = b$ then $(A,B) \sim (Z,0)$.
- Let $((A,B),D) \sim \zeta^{\otimes n}$. Properties:
    1. A, B are dependent.
    2. A, B are independent if conditioning on D.
    3. A, B are always disjoint.

Disjointness Lower Bound - Proof

Reduction to information: study $I(AB; T(A,B)|D)$, where $T(A,B)$ is the transcript.
Interesting because $I(AB; T(A,B)|D) \leq H(T) \leq |\text{comm}|$.
Will be done if we show $I(AB; T(A,B)|D) \geq \Omega(n)$.
Split to bits:

$$I(AB; T|D) \geq \sum_{i=1}^{n} I(A_i B_i; T|D)$$

Disjointness Lower Bound - Proof

Split to bits:

$$I(AB; T|D) \geq \sum_{i=1}^{n} I(A_iB_i; T|D)$$

Proof:

$$\begin{aligned}
&I(AB; T|D) \\
&= H(AB|D) - H(AB|TD) \\
&= \sum_{i=1}^{n} H(A_iB_i|D) - H(AB|TD) \\
&\geq \sum_{i=1}^{n} H(A_iB_i|D) - \sum_{i=1}^{n} H(A_iB_i|TD) \\
&= \sum_{i=1}^{n} I(A_iB_i; T|D)
\end{aligned}$$

Reduction Lemma

We reduce the problem to lower bounding the information obtained by a protocol computing AND of two bits.

Lemma (Reduction Lemma)

For any $i \in [n]$:

$$I(A_i B_i; T(A,B)|D) \geq \inf_P I(UV; P(U,V)|D)$$

where the infimum is over protocols P computing AND_2 (AND of two bits) with error $\leq \frac{1}{100}$. In RHS, $((U,V),D) \sim \zeta$.

Reduction Lemma - Proof

$$I(A_iB_i; T(A,B)|D) \geq \inf_P I(UV; P(U,V)|D)$$

Proof idea: construct protocol for AND$_2$ using T.
By definition of mutual information:

$$I(A_iB_i; T(A,B)|D) = \underset{d\sim\{a,b\}^{n-1}}{\mathbb{E}}[I(A_iB_i; T(A,B)|D_i, D_{-i} = d)]$$

For each fixing $D_{-i} = d$, the following is a protocol P for AND, given two input bits x, y:

1. Use this D_{-i}. Alice draws A_{-i}, Bob draws B_{-i}.
2. Set $A_i = x$ and $B_i = y$.
3. Run the INT protocol on these A, B.

Reduction Lemma - Proof Cont.

$$I(A_iB_i;T(A,B)|D) = \mathop{\mathbb{E}}_{d\sim\{a,b\}^{n-1}}[I(A_iB_i;T(A,B)|D_i,D_{-i}=d)]$$

Protocol P:

1. Use this D_{-i}. Alice draws A_{-i}, Bob draws B_{-i}.
2. Set $A_i = x$ and $B_i = y$.
3. Run the INT protocol on these A, B.

Now $(U, V, D, P(U, V)) \sim (A_i, B_i, D_i, T(A, B))$ conditioned on $D_{-i} = d$. So

$$I(A_iB_i;T(A,B)|D_i,D_{-i}=d) = I(UV;P(U,V)|D)$$

□

A Corollary

Corollary

$$\mathbf{BPP}(\mathrm{DISJ}) \geq n \cdot \inf_P I(UV; P(U,V)|D).$$

It remains to prove that $\inf_P I(UV; P(U,V)|D) > 0$. In other words, that there's $\kappa > 0$ such that $I(UV; P(U,V)|D) > \kappa$ for all P solving AND_2 with high probability.

Manipulation

We want $I(UV; P(U,V)|D) > \kappa > 0$.
By definition, letting Z be uniformly random from $\{0,1\}$:

$$\begin{aligned} & I(UV; P(U,V)|D) \\ & = \frac{1}{2} I(UV; P(U,V)|D = a) + \frac{1}{2} I(UV; P(U,V)|D = b) \\ & = \frac{1}{2} I(Z; P(0,Z)) + \frac{1}{2} I(Z; P(Z,0)) \end{aligned}$$

(No more D!)
For fixed $(u,v) \in \{0,1\}^2$ let p_{uv} be the distribution over $P(u,v)$. We will now relate the mutual informations above to distance between these distributions.

Hellinger Distance

We present a new notion of distance between distributions p, q:

Definition (Hellinger Distance)

For two distributions p, q over domain X, the squared Hellinger distance $h^2(p, q)$ is

$$h^2(p,q) = \frac{1}{2}\sum_{x\in X}(\sqrt{p(x)} - \sqrt{q(x)})^2 = 1 - \sum_{x\in X}\sqrt{p(x)q(x)}$$

Why is this useful?

- $h(p,q) \propto \|\sqrt{p} - \sqrt{q}\|_2$, hence it is a metric over $\sqrt{p}$ vectors. Can use triangle inequality.
- This metric comes from an inner product, so we can use appropriate Cauchy-Schwartz.
- And some other useful properties.

Back on Track

We need to show $I(Z; P(0, Z)) + I(Z; P(Z, 0))$ is bounded from below. We have:

$$I(Z; P(0, Z)) \geq h^2(p_{00}, p_{01})$$
$$I(Z; P(Z, 0)) \geq h^2(p_{00}, p_{10})$$

So their sum is at least

$$h^2(p_{00}, p_{10}) + h^2(p_{00}, p_{10})$$

Let $f = \sqrt{p_{00}} - \sqrt{p_{10}}$, $g = \sqrt{p_{00}} - \sqrt{p_{01}}$, then

$$h^2(p_{00}, p_{10}) + h^2(p_{00}, p_{10}) = \frac{1}{2}(\|f\|^2 + \|g^2\|)$$

$$= \frac{1}{4}(\|f - g\|^2 + \|f + g\|^2) \geq \frac{1}{4}\|f - g\|^2 = \frac{1}{2}h^2(p_{01}, p_{10})$$

Finishing

Enough to prove that $\frac{1}{2}h^2(p_{01}, p_{10})$ is bounded from below.

Lemma (Cut-and-Paste Lemma)

Let P be a randomize protocol over $X \times Y$. Then for every $x, x' \in X$ and every $y, y' \in Y$, we have

$$h(P_{xy}, P_{x'y'}) = h(P_{x,y'}, P_{x',y}).$$

Applying the Cut-and-Paste lemma, we have

$$h^2(p_{01}, p_{10}) = h^2(p_{00}, p_{11})$$

so it is enough to prove $h^2(p_{00}, p_{11})$ is bounded from below. This much is now intuitive.

Finishing - Cont.

The last bit:

Lemma (Distinguishing Lemma)

If P computes a function f with error at most ε on every input, and (x, y) and (x', y') are such that $f(x, y) \neq f(x', y')$, then

$$h^2(p_{xy}, p_{x'y'}) \geq 1 - 2\sqrt{\varepsilon}.$$

Recap

$$
\begin{aligned}
|communication| &\geq H(T(A,B)) \geq I(AB; T(A,B)|D) \\
&\geq \sum_{i=1}^{n} I(A_iB_i; T|D) \geq n \cdot \inf_P I(UV; P(U,V)|D) \\
&= n \cdot \frac{1}{2}(I(Z; P(0,Z)) + I(Z; P(Z,0))) \\
&\geq \frac{n}{2}(h^2(p_{00}, p_{01}) + h^2(p_{00}, p_{10})) \\
&\geq \frac{n}{4} h^2(p_{01}, p_{10}) \\
&= \frac{n}{4} h^2(p_{00}, p_{11}) \\
&\geq \frac{n}{4}(1 - 2\sqrt{\varepsilon})
\end{aligned}
$$

Proof of Cut-and-Paste Lemma

For any transcript P we can decompose

$$\mathbf{Pr}(P(x,y)=T)=q_A(x,T)q_B(y,T)$$

for some functions q_A, q_B. Now:

$$\begin{aligned}
&1-h^2(p_{xy},p_{x'y'})\\
&=\sum_T\sqrt{\mathbf{Pr}(P(x,y)=T)\,\mathbf{Pr}(P(x',y')=T)}\\
&=\sum_T\sqrt{q_A(x,T)q_B(y,T)q_A(x',T)q_B(y',T)}\\
&=\sum_T\sqrt{\mathbf{Pr}(P(x',y)=T)\,\mathbf{Pr}(P(x,y')=T)}\\
&=1-h^2(p_{x',y},p_{x,y'})
\end{aligned}$$

Divergence

In the expression for information $\mathbb{E}_{x,y\sim XY}\left(\log\frac{p(x,y)}{p(x)p(y)}\right)$, we are implicitly quantifying a relation between two distributions over pairs x, y, one being the joint distribution with probabilities $p(x, y)$, and one being the product distribution of marginals with probabilities $p(x)p(x)$. This quantity is more generally referred to as divergence.

For two distributions P, Q over a common domain:

$$(P||Q) = \mathbb{E}_{x\sim P}\log\frac{p(x)}{q(x)}$$

Divergence

We have the following:

$$(P||Q) = \mathbb{E}_{x\sim P} \log \frac{p(x)}{q(x)} = \mathbb{E}_{x\sim P} \log \frac{1}{q(x)} - H(P)$$

Recall that the optimal prefix-free encoding for q assigns x a string of length $\log \frac{1}{q(x)}$. So $\mathbb{E}_{x\sim P} \log \frac{1}{q(x)}$ can be seen as quantifying expected performance (codeword length) of using the optimal code for Q over the distribution P. Thus $(P||Q)$ quantifies the additional number of bits you need (in expectation) to encode samples from P when using an encoding optimized for Q, as opposed to one optimized for P.

Important Properties of Divergence

We see that the definition of divergence is not symmetric, and indeed we can have $(P||Q) \neq (Q||P)$. However an important fact is that it is always nonnegative:

$$\begin{aligned}\mathbb{E}_{x\sim P}\log\frac{p(x)}{q(x)} &= -\mathbb{E}_{x\sim P}\log\frac{q(x)}{p(x)}\\ &\geq -\log\left(\mathbb{E}_{x\sim P}\frac{q(x)}{p(x)}\right)\\ &= -\log 1 = 0\end{aligned}$$

In particular, since mutual information can be expressed as a divergence, it is always nonnegative as well.

Proof of Operational Interpretation

We first show that for any X distributed over a universe U, there is a prefix-free encoding scheme $f : U \to \{0,1\}^*$ such that

$$\mathbb{E}_{x \sim X}|f(x)| \leq H(X) + 1$$

Proof of Operational Interpretation

We will assume without loss of generality that $U = [n]$, and that the elements of U are arranged in decreasing order of probability so that $1 \geq p(1) \geq \cdots \geq p(n) \geq 0$. We will give a prefix-free encoding which assigns each $x \in [n]$ a codeword of length $\ell_x := \lceil \log \frac{1}{p(x)} \rceil$. If this holds, then we have:

$$\begin{aligned} \mathbb{E}_{x \sim X} |f(x)| &= \mathbb{E}_{x \sim X} \ell_x \\ &= \mathbb{E}_{x \sim X} \lceil \log \frac{1}{p(x)} \rceil \\ &\leq \mathbb{E}_{x \sim X} \left(1 + \log \frac{1}{p(x)} \right) = H(X) + 1 \end{aligned}$$

completing the proof.

Proof of Operational Interpretation

To construct such a code, we initialize a complete binary tree of depth n. Now, for each $x \in [n]$ in increasing order of x, we find a node of depth ℓ_x and delete all of its descendants so that this node becomes a leaf. We then give x the codeword specifying the root-to-leaf path of this node, and continue for the next value of x.

So long as we can always find a node of the appropriate depth, each x will be given a codeword of the appropriate length, and no codeword will be a prefix of another since all codewords are leaves.

Proof of Operational Interpretation

It suffices to show that after all the vertex deletions at step x, there is still a remaining vertex of depth ℓ_{x+1}. For $y < x$, the number of vertices of depth ℓ_x deleted at step y is exactly $2^{\ell_x - \ell_y}$. So the number of vertices of depth ℓ_x that are deleted before the x^{th} step is:

$$\sum_{y<x} 2^{\ell_x - \ell_y} = 2^{\ell_x} \sum_{y<x} 2^{-\ell_y} \leq 2^{\ell_x} \sum_{y<x} p(y) < 2^{\ell_x}$$

By definition there are 2^{ℓ_x} leaves of depth ℓ_x to begin with, so one must remain.

Proof of Operational Interpretation

We now show that no encoding can do better than $H(X)$ in expectation. Again, assume that X lies over the universe $[n]$, and say that $x \in [n]$ is given a codeword of length ℓ_x. So then we have:

$$\begin{aligned}
\mathbb{E}_{x\sim X}\left(\ell_x\right) &= \mathbb{E}_{x\sim X}\left(\log\frac{1}{p(x)} - \log(2^{-\ell_x}/p(x))\right)\\
&= H(X) - \mathbb{E}_{x\sim X}\left(\log(2^{-\ell_x}/p(x))\right)\\
&\geq H(X) - \log\left(\mathbb{E}_{x\sim X}\left(2^{-\ell_x}/p(x)\right)\right)\\
&= H(X) - \log\left(\sum_x 2^{-\ell_x}\right)
\end{aligned}$$

If we can show that $\sum_x 2^{-\ell_x} \leq 1$ then we are done.