Introduction to Entropy, Optimal Code-length and Brégman Theorem

Théophile Trunck

JCRAA2012

Juin 2012

Outline

Definition

Definition

The *entropy* of a discrete random variable X is

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x),$$

where $p(x) = P(X = x)$

- Think of entropy as the amount of randomness/surprise in X.
- It is the number of bits required on the average to describe X.
- For example, if $p(x) = 1$ for some x, then $H(X) = 0$.

Example

Consider a random variable X.

- which has a uniform distribution over 8 outcomes.

$$H(X) = -\sum_{i=1}^{8} \frac{1}{8} \log \frac{1}{8} = 3 \text{ bits}$$

- which has the following probabilities distribution over 8 outcomes, $\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\}$

$$\begin{aligned} H(X) &= -\frac{1}{2} \log \frac{1}{2} - \frac{1}{4} \log \frac{1}{4} - \frac{1}{8} \log \frac{1}{8} - \frac{1}{16} \log \frac{1}{16} - 4\frac{1}{64} \log \frac{1}{64} \\ &= 2 \text{ bits} \end{aligned}$$

Remarks

- During this talk $\log = \log_2$ so entropy will be expressed in *bits*.
- Adding terms of probability 0 does not change entropy. ($0 \log 0 = 0$ by continuity).
- Entropy does not depend on the actual values taken by X but only probability.
- $H(X) \geq 0$. ($0 \leq p(x) \leq 1$ so $-\log p(x) \geq 0$)

Example

Let

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$$

Then

$$H(X) = -p\log p - (1-p)\log(1-p) \stackrel{\text{def}}{=} H(p).$$

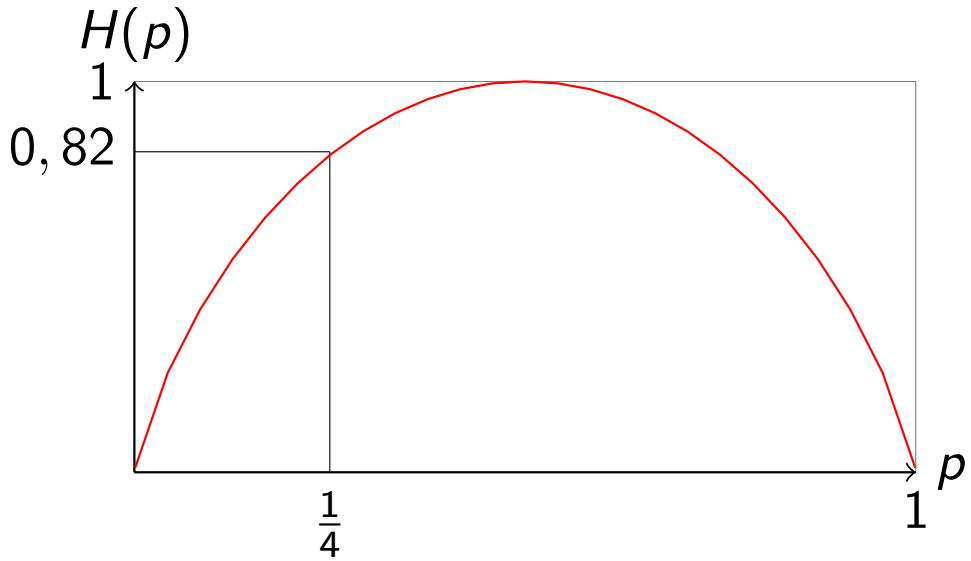


Figure : $p \mapsto H(p)$

Joint Entropy

Definition

The *joint entropy* $H(X, Y)$ of a pair of random variables X, Y is defined as

$$H(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y)$$

- An other expression is $H(X, Y) = -E \log p(X, Y)$
- If (X, Y) is seen as a vector-valued random variable we have the previous definition.

Conditional Entropy

Definition

If $(X, Y) \sim p(x, y)$, then the *conditional entropy* $H(Y|X)$ is defined as

$$\begin{aligned} H(Y|X) &= \sum_{x \in \mathcal{X}} p(x) H(Y|X = x) \\ &= -\sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x) \\ &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \\ &= -E_{p(x,y)} \log p(Y|X) \end{aligned}$$

Think of this as the amount of randomness/surprise in Y knowing X.

Chain rule

Theorem

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

Proof.

$$\begin{aligned}
H(X,Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x)p(y|x) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x) \\
&= -\sum_{x\in\mathcal{X}} p(x)\log p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x) \\
&= H(X) + H(Y|X)
\end{aligned}$$

Example

Let (X, Y) have the following joint distribution

Y \ X	1	2	3	4
1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
4	$\frac{1}{4}$	0	0	0

- Here $H(X) = \frac{7}{4}$ bits, $H(Y) = 2$ bits, $H(X|Y) = \frac{11}{8}$ bits, $H(Y|X) = \frac{13}{8}$ bits and $H(X, Y) = \frac{27}{8}$ bits.
- Note that $H(Y|X) \neq H(X|Y)$.

Mutual Information

Definition

If $(X, Y) \sim p(x, y)$, $X \sim p(x)$ and $Y \sim p(y)$, then the *mutual information* $I(X; Y)$ is

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

- Think about it as the amount of randomness/surprise in X contain by Y.
- By symmetry $I(X; Y) = I(Y; X)$

Mutual Information and Entropy

Theorem

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

Proof.

$$\begin{aligned} I(X;Y) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ &= \sum_{x,y} p(x,y) \log \frac{p(x|y)}{p(x)} \\ &= -\sum_{x,y} p(x,y) \log p(x) - (-\sum_{x,y} p(x,y) \log p(x|y)) \\ &= H(X) - H(X|Y) \end{aligned}$$

Mutual Information and Entropy

Theorem

$$I(X;Y) = H(X) - H(X|Y),$$
$$I(X;Y) = H(Y) - H(Y|X),$$
$$I(X;Y) = H(X) + H(Y) - H(X,Y),$$
$$I(X;Y) = I(Y;X),$$
$$I(X;X) = H(X).$$

Jensen's inequality

Theorem

If f is a concave function and X a random variable,

$$Ef(X) \leq f(EX).$$

Moreover, if f is strictly concave, we have equality if and only if $X = EX$, i.e. X is a constant.

Bound on Entropy

Theorem

$$H(X) \leq \log |\mathcal{X}|$$

with equality if and only if X has a uniform distribution.

Proof.

$$\begin{aligned}
-\log |\mathcal{X}| + H(X) &= -(\sum_{x \in \mathcal{X}} p(x) \log |\mathcal{X}| + \sum_{x \in \mathcal{X}} p(x) \log p(x)) \\
&= \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{|\mathcal{X}| p(x)} \\
&\leq \log \sum_{x \in \mathcal{X}} \frac{p(x)}{|\mathcal{X}| p(x)} \\
&= \log 1 = 0
\end{aligned}$$

Non-negativity of mutual information

Theorem

$$I(X;Y) \geq 0$$

with equality if and only if X and Y are independent.

Corollary

$$H(X|Y) \leq H(X)$$

with equality if and only if X and Y are independent.

Proof (Corollary).

$$0 \leq I(X;Y) = H(X) - H(X|Y)$$

□

Non-negativity of mutual information

Proof.

$$
\begin{aligned}
-I(X;Y) &= -\sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\
&= \sum_{x,y} p(x,y) \log \frac{p(x)p(y)}{p(x,y)} \\
&\leq \log \sum_{x,y} p(x)p(y) \\
&= \log(1) = 0
\end{aligned}
$$

Example

Let X, Y have the following joint distribution,

Y \ X	1	2
1	0	$\frac{3}{4}$
2	$\frac{1}{8}$	$\frac{1}{8}$

- We have $H(X) = 0,544$ bits, $H(X|Y=1) = 0$ bits, $H(X|Y=2) = 1$ bit and $H(X|Y) = 0,25$ bits.

- Note the the uncertainty in X is increased if we observe $Y = 2$ and decreased if $Y = 1$. But uncertainty decreases on the average.

Subadditivity

Theorem

$$H(X_1, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i)$$

With equality if and only if the X_i are independent.

Proof.

$$\begin{aligned} H(X_1, \ldots, X_n) &= \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1) \\ &\leq \sum_{i=1}^{n} H(X_i) \end{aligned}$$

Outline

Definition

Definition

A *source code* C for a random variable X is a mapping from $\mathcal{X}$ to $\{0,1\}^*$. $C(x)$ denote the codeword corresponding to x and $l(x)$ the length of $C(x)$

Definition

The *expected length* $L(C)$ of a source code C is

$$L(C) = \sum_{x \in \mathcal{X}} p(x) l(x)$$

Definition

A code is called an *instantaneous code* if no codeword is a prefix of any other codeword.

Kraft inequality

Theorem

Any instantaneous code satisfy

$$\sum_i 2^{-l_i} \leq 1$$

Proof.

- Consider a binary tree representing codeword
- l_{max} the length of the longest codeword
- A codeword at level l_i has $2^{l_{max}-l_i}$ descendants at level l_{max}
- These descendants sets are disjoint so

$$\sum_i 2^{l_{max}-l_i} \leq 2^{l_{max}}$$

Lower Bound on Optimal Code-length

Theorem

The expected length L of any instantaneous code for a random variable X satisfies,

$$L \geq H(X)$$

Proof.

$$\begin{aligned} H(X) - L &= \sum_i p_i \log \frac{1}{p_i} + \sum_i p_i \log 2^{-l_i} \\ &= \sum_i p_i \log \frac{2^{-l_i}}{p_i} \\ &\leq \log \sum_i 2^{-l_i} \\ &\leq \log 1 = 0 \end{aligned}$$

Huffman Coding

- We fusion the two lightest nodes to produce an heavier node.
- It is an instantaneous code.
- It is optimal, for all code C' the Huffman code C^* satisfies

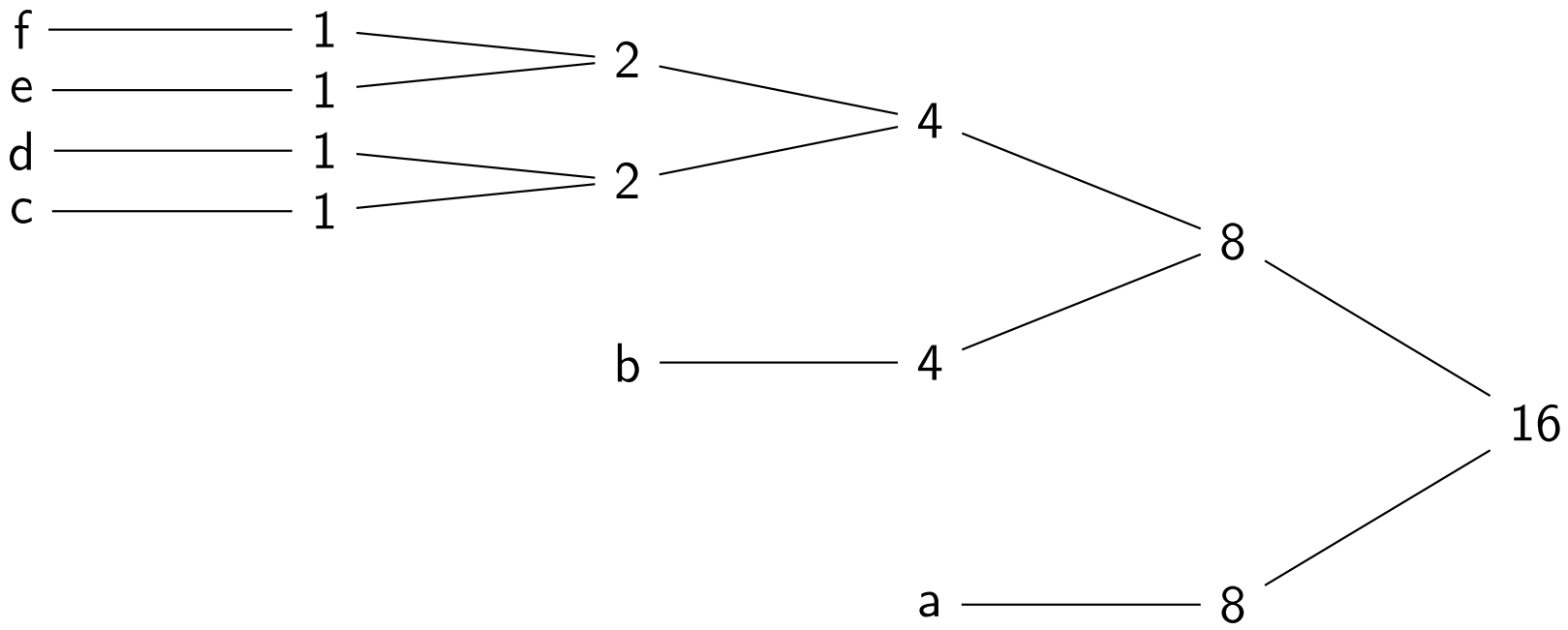
$$L(C^*) \leq L(C')$$

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

Example

Considere the following frequency apparition

a	b	c	d	e	f
8	4	1	1	1	1



So we have the following coding

a	b	c	d	e	f
0	10	1100	1101	1110	1111

Outline

Brégman Theorem

Theorem

Let $A = (a_{i,j})$ be an $n \times n$ 0-1 matrix, if the number of 1's in row i of A is r_i, then

$$perm(A) \leq \prod_{i=1}^{n} (r_i!)^{1/r_i}$$

Equivalently, let G be a bipartite graph on bipartition $\mathcal{E}, \mathcal{O}$ with each $v_i \in \mathcal{E}$ having degree d_i. Then

$$|\mathcal{M}_{perfect}(G)| \leq \prod_{i=1}^{n} (d_i!)^{\frac{1}{d_i}}$$

This is sharp because of n/d copies of $K_{d,d}$

Brégman Theorem ($|\mathcal{M}_{\text{perfect}}(G)| \leq \prod_{i=1}^{n}(d_i!)^{\frac{1}{d_i}}$)

- A perfect matching M can be view as a permutation σ with $\sigma(i) = j$ if and only if $v_i w_j \in M$ ($v_i \in \mathcal{E}$ $w_i \in \mathcal{O}$)
- Let X a random variable which represents a matching $\sigma \in \mathcal{M}_{\text{perfect}}(G)$ taken uniformly
- So we have
$$H(X) = \log |\mathcal{M}_{\text{perfect}}(G)|$$
- We will prove
$$H(X) \leq \sum_{i=1}^{n} \frac{\log d_i!}{d_i}$$

A Naive Approach

- Let's view X as a random vector $(\sigma(1), \ldots, \sigma(n))$
- By subadditivity we have

$$H(X) \leq \sum_{i=1}^{n} H(\sigma(i))$$

- There are at most d_i possible values for $\sigma(i)$

$$H(X) \leq \sum_{i=1}^{n} \log d_i$$

- By Stirling we are near because $\log(d_i!)/d_i \sim \log(d_i/e)$
- When we match v_i we need to take in account the already matched w_j

Brégman Theorem ($|\mathcal{M}_{\text{perfect}}(G)| \leq \prod_{i=1}^{n}(d_i!)^{\frac{1}{d_i}}$)

- Examine the v_i's in a random order τ

$$H(X) = \sum_{j=1}^{n} H(\sigma(\tau(j))|\sigma(\tau(1)), \ldots, \sigma(\tau(j-1)))$$

- For clarity use $k = k(\tau, i) = \tau^{-1}(i)$ and average on τ

$$\begin{aligned} H(X) &= \frac{1}{n!}\sum_{\tau}(\sum_{i=1}^{n} H(\sigma(i)|\sigma(\tau(1)), \ldots, \sigma(\tau(k-1)))) \\ &= \sum_{i=1}^{n}\frac{1}{n!}\sum_{\tau} H(\sigma(i)|\sigma(\tau(1)), \ldots, \sigma(\tau(k-1))) \end{aligned}$$

Brégman Theorem ($|\mathcal{M}_{\text{perfect}}(G)| \leq \prod_{i=1}^{n}(d_i!)^{\frac{1}{d_i}}$)

- Let us look at the inner sum $H(\sigma(i)|\sigma(\tau(1)),\dots\sigma(\tau(k-1)))$
- By definition

$$H(\sigma(i)|\sigma(\tau(1)),\dots\sigma(\tau(k-1))) = \sum_{x\in\sigma(\tau(1)),\dots,\sigma(\tau(k-1))} p(x)H(\sigma(i)|x)$$

- Note $D_i(\sigma,\tau)$ the number of neighbors of v_i not already taken by $\sigma(\tau(1)),\dots,\sigma(\tau(k-1))$. And regroup the x such that $D_i(\sigma,\tau)=j$. Given such a x $\sigma(i)|x$ has at most j outcome so using the bound

$$H(\sigma(i)|\sigma(\tau(1)),\dots\sigma(\tau(k-1))) \leq \sum_{j=1}^{d_i}\mathbb{P}(D_i(\sigma,\tau)=j)\log j$$

Brégman Theorem ($|\mathcal{M}_{\text{perfect}}(G)| \leq \prod_{i=1}^{n}(d_i!)^{\frac{1}{d_i}}$)

Remember, we had

$$H(X) = \sum_{i=1}^{n} \frac{1}{n!} \sum_{\tau} H(\sigma(i)|\sigma(\tau(1)), \ldots, \sigma(\tau(k-1)))$$

- So

$$H(X) \leq \sum_{i=1}^{n} \frac{1}{n!} \sum_{\tau} \sum_{j=1}^{d_i} \mathbb{P}(D_i(\sigma,\tau) = j) \log j$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{d_i} \log i \frac{1}{n!} \sum_{\tau} \mathbb{P}(D_i(\sigma,\tau) = j)$$

Brégman Theorem ($|\mathcal{M}_{\text{perfect}}(G)| \leq \prod_{i=1}^{n}(d_i!)^{\frac{1}{d_i}}$)

- Now we estimate the expected value of $\mathbb{P}(D_i(\sigma,\tau) = j)$
- $D_i(\sigma,\tau)$ depend only on when $\sigma(i)$ falls in the permutation of $N(v_i)$
- If it comes first, $D_i(\sigma,\tau) = d_i$; if second $D_i(\sigma,\tau) = d_i - 1$ and so on.
- By symmetry, $\sigma(i)$ is equally likely to appear in any position.

$$\frac{1}{n!}\sum_{\tau}\mathbb{P}(D_i(\sigma,\tau) = j) = \frac{1}{d_i}$$

Brégman Theorem ($|\mathcal{M}_{\text{perfect}}(G)| \leq \prod_{i=1}^{n}(d_i!)^{\frac{1}{d_i}}$)

Remember, we had

$$H(X) \leq \sum_{i=1}^{n}\sum_{j=1}^{d_i} \log i \frac{1}{n!}\sum_{\tau}\mathbb{P}(D_i(\sigma,\tau)=j)$$

- So we get

$$\begin{aligned} H(X) &\leq \sum_{i=1}^{n}\frac{1}{d_i}\sum_{j=1}^{d_i}\log j \\ &= \sum_{i=1}^{n}\frac{\log d_i!}{d_i} \end{aligned}$$

Kahn-Lovász Theorem

Theorem

Let G be a graph on $2n$ vertices v_i of degree d_i, then

$$|\mathcal{M}_{perfect}(G)| \leq \prod_{i=1}^{2n} (d_i!)^{1/2d_i}$$

Kahn-Lovász Theorem $|\mathcal{M}_{perfect}(G)| \leq \prod_{i=1}^{2n}(d_i!)^{1/2d_i}$

Proof.

Let $\mathcal{S}_e$ the set of all even cycle covers, $\mathcal{S}$ the set of all cycle covers and $c(S)$ the number of cycles in S.

- Two perfect matching form an even cycle cover.

$$|\mathcal{M}_{\text{perfect}} \times \mathcal{M}_{\text{perfect}}| = \sum_{S \in \mathcal{S}_e} 2^{c(S)}$$

- Any permutation contributing to $\text{perm}(\text{Adj}(G))$ form a cycle cover.

$$\text{perm}(\text{Adj}(G)) = \sum_{S \in \mathcal{S}} 2^{c(S)}$$

- So by Brégman's Theorem

$$|\mathcal{M}_{\text{perfect}}(G)| \leq \sqrt{\text{perm}(\text{Adj}(G))} \leq \prod_{i=1}^{2n}(d_i!)^{1/2d_i}$$

Open Questions

Conjecture

For d-regular G on $2n$ vertices

$$|\mathcal{M}(G)| \leq |\mathcal{M}(K_{n,d})|$$

Conjecture (Upper Matching conjecture)

For bipartite d-regular G on $2n$ vertices, and for all t,

$$|\mathcal{M}_t(G)| \leq |\mathcal{M}_t(K_{n,d})|$$

Known results

Conjecture (Upper Matching conjecture)

For bipartite d-regular G on $2n$ vertices, and for all t,

$$|\mathcal{M}_t(G)| \leq |\mathcal{M}_t(K_{n,d})|$$

- For $t = n$ Upper Matching conjecture if Brégman's theorem
- It is easy for $t = 0, 1, 2$.
- It is prove by Friedland, Krop and Markström for $t = 3, 4$
- It we fixed $\alpha \in [0, 1]$, Carroll, Galvin and Tetali prove the following

$$\log |\mathcal{M}_{\alpha n}(G)| \leq n(\alpha \log d + H(\alpha))$$

Summary

- $H(X) \stackrel{\text{def}}{=} -\sum_{x \in \mathcal{X}} p(x) \log p(x) = -E \log p(x)$
- (Chain rule) $H(X, Y) = H(X) + H(Y|X)$
- (Mutual Information) $0 \leq I(X; Y) = I(Y; X) = H(X) - H(X|Y)$
- (Bounds) $0 \leq H(X) \leq \log |\mathcal{X}|$
- (Conditioning reduces Entropy) $H(X|Y) \leq H(X)$
- (Subadditivity) $H(X_1, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i)$