Lecture 2

- Definition, Properties of H(X), I(X;Y)
- Differential Entropy
- Asymptotic Equipartition Property (AEP)

Entropy Function

- Entropy measures <u>uncertainty</u> in a r.v.

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

- Properties:

1. $H(X) \geq 0$
2. $H_b(X) = \log_b(a)\, H_a(X)$
3. $H(p)$ concave in prob. distribution p

Joint Entropy

$$H(X,Y) = -\sum_{x,y} p(x,y)\log p(x,y) = -E[\log p(x,y)]$$

$$H(Y \mid X) = -\sum_{x} p(x) H(Y \mid X = x) = -E[\log p(y \mid x)]$$

Note: $H(X \mid Y) \neq H(Y \mid X)$ **in general**

Chain Rule: $H(X,Y) = H(X) + H(Y \mid X)$

$$= H(Y) + H(X \mid Y)$$

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

$$H(X,Y \mid Z) = H(X \mid Z) + H(Y \mid Z, X)$$

Conditioning: $H(Y \mid X) \leq H(Y)$

$$H(X_1, \ldots X_n) \leq \sum_{i=1}^{n} H(X_i) \quad \textbf{equality} \leftrightarrow X_i \textbf{ independent}$$

Mutual Information

- Amount of information one r.v. has about another r.v

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

$$= H(X) - H(X \mid Y)$$

$$= H(Y) - H(Y \mid X) = I(Y;X)$$

$I(X;Y) \geq 0$ **(with equality** $\leftrightarrow$ **X, Y independent)**

$I(X;X) = H(X) - H(X \mid X) = H(X)$ **(self-information)**

$$I(X;Y \mid Z) = H(Y \mid Z) - H(Y \mid X, Z)$$

Chain Rule: $I(X_1, X_2; Y) = I(X_1; Y) + I(X_2; Y \mid X_1)$

$$I(X_1, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y \mid X_1, \ldots, X_{i-1})$$

$I(X;Y)$ **concave in p(x) for fixed p(y | x)**

convex in p(y | x) for fixed p(x)

Example

- X_n = 1 if snows on n-th day, 0 otherwise
- $P(X_n=1) = 1/4$ for all n
- X_n correlated with previous, next day
- $P(X_n=1|X_{n-1}=1)=1/2$, $P(X_n=1|X_{n-1}=0)=1/6$

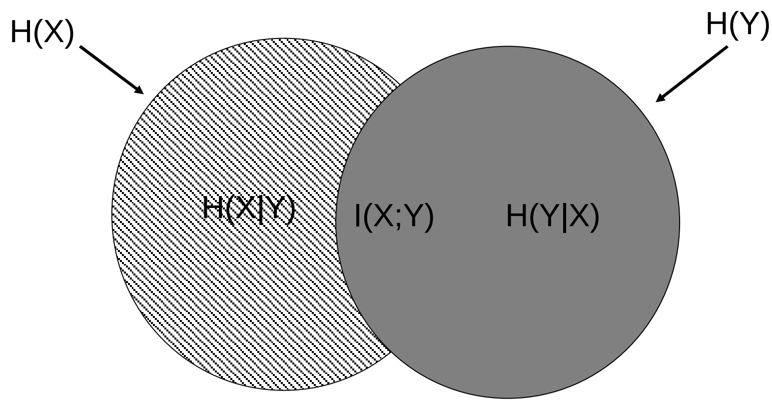
- $H(X_n) = H(1/4) = .8113$
- $H(X_n|X_{n-1}=1) = H(1/2) = 1$
- $H(X_n|X_{n-1}=0) = H(1/6) = .65$
- $H(X_n|X_{n-1})=1/4 * 1 + 3/4 * .65 = .7375 < H(X_n)$
- Conditioning <u>always</u> reduces entropy ($H(Y|X) <= H(Y)$) but can have more entropy for specific choice of X (i.e. $H(Y|X=x) > H(Y)$ as have above when $X_{n-1}=1$)

Example (cont)

- $I(X_n; X_{n-1}) = H(X_n) – H(X_n|X_{n-1})$
  $= .8113 - .7375$
  $= .0738$

- Same model with $P(X_n=1|X_{n-1}=1)=1, P(X_n=1|X_{n-1}=0)=0$
- $I(X_n; X_{n-1}) = H(X_n) – H(X_n|X_{n-1})$
  $= .8113 - 0$
  $= .8113 = H(X_n)$

Venn Diagram



Kullback-Leibler Distance

- K-L distance measures "distance" between two prob. distributions

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

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

Distance betwen p(x, y) and p(x)p(y)

$D(p \| q) \geq 0$ with equality iff p(x) = q(x) for all x

Asymmetric: $D(p \| q) \neq D(q \| p)$

Inequalities

- Jensen's Inequality:
  For any convex function f and any r.v. X
  E[f(X)] >= f(E[X])

- Data Processing Inequality:
  If X -> Y -> Z form a Markov chain, then
  I(X;Y) >= I(X;Z)
  I(Y;Z) >= I(X;Z)

  i.e. closer you are in chain, higher the mutual inf

Differential Entropy

- For continuous r.v. X with pdf f(x)

$$h(X) = -\int f(x)\log f(x)dx$$

- Properties:

$h(X)$ not necessarily non-negative

Translation: $\mathrm{h(X+a) = h(X)}$

Scaling: $\mathrm{h(aX) = h(X) + \log|a|}$

Conditioning: $\mathrm{h(X|Y) \le h(X)}$

Mutual Inf: $\mathrm{I(X;Y) = h(X) - h(X|Y)}$

Asymptotic Equipartition Property

$X_1, X_2, \ldots$ are iid ~ p(x)

AEP: $-\frac{1}{\mathrm{n}} \log p(X_1, \ldots, X_n) \to H(X)$

Typical set $A_\varepsilon^{(n)}$: sequences $(x_1, \ldots, x_n)$ satisfying

$2^{-\mathrm{n(H(X)}+\varepsilon)} \leq p(x_1, \ldots, x_n) \leq 2^{-\mathrm{n(H(X)}-\varepsilon)}$

- As n gets large, almost all sequences are typical
- Roughly $2^{nH(x)}$ typical sequences, each with probability roughly equal to $2^{-nH(x)}$

Typical sequences

Typical set $A_\varepsilon^{(n)}$: sequences $(x_1, \ldots, x_n)$ satisfying

$2^{-\mathrm{n(H(X)}+\varepsilon)} \leq p(x_1, \ldots, x_n) \leq 2^{-\mathrm{n(H(X)}-\varepsilon)}$

Properties:

1. $P(A_\varepsilon^{(n)}) > 1 - \varepsilon$ for sufficiently large n
2. $|A_\varepsilon^{(n)}| \leq 2^{\mathrm{n(H(X)}+\varepsilon)}$
3. $|A_\varepsilon^{(n)}| \geq (1-\varepsilon) 2^{\mathrm{n(H(X)}-\varepsilon)}$