Data Compression

Entropy

中央大學資工系
蘇柏齊

Information

- What is information??
  - Knowledge derived from study, experience, or instruction
  - Knowledge of a specific event or situation
  - A collection of facts or data
  - Communication of knowledge
- Information measurement
  - Quantify information:
    how much information in this piece of data?

Intuitive Properties of Information Measurement

- Property 1
  - Information contained in events should be defined in terms of some measurement of the uncertainty of the events
- Property 2
  - “Less certain events” should contain more information than “more certain events.”
    (The amount of information in an event = the degree of surprise)
- Property 3
  - Information obtained from the occurrence of two independent events is the sum of the information obtained from the individual events
- Property 4
  - The information amount should be a positive number

Self-information of an Outcome x_i in Random Variable (Experiment) X

- Shannon defined the "self information" as $I(x_i) = -\log_b P(x_i)$
  - The base, b, of the logarithm depends on the unit of information
    - $\log_2$: bit ($\log_e$: nat, $\log_{10}$: Hartley)
    - Base conversion: $\log_2 k = \log_{10} k / \log_{10} 2$
- Compared with intuitive properties
  - $P(x_i)$ is the probability of the occurrence of x_i

    …… Property (1)
  - $I(x_i)$ is a continuous function of $P(x_i)$ and increases as $P(x_i)$ goes from 1 to 0

    …… Property (2)
  - If x_i, x_j are independent events,
    - $I(x_i,x_j) = I(x_i) + I(x_j)$

      $I(x_i) = -\log P(x_i) = \log 1/P(x_i)$

      $I(x_i,x_j) = -\log P(x_i,x_j) = \log 1/\{P(x_i)P(x_j)\} = \log 1/P(x_i) + \log 1/P(x_j) = I(x_i) + I(x_j)$

    …… Property (3)
  - $I(x_i) >= 0$

    …… Property (4)

Alphabet or Sample Space of X: $S_X=\{x_0, x_1,\ldots, x_{m-1}\}$

- Letters or samples represent all the possible outcomes (events)
  - Text {a, b, c,… z, A, B,…}
  - Binary {0, 1}
  - Gray Level Image {0, 1,… 255}
  - Speech signal $\{-2^{15}\ldots 2^{15}-1\}$

Entropy

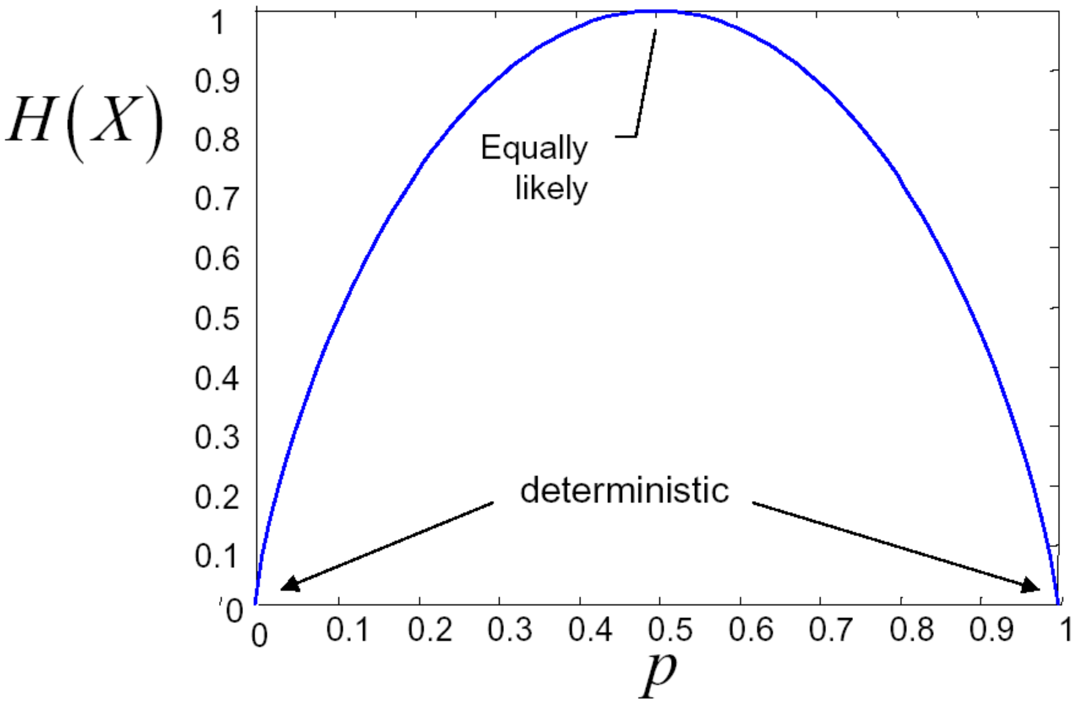
- Average (expected) information amount over the whole alphabet:

$$H(X) = \mathop{E}_{x \in S_X} \{I(x)\} = -\sum_{x \in S_X} P(x) \cdot \log P(x)$$

- Example: weather
  - S_X = {Rain, Fine, Cloudy, Snow}
    - P(Rain) =½, P(Fine) =¼,
      P(Cloudy) = ¼, P(Snow) = 0
      - H(X) = 1.5 bits/symbol
    - If ¼ for each case (equal probability)
      - H(X) = 2 bits/symbol (>1.5 bits)
    - H(X)= 0 for a certain experiment with P=1
      - X:The Sun rises from ? East (P=1), West (P=0), H(X)=0.

Binary Random Variable

- Consider a binary memoryless source {0, 1}, P_0=p, P_1=1-p
  $\rightarrow H = -p \log p - (1-p) \log (1-p)$



Properties of Entropy

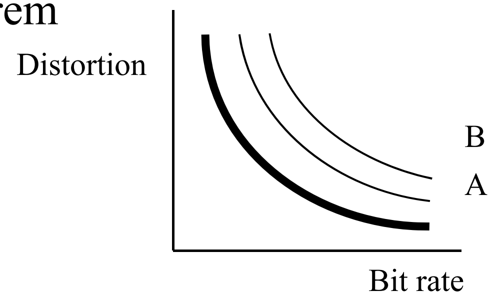
- Bound of entropy
  - $0 <= H(X) <= \log_2(\text{Size of Alphabet})$
    - Lower bound achieved when only one outcome can occur
    - Higher bound achieved when all outcomes are equally likely
- Very likely and very unlikely outcomes do not substantially change entropy of a random variable
  - $-p \log_2 p \rightarrow 0$ for $p \rightarrow 0$ or $p \rightarrow 1$

Noiseless Source Coding Theorem

- For a source with Entropy H bits/message, it is possible to find a distortionless (lossless) coding scheme using an average of H+e bits/message. e>0 is an arbitrarily small quantity.
- The entropy H(X) is a lower bound for the average word length *R* of a decodable variable-length code for the symbols.
- Redundancy of a code: R-H(X)>=0

Information Theory

- Information theory
  - Use a probabilistic model to measure the amount of information associated with a data source.
  - Characterize performance bounds of source and channel codes under various circumstances.
- Shannon's three theorems
  - Shannon's 1st Theorem
    – Source coding (noiseless) theorem
    - It characterizes the minimum average codeword length per source symbol that can be achieved.
  - Shannon's 2nd Theorem
    – Channel coding (noisy) theorem
    - It characterizes the probability of error transmitted through noisy channel.
    - It tells us that the prob. of error can be made arbitrarily small if the coded message rate is less than the capacity of the channel.
  - Shannon's 3rd Theorem
    – Rate-distortion (lossy compression) theorem
    - By constraining the average error rates (distortion) to same maximum level D, we determine the smallest bit rate to represent the information source. This is known as Rate-Distortion function.



Joint Entropy

- Consider random vectors (with discrete, finite-alphabet components) $\mathbf{X} = [X_0, X_1, \ldots, X_{n-1}]$
- Joint Entropy:

$$H(\mathbf{X}) = H(X_0, X_1, \ldots, X_{n-1}) = -\sum_{x_0}\sum_{x_1}\cdots\sum_{x_{n-1}} P(x_0, x_1, \ldots, x_{n-1}) \log P(x_0, x_1, \ldots, x_{n-1})$$

- Example: guess or transmit a "four letter word":
  - $H(\mathbf{X}) = H(X_0, X_1, X_2, X_3)$

$$= -\sum_{x_0=a}^{z}\sum_{x_1=a}^{z}\sum_{x_2=a}^{z}\sum_{x_3=a}^{z} P(x_0, x_1, x_2, x_3) \log P(x_0, x_1, x_2, x_3)$$

Shannon's Noiseless Source Coding Theorem Expressed in Joint Entropy

- Consider a "vector source" **X**. Joint entropy $H(\mathbf{X})$ is the achievable lower bound of the bit-rate for encoding **X.**
- If a source that puts out symbols from a set A, then the entropy is a measure of the average number of binary symbols needed to code the output of the source. (The best that a lossless compression scheme can do is to encode the output of a source with an average number of bits equal to the entropy of the source.)
- Entropy of the source: $\lim_{n\to\infty} \frac{1}{n} G_n$

$$G_n = -\sum_{i_1=1}^{m}\sum_{i_2=1}^{m}\cdots\sum_{i_n=1}^{m} P(X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n) \log P(X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n)$$

- If each element of the vector source is i.i.d. (identically independent distributed), the average code length for one symbol:

$$G_n = -n\sum_{i_1=1}^{m} P(X_1 = i_1)\log P(X_1 = i_1) \Longrightarrow H(X) = -\sum_{i_1=1}^{i_1=m} P(X_1 = i_1)\log P(X_1 = i_1)$$

First-order entropy

- Entropy of the source is basically *unknowable*. If we know more about the source, we may *estimate* the actual source entropy more accurately via a good modeling.

Examples

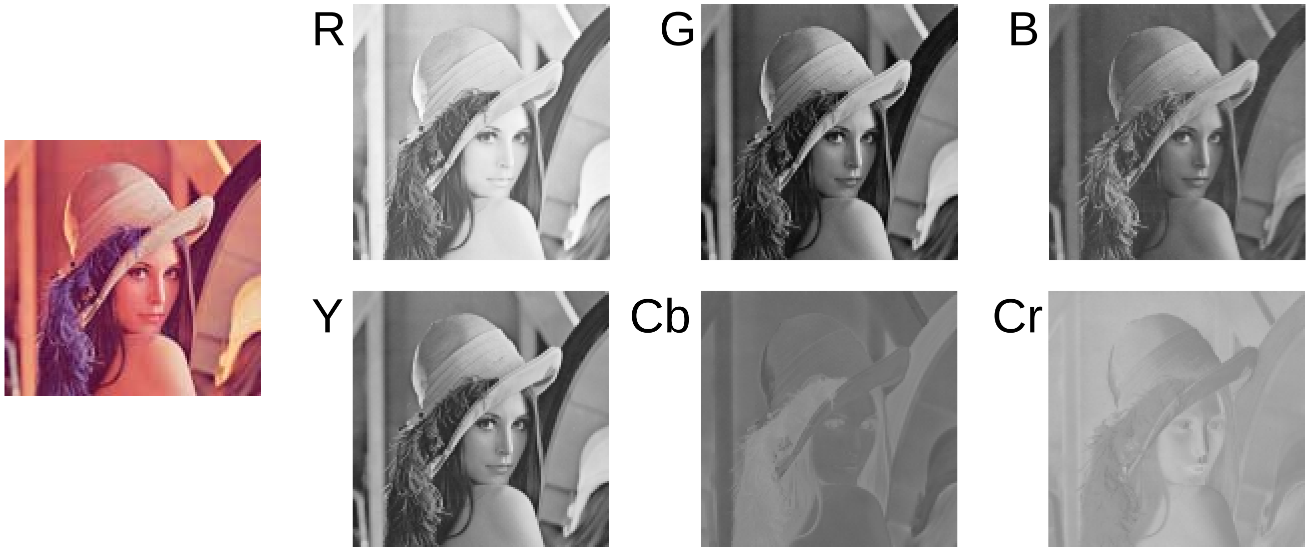
- Consider the following sequence:
  1 2 3 2 3 4 5 4 5 6 7 8 9 8 9 10
  - Assume I.I.D.
    - P(1)=P(6)=P(7)=P(10)=1/16
      P(2)=P(3)=P(4)=P(5)=P(8)=P(9)=2/16
      $$H(X)=-\sum_{x=1}^{10} P(x)\cdot \log P(x) = 3.25(\mathit{bits})$$
  - However, consider residual sequence:
    1 1 1 -1 1 1 1 -1 1 1 1 1 1 -1 1 1
    - P(1)=13/16, P(-1)=3/16
      H(X)= 0.7 (bits)
- 12123333123333123312
  - Assume I.I.D.
    - P(1)=P(2)=1/4, P(3)=1/2 => 1.5 bits/symbol => 30 bits
  - 12, 12, 33… P(12)=P(33)=1/2 => 1 bits/symbol => 10 bits

Joint Entropy and Statistical Dependence

- Theorem: $H(X_0,X_1,\ldots,X_{n-1}) <= H(X_0)+H(X_1)+\ldots+H(X_{n-1})$
  - Equality for statistical independence of $X_0,X_1,\ldots,X_{n-1}$
  - Exploiting statistical dependence can reduce bit-rate
  - Statistically independent components can be compressed and decompressed separately without loss

Statistical Dependence among Color Components

- Image: 'Lena', 512 x 512 pixels



Calculate 1st-order statistics

$H(R)$=7.25 bpp
$H(G)$=7.59 bpp
$H(B)$=6.97 bpp

$H(Y)$=7.23 bpp
$H(Cb)$=5.47 bpp
$H(Cr)$=5.42 bpp

R fixed	3 x 8 = 24 bpp
$H(Y,Cb,Cr)$	15.01 bpp
$H(Y)+H(Cb)+H(Cr)$	18.12 bpp
ΔH	3.11 bpp

R fixed	3 x 8 = 24 bpp
$H(R,G,B)$	16.84 bpp
$H(R)+H(G)+H(B)$	21.82 bpp
ΔH	4.98 bpp

Conditional Entropy

- Consider two discrete finite-alphabet r.v. *X* and *Y*

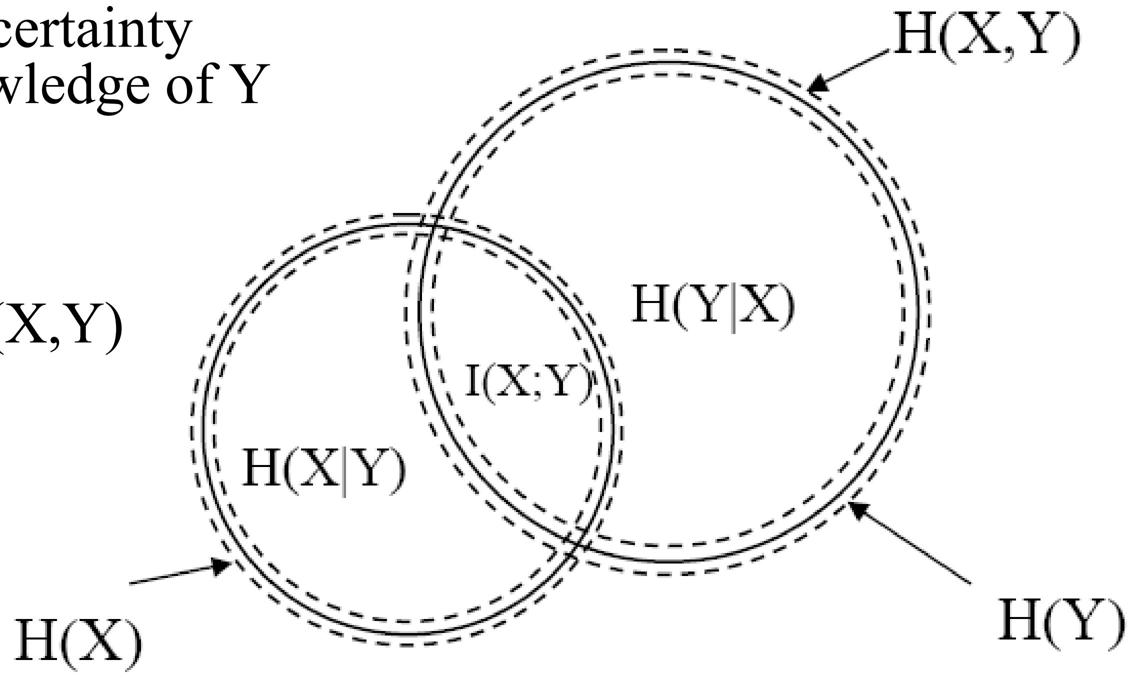
$$H(X\,|\,Y) = E\left[-\log_2 f_{X|Y}(x,y)\right] = -\sum_y \sum_x f_{X,Y}(x,y)\log_2 f_{X|Y}(x,y)$$

- Conditional entropy *H(X|Y)* is average additional information, if *Y* is already known

$$\begin{aligned} H(X,Y) &= E\left[-\log_2 f_{X,Y}(X,Y)\right] \\ &= E\left[-\log_2\left(f_Y(Y)\,f_{X|Y}(X,Y)\right)\right] \\ &= E\left[-\log_2 f_Y(Y)\right] + E\left[-\log_2 f_{X|Y}(X,Y)\right] \\ &= H(Y) + H(X\,|\,Y) \end{aligned}$$

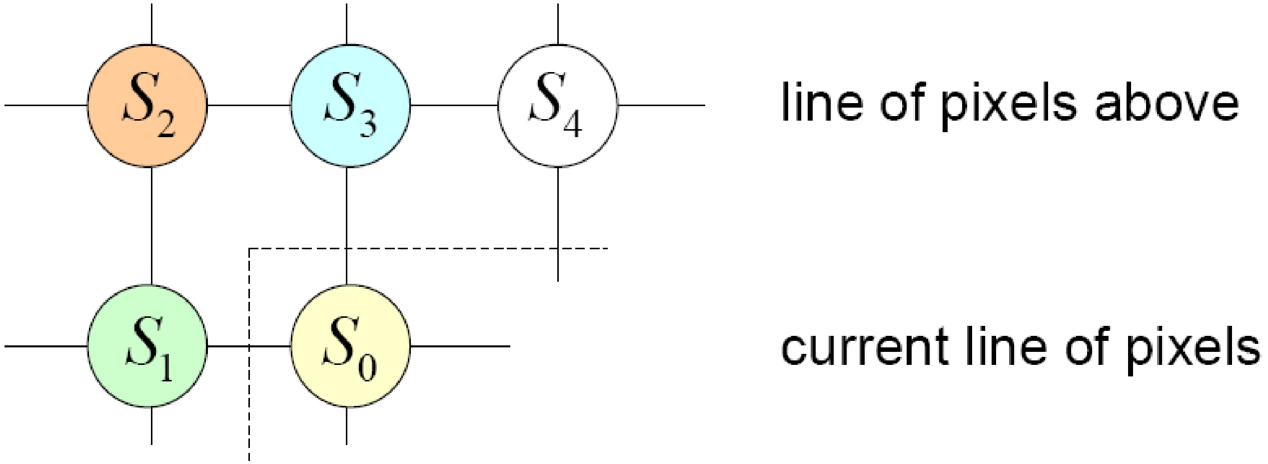
Entropy and Mutual Information

- $H(X, Y) = H(X) + H(Y|X)$
- $H(Y|X) \neq H(X|Y)$
- $H(X|Y) \leq H(X)$
- Mutual Information
  - Reduction in the uncertainty of X due to the knowledge of Y
  - $I(X;Y)$
    $= H(X) - H(X|Y)$
    $= H(Y) - H(Y|X)$
    $= H(X) + H(Y) - H(X,Y)$
  - $I(X;X)$
    $= H(X) - H(X|X)$
    $= H(X)$
    Self-information of X



Conditional Entropy of Images

- Image: 'Lena', 512 x 512 pixels



component	$H(S_0)$	$H(S_0 \mid S_1)$	$H(S_0 \mid S_3)$	$H(S_0 \mid S_2)$
Y	7.23	4.67	4.32	4.86
Cb	5.47	3.80	3.58	3.85
Cr	5.42	3.69	3.55	3.82

Markov Model

- Having a good model for the data can be useful in estimating the entropy of the source.
- One of the most popular ways of representing dependence in the data is through the use of Markov models.
- Let $\{x_{n-1}, x_{n-2}, \ldots, x_{n-k}, \ldots\}$ be a sequence of observations (outputs). The sequence is said to follow a kth-order Markov model if $P(x_n | x_{n-1}, \ldots, x_{n-k}) = P(x_n | x_{n-1}, \ldots, x_{n-k}, \ldots)$.
- $\{x_{n-1}, x_{n-2}, \ldots, x_{n-k}, \ldots\}$ are called the state of the process.
- Knowledge of the past k symbols is equivalent to the knowledge of the entire past history of the process.
- 1st-order Markov model is commonly used. Different forms exist:
  - $x_n = ax_{n-1} + e$
- Shannon used a 2nd-order model for English text consisting of the 26 letters and one space => 3.1 bits/letter.
- An experiment of using 100 letters => 1.3 and 0.6 bits/letter.

Entropy for kth-Order Markov Model (Finite Context Model)

Read k symbols, $s_{k-1}, s_{k-2}, \ldots, s_0$, $P(s_k | s_{k-1}, s_{k-2}, \ldots, s_0)$

$I(s_k | s_{k-1}, s_{k-2}, \ldots, s_0) = -\log_2 P(s_k | s_{k-1}, s_{k-2}, \ldots, s_0)$

$$H(S | s_{k-1}, s_{k-2}, \ldots, s_0)$$

$$= \sum_{S} P(s_k | s_{k-1}, s_{k-2}, \ldots, s_0) \log \frac{1}{P(s_k | s_{k-1}, s_{k-2}, \ldots, s_0)}$$

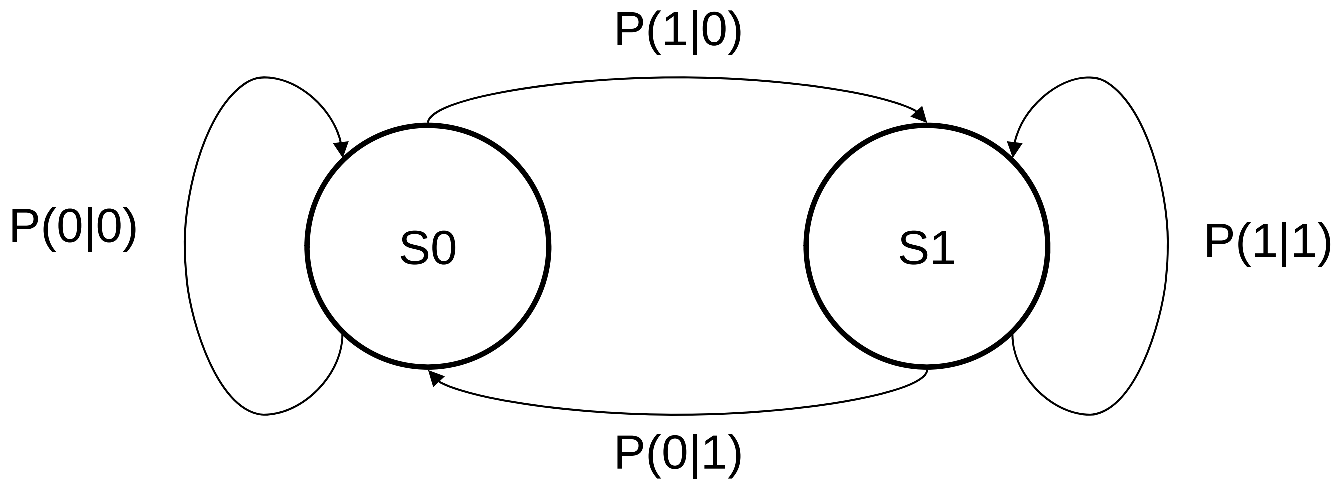
$$H(S) = \sum_{S^k} P(s_{k-1}, s_{k-2}, \ldots, s_0) H(S | s_{k-1}, s_{k-2}, \ldots, s_0)$$

$$= \sum_{S^k} P(s_{k-1}, s_{k-2}, \ldots, s_0) \sum_{S} P(s_k | s_{k-1}, s_{k-2}, \ldots, s_0) \log \frac{1}{P(s_k | s_{k-1}, s_{k-2}, \ldots, s_0)}$$

$$= \sum_{S^k} \sum_{S} P(s_{k-1}, s_{k-2}, \ldots, s_0) P(s_k | s_{k-1}, s_{k-2}, \ldots, s_0) \log \frac{1}{P(s_k | s_{k-1}, s_{k-2}, \ldots, s_0)}$$

$$= \sum_{S^{k+1}} P(s_k, s_{k-1}, s_{k-2}, \ldots, s_0) \log \frac{1}{P(s_k | s_{k-1}, s_{k-2}, \ldots, s_0)}$$

Example: Binary Data



- P(0|0)=0.99, P(1|0)=0.01, P(1|1)=0.70, P(0|1)=0.3, (P(S0)=30/31, P(S1)=1/31)
- I.I.D.
  - H=-(30/31)log(30/31)-(1/31)log(1/31)=.206 bits
- 1st-order
  - -P(0,0)logP(0|0)-P(1,0)logP(1|0)-P(1,1)logP(1|1)-P(0,1)logP(0|1)
    = -(30/31*0.99)log(0.99) -(30/31*0.01) log(0.01)
    -(1/31*0.7)log(0.7) -(1/31*0.3)log(0.3)
    = 0.107 bits

ps. P(0,0)+P(1,0)+P(1,1)+P(0,1)=1

Another Look at Joint and Conditional Entropy

- **The joint Entropy H(X, Y) of a pair of discrete random variables (X, Y) with a joint distribute P(X, Y) is defined as:**

$$H(x, y) = -\sum_{x \in X} \sum_{y \in Y} P(x, y) \cdot \log P(x, y) \ {}^{\text{bits}}\!/\!_{\text{message}} = -E\left\{\log \frac{1}{P(x, y)}\right\}$$

- **The conditional Entropy:**

$$\begin{aligned}
H(Y|X) &= \sum_{x \in X} P(x) \cdot H(Y|X{=}x) \\
&= -\sum_{x \in X} P(x) \sum_{y \in Y} P(y|x) \cdot \log P(y|x) \\
&= -\sum_{x \in X} \sum_{y \in Y} P(x) \cdot P(y|x) \cdot \log P(y|x) \\
&= -\sum_{x \in X} \sum_{y \in Y} P(x \cdot y) \cdot \log P(y|x) \\
&= -E\{\log P(y|x)\}
\end{aligned}$$

- **H(XY) = H(X)+H(Y|X)**

$$\begin{aligned}
H(X \cdot Y) &= -\sum_{x \in X} \sum_{y \in Y} P(x \cdot y) \cdot \log P(x \cdot y) \\
&= -\sum \sum P(x \cdot y) \cdot \underline{\log \{P(x) \cdot P(y|x)\}} \\
&= -\sum \sum P(x \cdot y) \cdot \underline{\log P(x)} \\
&\quad -\sum \sum P(x \cdot y) \cdot \underline{\log P(y|x)} \\
&= -\sum P(x) \cdot \log P(x) \\
&\quad -\sum \sum P(x \cdot y) \cdot \log P(y|x) \\
&= H(X) + H(Y|X)
\end{aligned}$$