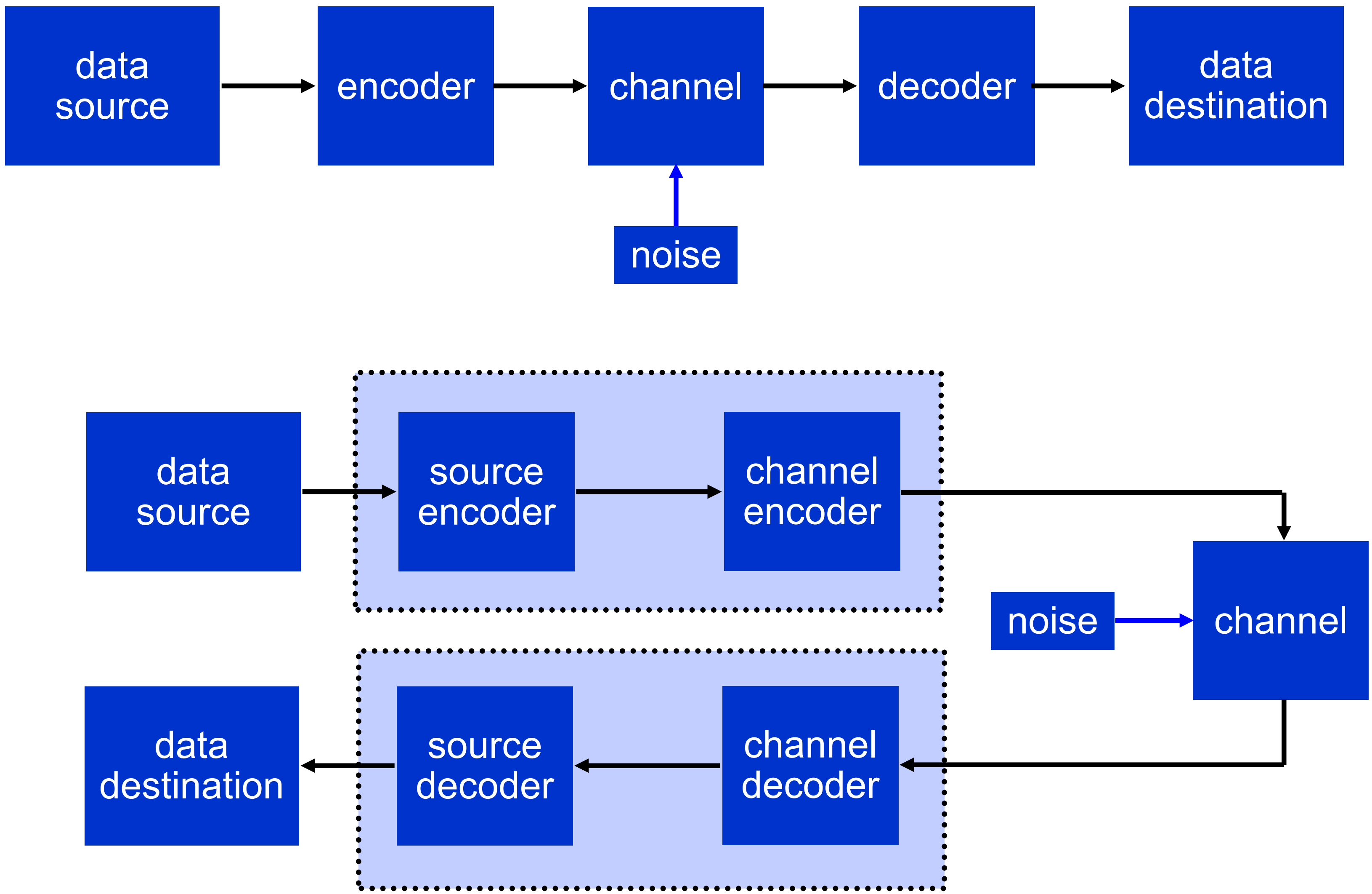
Applications of Information Theory in Image Processing

1. Review of Information Theory and Lossless Source Coding

Notation

- A : discrete (usually finite) alphabet; $\alpha = |A|$: size of A (when finite)
- A^n : set of strings of length n over A; A^* : set of finite strings over A
- λ : empty string
- $x_1^n = x^n = x_1 x_2 \ldots x_n$: finite sequence over A
- $x_1^\infty = x^\infty = x_1 x_2 \ldots x_t \ldots$: infinite sequence over A
- $x_i^j = x_i x_{i+1} \ldots x_j$: sub-sequence (i sometimes omitted if $= 1$)
- $p_X(x)$ or $P_X(x)$: $\mathrm{Prob}(X = x)$ probability mass function (PMF) on A
  (subscript X dropped if clear from context)
- $X \sim p(x)$: X obeys PMF $p(x)$
- $E_p[F]$: expectation of F w.r.t. PMF p (subscript and [ ] may be dropped)
- $\hat{p}_{x^n}(x)$: empirical distribution obtained from x^n
- $\log x$: logarithm to base 2 of x, unless base otherwise specified
- $\ln x$: natural logarithm of x
- $H(X), H(p)$: entropy of a r.v. X or PMF p, in bits (usually per-symbol)
- $\mathbf{H}(X^n)$: joint entropy of $X_1, X_2, \ldots, X_n$ (unnormalized)
- $H_2(p) = -p \log p - (1-p)\log(1-p)$, $0 \le p \le 1$: binary entropy function
- $D(p||q)$: relative entropy (information divergence) between PMFs p and q

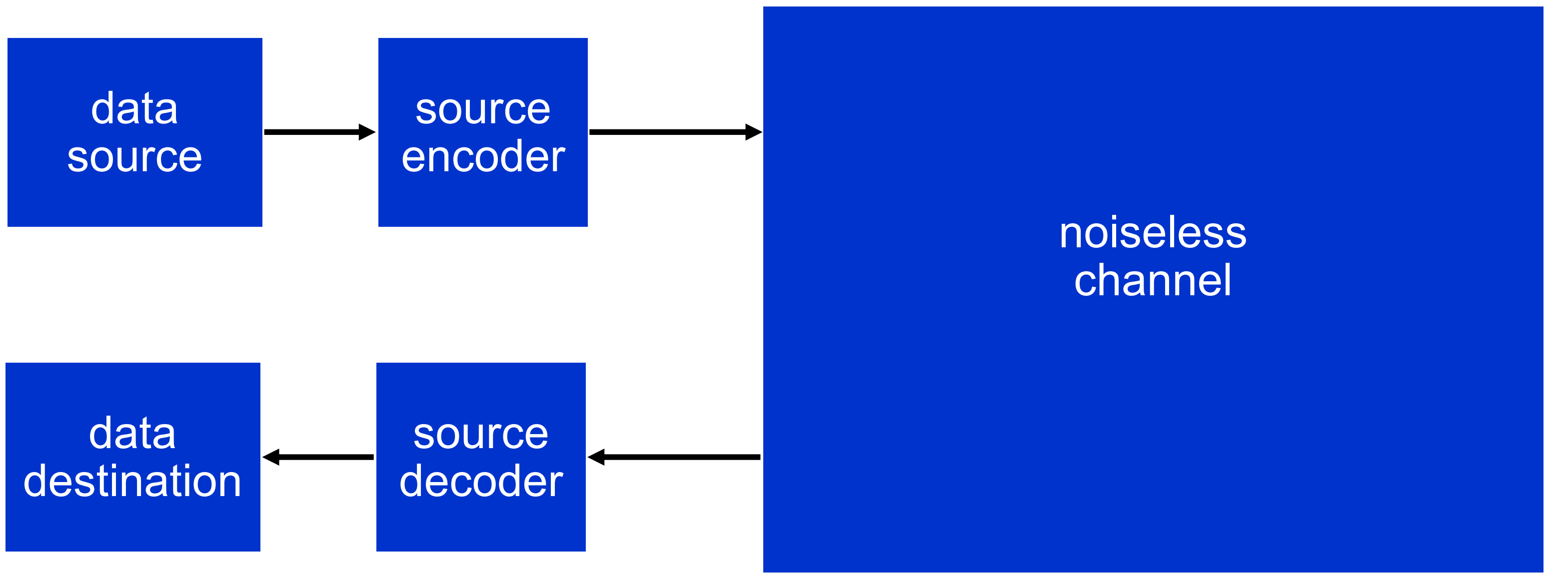
Coding in a communication/storage system



Information Theory

- **Shannon, *"A mathematical theory of communication,"* *Bell Tech. Journal,* 1948**
  - Theoretical foundations of source and channel coding
  - Fundamental bounds and coding theorems in a probabilistic setting
    - in a nutshell: perfect communication in the presence of noise is possible as long as the *entropy rate* of the source is below the *channel capacity*
  - Fundamental theorems essentially non-constructive: we've spent the last 70 years realizing Shannon's promised paradise in practice
    - very successful: enabled current digital revolution (multimedia, internet, wireless communication, mass storage, …)
  - Separation theorem: *source and channel coding can be done independently*

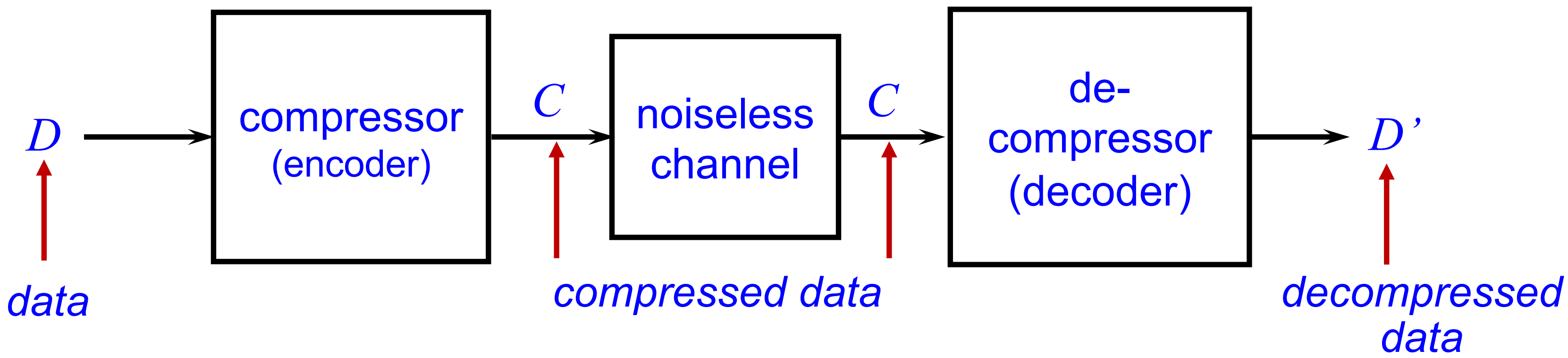
Source Coding



Source coding = Data compression

$\Rightarrow$ efficient use of bandwidth/space

Data Compression

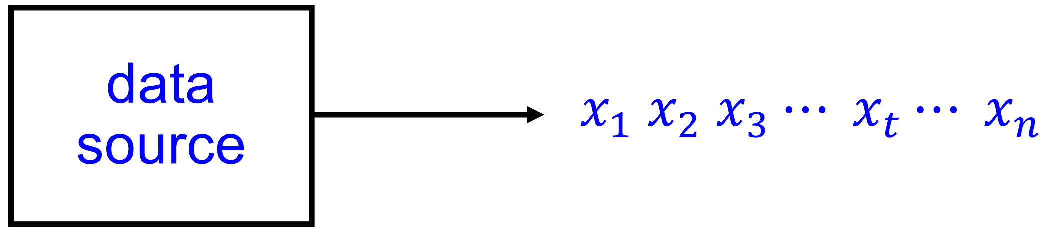


the goal: $\mathrm{size}(C) < \mathrm{size}(D)$

compression ratio: $\rho = \dfrac{\mathrm{size}(C)}{\mathrm{size}(D)}$ in appropriate units, e.g., bits/symbol or unitless bits/bit

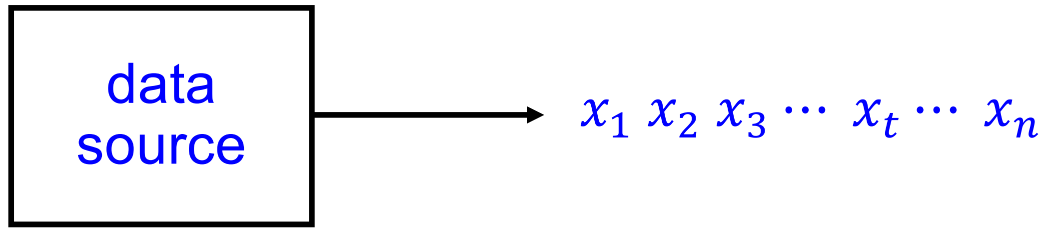
- *Lossless* compression: $D = D'$ (the case of interest here)
- *Lossy* compression: $D' \approx D$; D' is an *approximation* of D under some metric

Data Sources



- Symbols $x_i \in A$ $=$ a *countable* (usually *finite*) *alphabet.*
- *Probabilistic source:* x_i are realizations of *random variables* X_i ; X_1^n obeys some probability distribution P on A^n.
- We are often interested in $n \to \infty$: X_1^∞ is a *random process.*
    - *stationary* (*time-invariant*): $X_i^\infty = X_j^\infty$, as random processes, $\forall i, j \geq 1$
    - *ergodic*: time averages converge to ensemble averages
    - *memoryless*: X_i are statistically independent
    - *independent, identically distributed* (*i.i.d.*): memoryless, and $X_i \sim p_X \;\; \forall i$

Data Sources



- Symbols $x_i \in A$ = a *countable* (usually *finite*) *alphabet.*
- *Individual sequence:* x_i are just symbols, not assumed to be a realization of a random process. We will talk about *probability assignments*, but they will be derived from the *data* under certain constraints, and with certain objectives.
- Here too, we will often be interested in asymptotic behavior as $n \to \infty$.

Statistics on Individual Sequences

- Empirical distributions derived from an individual sequence x^n.

$$\hat{p}_{x^n}(a) = \frac{1}{n}\left|\{\, i :\ 1 \le i \le n,\ x_i = a \,\}\right|,\quad a \in A$$

Memoryless (zero-th order)

  - We can compute empirical statistics of *any order* (joint, conditional, etc.)
  - Sequence probability according to its own empirical distribution

$$\hat{P}_{x^n}(x^n) = \prod_{i=1}^{n} \hat{p}_{x^n}(x_i)$$

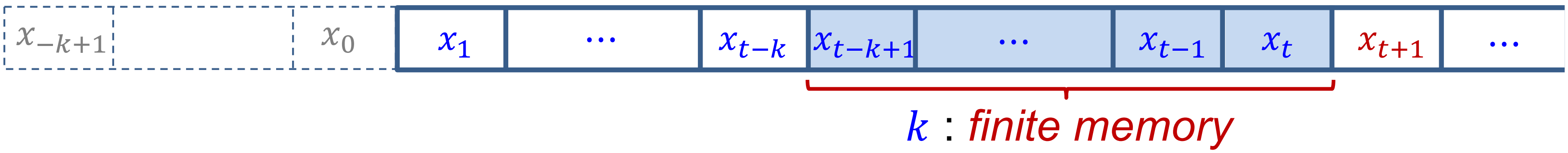
  - This is the *highest probability* assigned to the sequence by any distribution from the model class: *maximum likelihood (ML) probability*
  - Example: *Bernoulli model* (binary, i.i.d.)

$$A = \{0,1\},\quad n_0 = \left|\{\, i :\ x_i = 0\}\right|,\quad n_1 = n - n_0 = \left|\{\, i :\ x_i = 1\}\right|$$

$$\hat{p}(0) = \frac{n_0}{n},\ \hat{p}(1) = \frac{n_1}{n}\ :\ \hat{P}_{x^n}(x^n) = \hat{p}(0)^{n_0}\,\hat{p}(1)^{n_1} = \frac{n_0^{\,n_0} n_1^{\,n_1}}{n^n}$$

  - Notice that if x^n is in fact the outcome of a random process, then its empirical distribution (and other statistics) are themselves random variables
    - e.g., expressions of the type $P_{X^n}(\,|\hat{p}(a) - p(a)| \ge \epsilon\,)$

Statistical Models for Data Sources: Finite Memory



i.i.d = Markov of order 0 (*memoryless*)

- *Markov* (or *finite memory*) of order $k \geq 0$

$$P(x_{t+1}|x_1^t) = P(x_{t+1}\big|x_{t-k+1}^t), \quad t \geq k$$

  - We refer to x_{t-k+1}^t as the *context*, or the *state* x_{t+1} occurs in. A^k is referred to as the *state space*.
  - Some convention is needed for $t < k$: *initial state*
    - for example, $x_{-k+1} \cdots x_{-1} x_0 =$ some fixed string
  - Sequence probability

$$P(x_1^n) = \prod_{t=1}^{n} p(x_t|x_{t-k}^{t-1})$$

    - Say $|A| = \alpha$. The $\alpha \cdot \alpha^k$ numbers $p(a|s),\ \ a \in A,\ \ s \in A^k$, together with the fixed initial state, completely define the source.
    - In fact, there are only $(\alpha - 1) \cdot \alpha^k$ *independent parameters*; once we have $p(a|s)$ for $\alpha - 1$ symbols a , the probability of the remaining symbol is fully determined.

Statistical Models for Data Sources: Finite Memory

- Sequence probability

$$P(x^n) = \prod_{t=1}^{n} p(x_t|x_{t-k}^{t-1})$$

- *k-th order Markov empirical distribution*

number of times we see a in state s

$$\hat{p}_{x^n,k}(a|s) = \frac{\left|\{t : 1 \le t \le n,\ \ x_t = a,\ \ x_{t-k}^{t-1} = s\}\right|}{\left|\{t : 1 \le t \le n,\ x_{t-k}^{t-1} = s\}\right|}$$

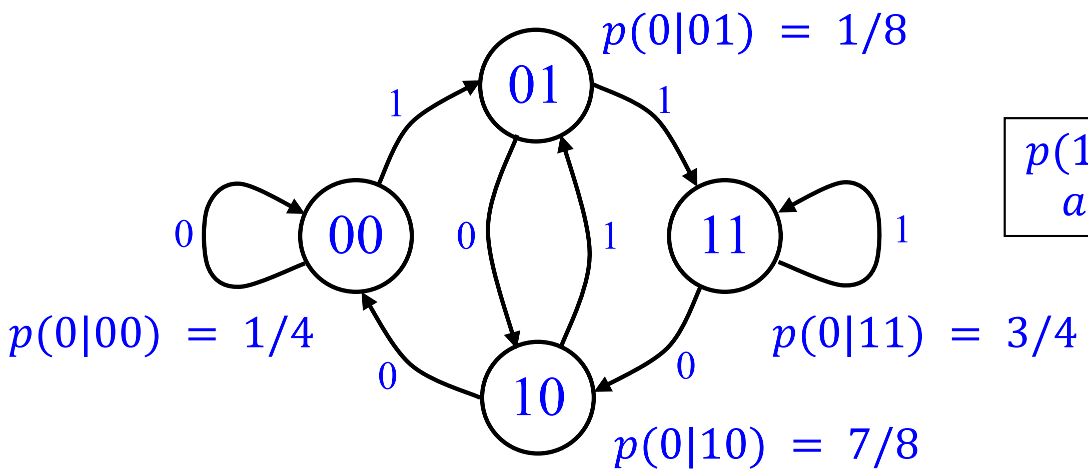
number of times we see state s

- As before, we can ask what probability this distribution assigns to x^n

$$\hat{P}_{x^n,k}(x^n) = \prod_{t=1}^{n} \hat{p}_{x^n,k}(x_t|x_{t-k}^{t-1})$$

  - This is the ML probability of x^n for the class of k-th order Markov models.

Example: Binary finite memory source, $k = 2$



$p(1|ab) = 1 - p(0|ab)$,
$ab \in \{00, 01, 10, 11\}$

- **Steady state:** $\pi_{ab} \stackrel{\text{def}}{=} p_s(ab)$ stationary state probabilities

$$\pi_{ab} = \sum_{cd} \pi_{cd} P(ab|cd), \qquad ab, cd \in \{00, 01, 10, 11\}$$

$$[\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}] = [\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}] \cdot \begin{bmatrix} \frac{1}{4} & \frac{3}{4} & 0 & 0 \\ 0 & 0 & \frac{1}{8} & \frac{7}{8} \\ \frac{7}{8} & \frac{1}{8} & 0 & 0 \\ 0 & 0 & \frac{3}{4} & \frac{1}{4} \end{bmatrix}$$

$$\Rightarrow [\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}] = [\tfrac{7}{26} \quad \tfrac{6}{26} \quad \tfrac{6}{26} \quad \tfrac{7}{26}], \qquad [p_s(0), p_s(1)] = [\tfrac{1}{2} \quad \tfrac{1}{2}]$$

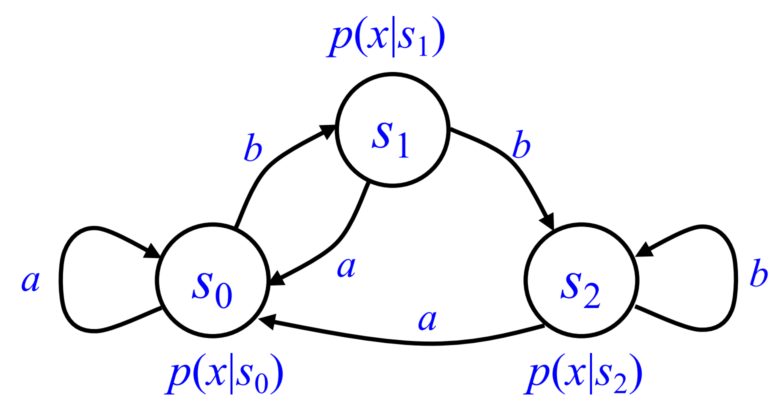
$p_s(b) = \sum_{cd} \pi_{cd} \cdot P(b|cd)$ stationary symbol probabilities

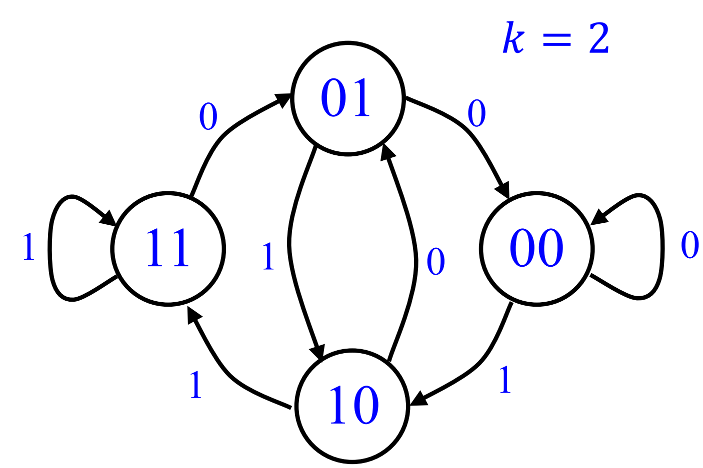
Statistical Models for Data Sources: FSM

- *Finite State Machine (FSM)*
  - *state space* $S = \{s_0, s_1, \dots, s_{K-1}\}$
  - *initial state* s_0
  - *output probability*

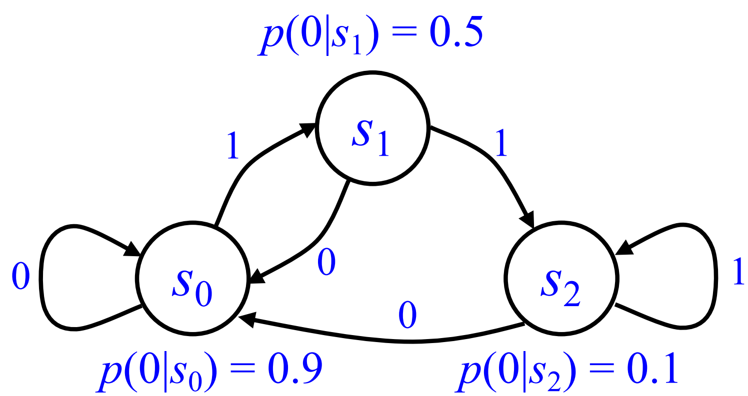
    $p(a|s), \quad a \in A, s \in S$
  - *state transition probability*

    $q(s|s', x), \quad s, s' \in S, \quad x \in A$
  - *unifilar* $\Leftrightarrow$ deterministic transitions: *next-state function* $f: S \times A \to S$
  - every finite memory source is equivalent to a unifilar FSM with $K \leq |A|^k$, but in general, finite state $\neq$ finite memory





Example: Binary FSM



- **Steady state:** $\pi_i \stackrel{\text{def}}{=} p_{\text{stat}}(s_i)$

stationary state probs.

$$[\pi_0\ \pi_1\ \pi_2]\begin{bmatrix} .9 & .1 & 0 \\ .5 & 0 & .5 \\ .1 & 0 & .9 \end{bmatrix} = [\pi_0\ \pi_1\ \pi_2]$$

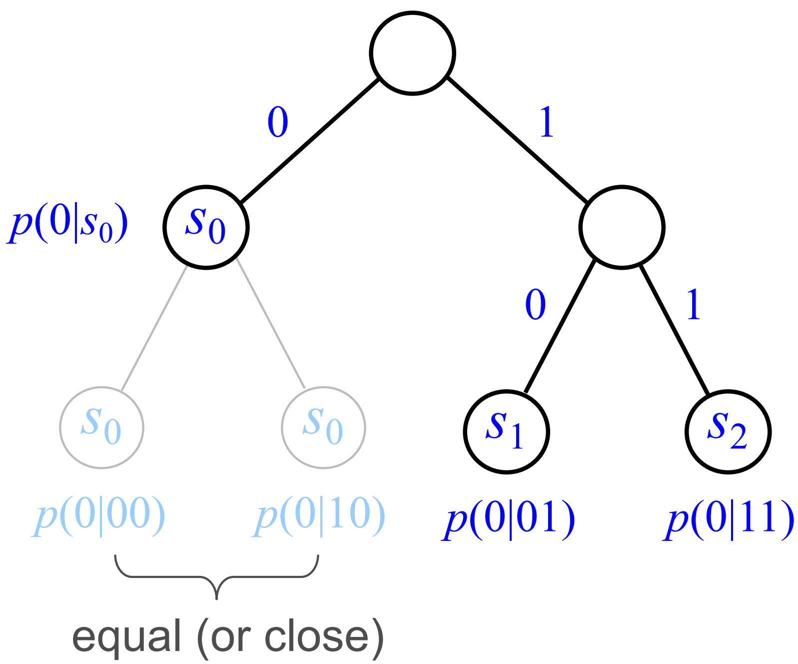
$$\Rightarrow [\pi_0\ \pi_1\ \pi_2] = \begin{bmatrix} \frac{5}{8} & \frac{1}{16} & \frac{5}{16} \end{bmatrix},\ [p_0\ p_1] = \begin{bmatrix} \frac{5}{8} & \frac{3}{8} \end{bmatrix}$$

stationary symbol probs.

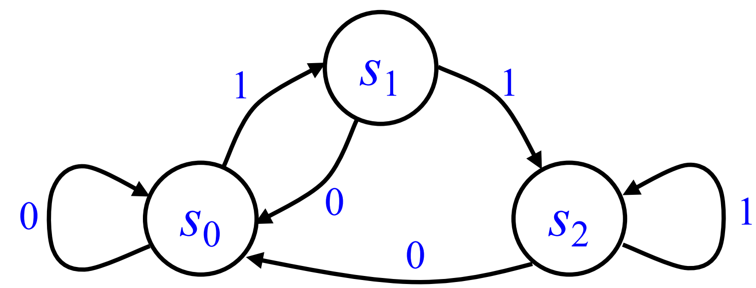
$$p_{\text{stat}}(b) = \sum_i \pi_i \cdot p(b|s_i)$$

Statistical Models for Data Sources: Trees

- *Tree sources*
  - finite memory $\leq k$ (Markov)
  - # of past symbols needed to determine the state might be $< k$ for some states



- by merging nodes from the full Markov tree, we get a model with a *smaller number of free parameters*
- the set of tree sources with unbalanced trees has *measure zero* in the space of Markov sources of any given order
- yet, tree source models have proven very useful in practice, and are associated with some of the best compression algorithms to date



$$s_0 \leftrightarrow 0, \quad s_1 \leftrightarrow 01, \quad s_2 \leftrightarrow 11$$

this tree has an FSM representation (not always the case)

Entropy

$$\boxed{X \sim p(x): \quad H(X) = -\sum_{x \in A} p(x) \log p(x)} \qquad [\log = \log_2,\ 0 \log 0 \stackrel{\text{def}}{=} 0\,]$$

$$H(X) = E_p[-\log p(X)]$$

entropy of X (or of the PMF p), measured in *bits*.

- H measures the *uncertainty* or (the average of the) *self-information* of X.
- We also write $H(p)$: a random variable is not actually needed; $p(\cdot)$ could be an empirical distribution.

Entropy: example

$$\boxed{X \sim p(x): \quad H(X) = -\sum_{x\in A} p(x)\log p(x)}$$ $[\, 0 \log 0 \overset{\text{def}}{=} 0 \,]$

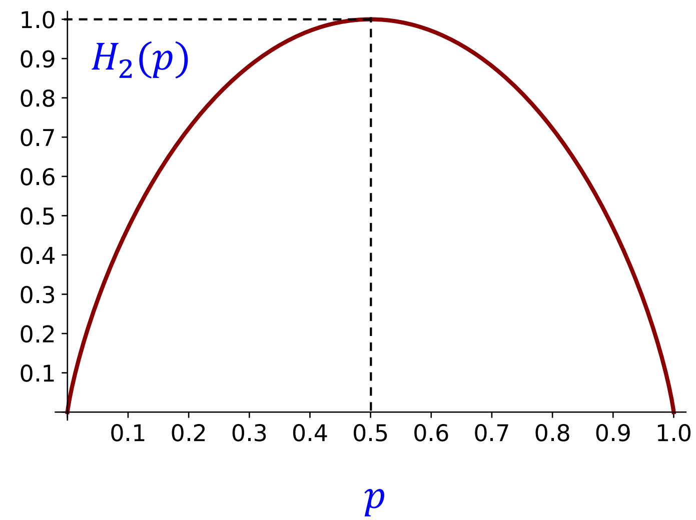
Example: $A = \{0, 1\}$ $\quad P(0) = p, \quad P(1) = 1 - p$

$$\boxed{H_2(p) = -p\log p - (1-p)\log(1-p)}$$ *binary entropy function*

Main properties:

- $H_2(p) \geq 0$, $H_2(p)$ is ∩-convex, $0 \leq p \leq 1$
- $H_2(p) \to 0$ as $p \to 0$ or 1, with slope ∞
- $H_2(p)$ is maximal at $p = 0.5$, $H_2(0.5) = 1$

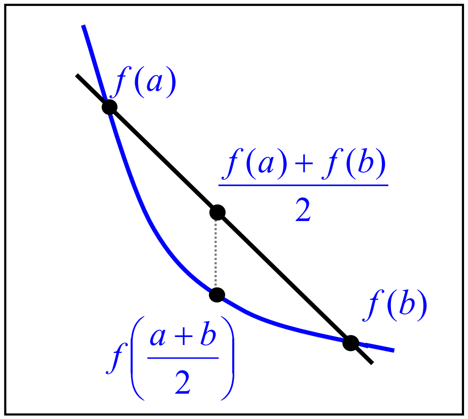
  $\Rightarrow$ the entropy of an unbiased coin is 1 bit



Entropy (cont.)

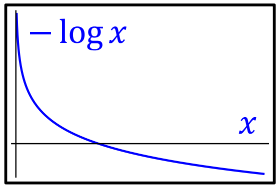
- For a general finite alphabet A, $H(X)$ is maximal when X is *uniformly distributed* , i.e.,

$$X \sim p_u\,, \text{ where } p_{u(a)} = \frac{1}{|A|} \quad \forall a \in A \;.$$



- Proof:
  - Jensen's inequality:

    if f is a $\cup$-convex function and Y a r.v., then $Ef(Y) \geq f(EY)$.
  - $-\log x$ is a $\cup$-convex function of x; set $Y = 1/p(X)$ and $f(Y) = -\log Y$



$$H(X) = E\left[\log \frac{1}{p(X)}\right] = E[\log Y] = -E[-\log Y] \leq -(-\log EY)$$

$$= \log E\left[\frac{1}{p(X)}\right] = \log|A| = H(p_u)$$

Jensen

Joint Entropy

- The *joint entropy* of random variables $(X, Y) \sim p(x, y)$ is defined as

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

  - This can be extended to any number of random variables: $\mathbf{H}(X_1, X_2, \dots, X_n)$.

  Notation:

  $\mathbf{H}(X_1, X_2, \dots, X_n)$ = joint entropy of $X_1, X_2, \dots, X_n$ $\quad (0 \leq \mathbf{H} \leq n \log |A|\,)$

  $$H(X_1, X_2, \dots, X_n) = \frac{1}{n}\mathbf{H}(X_1, X_2, \dots, X_n)$$

  = *normalized per-symbol entropy* $(0 \leq H \leq \log |A|)$

  - If (X, Y) are statistically independent, then $\mathbf{H}(X, Y) = \mathbf{H}(X) + \mathbf{H}(Y)$.

Conditional Entropy

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

- The *conditional entropy* (of Y conditioned on X) is defined as

$$H(Y|X) = \sum_{x} p(x) H(Y|X=x)$$

$$= -\sum_{x} p(x) \sum_{y} p(y|x)\log p(y|x)$$

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

- *Chain rule*

$$\mathbf{H}(X,Y) = H(X) + H(Y|X)$$

- *Conditioning reduces uncertainty*

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

  - but $H(X|Y=y) \ge H(X)$ is possible

Entropy Rates

- *Entropy rate of a random process*

$$H(X_1^\infty) = \lim_{n\to\infty} \frac{1}{n}\mathbf{H}(X_1^n)$$

in bits/symbol,

if the limit exists!

- A related limit based on *conditional entropy*

$$H^*(X_1^\infty) = \lim_{n\to\infty} H(X_n|X_{n-1}, X_{n-2}, \dots, X_1)$$

in bits/symbol,

if the limit exists!

Theorem: For a stationary random process, both limits exist, and

$$H^*(X_1^\infty) = H(X_1^\infty)$$

Entropy rates (examples)

- $X_1, X_2, \ldots$ i.i.d.:

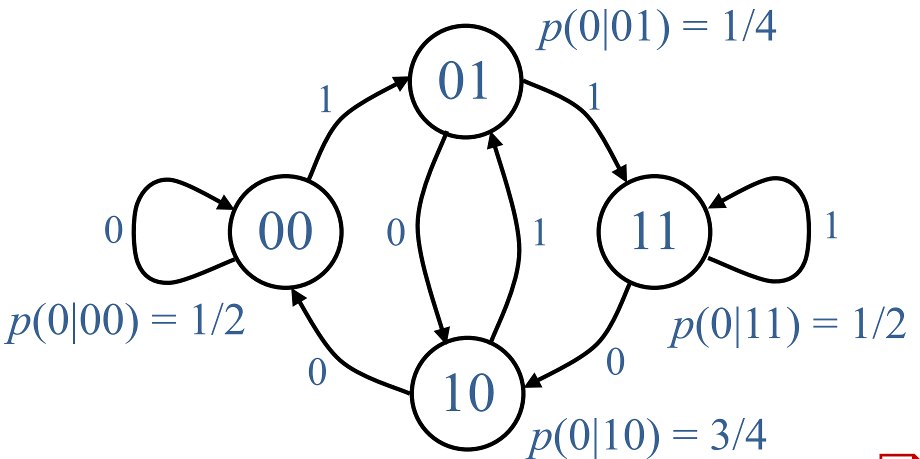
$$H(X_1^\infty) = \lim_{n\to\infty} \frac{1}{n}\mathbf{H}(X_1, X_2, \ldots, X_n) = \lim_{n\to\infty} \frac{1}{n} nH(X_1) = H(\mathrm{X}_1)$$

- X_1^∞ stationary k-th order Markov:

$$\begin{aligned} H(X_1^\infty) &= H^*(X_1^\infty) && \text{theorem} \\ &= \lim_{n\to\infty} H(X_n | X_{n-1}, \ldots, X_1) && \text{definition} \\ &= \lim_{n\to\infty} H(X_n | X_{n-1}, \ldots, X_{n-k}) && \text{Markov} \\ &= H(X_{k+1} | X_k, \ldots, X_1) && \text{stationary} \end{aligned}$$

The theorem provides a very useful tool to compute entropy rates for a broad family of source models

Entropy Rates - Example



steady state:

$$[\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}] = [\tfrac{7}{26} \quad \tfrac{6}{26} \quad \tfrac{6}{26} \quad \tfrac{7}{26}],$$

$$[p_s(0), p_s(1)] = [\tfrac{1}{2} \quad \tfrac{1}{2}]$$

- Markov process entropy

$$H(X|S) = \sum_{ab} \pi_{ab} H\big(p(0|ab)\big) =$$

$$\frac{7}{26}H\left(\frac{1}{4}\right) + \frac{6}{26}H\left(\frac{1}{8}\right) + \frac{6}{26}H\left(\frac{7}{8}\right) + \frac{7}{26}H\left(\frac{3}{4}\right) \approx 0.688$$

- Zero-order entropy

$$H\left(\tfrac{1}{2}\right) = 1$$

- Individual sequence—fitted with Markov model of order $k = 2$

00 | 0000001111111111100000100000001111111101

initial state

state 00

state 11

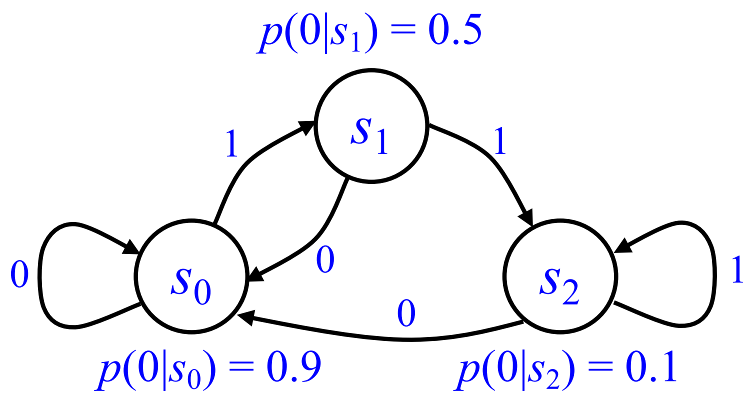
state 01

state 10

Empirical: $\hat{p}(0|s) = [\frac{14}{17}, \frac{2}{4}, \frac{1}{2}, \frac{2}{17}]$, $\quad \hat{p}(s) = [\frac{17}{40}, \frac{4}{40}, \frac{2}{40}, \frac{17}{40}]$, $\quad s = 00, 01, 10, 11$

$$\hat{H}(x^{40}|S) = \frac{17}{40}H\left(\frac{14}{17}\right) + \frac{4}{40}H\left(\frac{1}{2}\right) + \frac{2}{40}H\left(\frac{1}{2}\right) + \frac{17}{40}H\left(\frac{2}{17}\right) \approx 0.657$$

Entropy Rates - Example



Steady state:

$$[\pi_0\ \pi_1\ \pi_2] = \begin{bmatrix} \frac{5}{8} & \frac{1}{16} & \frac{5}{16} \end{bmatrix}, \quad [p_0\ p_1] = \begin{bmatrix} \frac{5}{8} & \frac{3}{8} \end{bmatrix}$$

state probs. symb. probs.

- Zero-order entropy

$$H(0.375) = \ 0.954$$

- Markov process entropy

$$H(X\,|\,S) = \sum_{i=0}^{2} p(s_i) H(p(0\,|\,s_i)) =$$

$$\frac{5}{8} H(0.9) + \frac{1}{16} H(0.5) + \frac{5}{16} H(0.1) \ \approx \ 0.502$$

- Individual sequence - fitted with FSM model

00000011111111111000001000000001111111110 s_0 s_1 s_2

Empirical entropy:

$$\hat{p}(0\,|\,s_0) = \frac{16}{19},\ \hat{p}(0\,|\,s_1) = \frac{1}{3},\ \hat{p}(0\,|\,s_2) = \frac{1}{9}, \quad [\hat{\pi}_0\ \hat{\pi}_1\ \hat{\pi}_2] = \begin{bmatrix} \frac{19}{40} & \frac{3}{40} & \frac{18}{40} \end{bmatrix}, \quad \hat{H}(x\,|\,S) = 0.594$$

Empirical entropy

- Defined for a sequence x^n , relative to a class of models, as

$$\hat{H}(x^n) = -\frac{1}{n}\log(\text{ML probability of } x^n)$$

normalized,
in bits/symbol

- Example: Bernoulli model. Recall

$$\hat{P}_{x^n}(x^n) = \hat{p}(0)^{n_0}\,\hat{p}(1)^{n_1} = \frac{n_0^{n_0} n_1^{n_1}}{n^n} = \left(\frac{n_0}{n}\right)^{n_0}\left(\frac{n_1}{n}\right)^{n_1}$$

using
$n_0 + n_1 = n$

This is the ML probability of x^n relative to the class of Bernoulli models (zero-order Markov).

- We have

$$\hat{H}(x^n) = -\frac{1}{n}\log\hat{P}_{x^n}(x^n) = -\frac{n_0}{n}\log\frac{n_0}{n} - \frac{n_1}{n}\log\frac{n_1}{n} = H_2(\hat{p}(0))$$

in fact, "empirical entropy = entropy of empirical probability" holds for most probability models we are interested in, including Markov models of any order

Relative Entropy

- The *relative entropy* (or *Kullback-Leibler distance*, or *information divergence*) between two PMFs $p(x)$ and $q(x)$ is defined as

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

  - Theorem: $D(p||q) \geq 0$, with equality iff $p = q$.
    - Proof (using strict concavity of $\log$, and Jensen's inequality):

$$-D(p||q) = \sum_x p(x)\log\frac{q(x)}{p(x)} \leq \log\sum_x p(x)\frac{q(x)}{p(x)} = \log\sum_x q(x) \leq 0$$

the summations are over values of x where $p(x)q(x) \neq 0$; other terms contribute either 0 or ∞ to D. Since $\log$ is strictly concave, equality holds iff $\frac{p(x)}{q(x)} = 1\ \forall x$. ■

  - D is not symmetric, and therefore not a distance in the metric sense.
  - However, it is a very useful way to express 'proximity' of distributions.

in a sense, $D(p||q)$ measures the inefficiency of assuming that the distribution is q when it is actually p