Digital vs. Analog Data

Digital data: bits.

$\longrightarrow$ discrete signal

$\longrightarrow$ both in time and amplitude

Analog data: audio/voice, video/image

$\longrightarrow$ continuous signal

$\longrightarrow$ both in time and amplitude

Both forms used in today's network environment.

$\longrightarrow$ burning CDs

$\longrightarrow$ audio/video playback

In broadband networks:

$\longrightarrow$ use analog data to carry digital data

Important task: analog data is often digitized

$\longrightarrow$ useful: why?

$\longrightarrow$ basics: digital signal processing

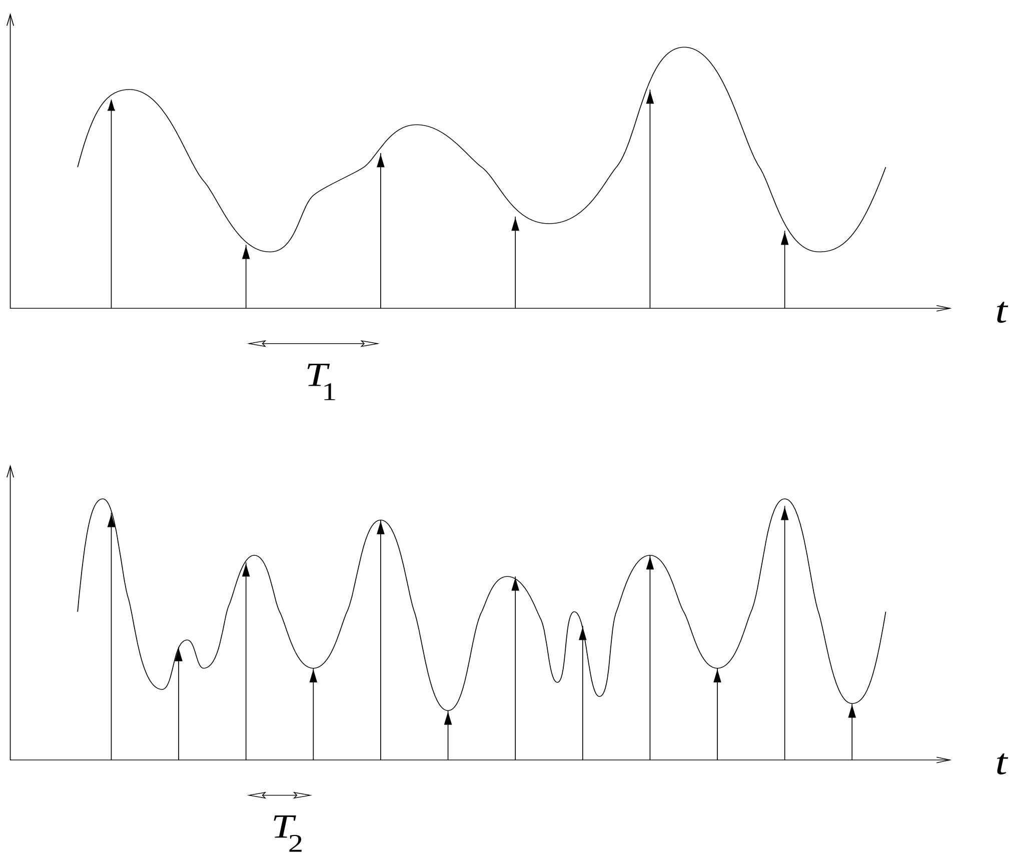
How to digitize such that digital representation is faithful?

$\longrightarrow$ sampling

$\longrightarrow$ interface between analog & digital world

Intuition behind sampling:

$\longrightarrow$ slowly vs. rapidly varying signal



If a signal varies quickly, need more samples to not miss details/changes.

$$\nu_1 = 1/T_1 \;\; < \;\; \nu_2 = 1/T_2$$

Are regularly spaced samples of fixed interval the best?

$\longrightarrow$ perhaps "savings" possible

$\longrightarrow$ the more samples, the more bits

Application: network probing

$\longrightarrow$ goal: ascertain network state

$\longrightarrow$ send sequence of probe packets

$\longrightarrow$ from arriving probes infer/estimate congestion

But, do not want to disturb Schrödinger's cat . . .

$\longrightarrow$ networking: minimize overhead

$\longrightarrow$ irregular probing

$\longrightarrow$ assurance: probabilistic

Sampling criterion for guaranteed faithfulness:

Sampling Theorem (Nyquist): Given continuous bandlimited signal $s(t)$ with $S(\omega) = 0$ for $|\omega| > W$, $s(t)$ can be reconstructed from its samples if

$$\nu \; > \; 2W$$

where ν is the sampling rate.

$\longrightarrow$ ν: samples per second

Issue of digitizing amplitude/magnitude ignored

$\longrightarrow$ problem of quantization

$\longrightarrow$ possible source of information loss

$\longrightarrow$ exploit limitations of human perception

$\longrightarrow$ logarithmic scale

Compression

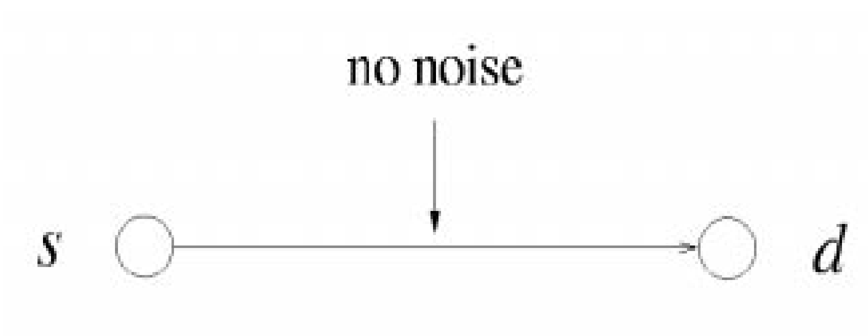
Information transmission over noiseless medium

$\longrightarrow$ medium or "channel"

Sender wants to communicate information to receiver over noiseless channel.

$\longrightarrow$ receive exactly what is sent

$\longrightarrow$ idealized scenario



Set-up:

$\longrightarrow$ take a system perspective

$\longrightarrow$ e.g., modem manufacturer

Need to specify two parts: property of source and how compression is done.

Part I. What does the source look like:

- source s emits symbols from finite alphabet set Σ

  $\rightarrow$ e.g., $\Sigma = \{0, 1\}$; $\Sigma =$ ASCII character set

- symbol $a \in \Sigma$ is generated with probability $p_a > 0$

  $\rightarrow$ e.g., books have known distribution for 'e', 'x' ...

  $\rightarrow$ let's play "Wheel of Fortune"

Part II. Compression machinery:

- code book F assigns code word $w_a = F(a)$ for each symbol $a \in \Sigma$

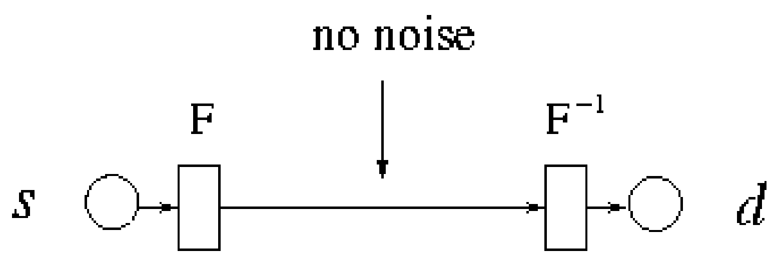
  $\rightarrow$ w_a is a binary string of length $|w_a|$

  $\rightarrow$ F could be just a table

- F is invertible

  $\rightarrow$ receiver d can recover a from w_a

  $\rightarrow$ F^{-1} is the same table, different look-up



Ex.: $\Sigma = \{A, C, G, T\}$; need at least two bits

- F^1: $w_A = 00$, $w_C = 01$, $w_G = 10$, $w_T = 11$
- F^2: $w_A = 0$, $w_C = 10$, $w_G = 110$, $w_T = 1110$

$\longrightarrow$ pros & cons?

Note: code book F is not unique

$\longrightarrow$ find a "good" code book

$\longrightarrow$ when is a code book good?

Performance (i.e., "goodness") measure: average code length L

$$L = \sum_{a \in \Sigma} p_a |w_a|$$

$\longrightarrow$ average number of bits consumed by given F

Ex.: If DNA sequence is 10000 letters long, then require on average $10000 \cdot L$ bits to be transmitted.

$\longrightarrow$ good to have code book with small L

Optimization problem: Given source $\langle \Sigma, \mathbf{p} \rangle$ where $\mathbf{p}$ is a probability vector, find a code book F with least L.

A fundamental result on what is achievable to attain small L.

Entropy H of source $\langle \Sigma, \mathbf{p} \rangle$ is defined as

$$H = \sum_{a \in \Sigma} p_a \log \frac{1}{p_a}$$

Ex.: $\Sigma = \{A, C, G, T\}$; H is maximum if $p_A = p_C = p_G = p_T = 1/4$.

$\longrightarrow$ when is it minimum?

Source Coding Theorem (Shannon): For all F,

$$H \leq L.$$

Moreover, L can be made to approach H.

Remark:

- To approach minimum H use blocks of k symbols

  $\rightarrow$ extension code

- entropy is innate property of source s

- Ensemble limitation

  $\rightarrow$ e.g., sending number $\pi = 3.1415927\ldots$

  $\rightarrow$ better way?