ECE 8873
Data Compression and Modeling

Lecture 2:
Information Source & Modeling

School of Electrical and Computer Engineering
Georgia Institute of Technology
Spring, 2004

Information Source

- A source is an origin of information. A random source is equivalent to a random experiment, which generates outcomes for observation or reception.
- The mechanism that a random source uses to generate information is usually unknown to the observer, who sees only the outcomes of the experiment or the signals the source puts out.
- As in random experiments, an information source is associated with a probability measure, from which one can calculate the entropy of the source.
- When symbols or signals are generated in sequence, the sequential experiments may or may not be independent.

Defining a Source – Parallel to Pr Space

- Sample space, observation space, or signal space built upon a symbol set $A = \{\alpha_i\}_{i=1}^{M}$
  which is also called an alphabet without loss of generality, the symbols α_i are referred to as letters, and m the size of the alphabet.
- Let $\mathbf{X} = (X_1, X_2, X_3, \cdots, X_n)$ be a signal sequence generated by the source. A sequence of length n so generated can be considered as an outcome of a combined experiment with the observation space formed by the cartesian product of the original alphabet: $A^n = A \times A \times \cdots \times A$ and $X_i = \alpha_j \in A$

  Again, the experiments may not be independent.

Shannon's Self-Information

- Let X be an event of a random experiment and $P(A)$ denotes the probability that event X will occur.
- Self-information associated with event X is given by

$$i(X) = -\log_b P(X)$$

- If X and Y are independent events,

$$P(XY) = P(X)P(Y)$$

and thus

$$i(XY) = -\log_b P(X)P(Y) = -\log_b P(X) - \log_b P(Y) = i(X) + i(Y)$$

- When b=2, the unit of information is called bit; if the base is e, the unit is nat; if b=10, the unit is hartley.

Source Entropy

- If X_i are iid (independent & identically distributed), with X denoting a generic random variable as X_i

$$G_n = -\sum_{i_1=1}^{m}\sum_{i_2=1}^{m}\cdots\sum_{i_n=1}^{m}\Pr(X_1=\alpha_{i_1}, X_2=\alpha_{i_2},\cdots,X_n=\alpha_{i_n})\bullet$$
$$\log\Pr(X_1=\alpha_{i_1}, X_2=\alpha_{i_2},\cdots,X_n=\alpha_{i_n})$$
$$= -\sum_{i_1=1}^{m}\sum_{i_2=1}^{m}\cdots\sum_{i_n=1}^{m}\Pr(X_1=\alpha_{i_1})\Pr(X_2=\alpha_{i_2})\cdots\Pr(X_n=\alpha_{i_n})\bullet\sum_{k=1}^{n}\log\Pr(X_k=\alpha_{i_k})$$
$$= -n\sum_{i=1}^{m}\Pr(X=\alpha_i)\log\Pr(X=\alpha_i)$$
$$H(S)=\lim_{n\to\infty}\frac{G_n}{n} = -\sum_{i=1}^{m}\Pr(X=i)\log\Pr(X=i)$$

If the condition of iid is assumed, rather than a given fact, then the above $H(S)$ is called 1st order entropy.

Source Entropy

- The average self-information of such a length-n sequence is

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

- The entropy of the source (per symbol) is defined as

$$H(S)=\lim_{n\to\infty}\frac{G_n}{n}$$

- In the lack of complete knowledge of the experiments, assumptions are often made to facilitate entropy calculation; e.g., iid, Markov, …

Source Entropy

- ❖ True source entropy
  - defined over the true probability space (and the true probability measure of the source); the true amount of information the source produces per "trial", a characteristic quantity of the source.
- ❖ Estimated source entropy
  - Source distribution is usually not completely or precisely known (particularly in sequences resulted from non-independent combined experiments)
    - Limited observations (time, space, …)
    - Limited tools (math, computing implement, …)
  - Source entropy is normally calculated using an estimated source distribution with certain assumed conditions – "communication entropy"

Communication Entropy

- Practical concept of entropy, particularly for information transmission – not exactly for finding bounds or asymptotic results.
- Communication involves "time" – when and what information is being communicated – a departure from Shannon's fundamental framework in which the amount of information may be certified, at times, only after infinite observations.
- The notion of "time" may come from the human perspective (e.g., for real time interactions), or the implementation perspective (e.g., the cost of storing the signal before being coded and sent). This "time element" will put the discussion of "entropy" in a practical light, e.g.;
  - How long is the observation window supposed to be in the configuration? (Also, is the window length fixed?)
  - How much dependency in the sequence of trials can be taken into account?

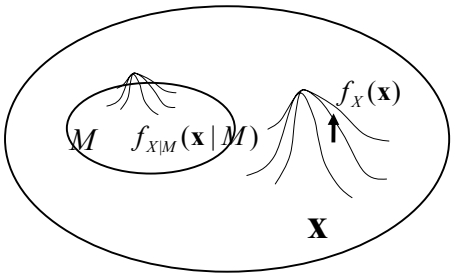
Source Entropy

- Critical question: what is the source model?
- We often assume stationarity and ergodicity in order to be able to analyze the performance of a coding scheme, particularly in terms of the expected rate or expected distortion.
- In reality, we often only deal with finite sequences – i.e. coding of a signal within a finite time span.
- The statistical property of the signal at hand may be "local" – more like a conditional probability space – and a superior performance (again within the limited scope) than that based on the global model may be attainable.

Global vs. Local Models & Adaptive Methods

- Suppose we are interested in a compression scheme for the general English language. The entropy of a source that puts out all possible English sentences would be the lower bound of the compression performance of any code. Suppose we have a code C that achieves this bound.
- Now, if we use this code to compress the Wall Street Journal, is it still going to achieve the optimal performance? The answer is no. A better code probably can be designed using a source model based on the Wall Street Journal.
- We can do at least two things here to improve the original "optimal code":
  - Collect the source material (the WSJ in this example) that best reflects the source statistics and design code based on a model estimated from the collected material;
  - Adapt the global (i.e., general English) model to the local (i.e., the WSJ) model as more of the new signals appear and modify the code accordingly.

Global vs. Local Model



Compare a code design based on $f_X(\mathbf{x})$ - and a code design based on $f_{X|M}(\mathbf{x}\,|\,M)$. If we only deal with signals in M, the latter is very likely to perform better.

Key considerations:
- How easy and reliable is it to determine locality?
- How sensitive is it in terms of performance if such determination is never completely reliable?
- How to represent or inform the receiver of the locality information?

Fundamental Dimensions of Source Coding

- Structure of information (modeling)
  - How is information generated by the source?
  - How to approximate the information-generation process?
  - How to represent this process?
- Random nature of information
  - Efficiency of codes depends on how precise the knowledge the encoder has about the source.
  - How to estimate the source distribution?
  - How to design codes to achieve maximum efficiency given prescribed constraints?

Discovery of Information Structure

- Prior knowledge of the physical process that produces the information
  - Speech: articulatory apparatus, underlying phonetics, syntax, semantics, grammar,…
  - Video: movie settings, scene description, motion, kinematics, …
- Use function approximation to fit data; need to decide on form first; e.g.
  - Linear models: AR, MA, ARMA,…
  - Non-linear models: neural networks, polynomials,..
  - Non-parametric models:
  - Time-varying models:

Structure of Information Source

- The term "structure" refers to the functional relationship among the symbols or signals in the sequence. For example,

$$x_i = f(t_i)$$

$$x_i = f(t_i) + v_i \quad \text{where } v_i \text{ is the noise}$$

$$x_i = \sum_{j=1}^{K} a_j x_{i-j} + v_i \qquad \text{Use regression to find } f$$

Or in terms of Markovian memory

$$\Pr(X_n, X_{n-1}, \cdots, X_1) = \Pr(X_n \mid X_{n-1}, X_{n-2}, \cdots, X_1)\Pr(X_{n-1} \mid X_{n-2}, \cdots, X_1)\cdots\Pr(X_2 \mid X_1)\Pr(X_1)$$
$$= \Pr(X_n \mid X_{n-1})\Pr(X_{n-1} \mid X_{n-2})\cdots\Pr(X_2 \mid X_1)\Pr(X_1)$$

$$\Pr(X_n, X_{n-1}, \cdots, X_1) = \Pr(X_n \mid X_{n-1}, X_{n-2}, \cdots, X_1)\Pr(X_{n-1} \mid X_{n-2}, \cdots, X_1)\cdots\Pr(X_2 \mid X_1)\Pr(X_1)$$
$$= \Pr(X_n \mid X_{n-1}, X_{n-2})\Pr(X_{n-1} \mid X_{n-2}, X_{n-3})\cdots\Pr(X_2 \mid X_1)\Pr(X_1)$$

Function Approximation

$x_i = f(t_i) + v_i$ where v_i is the approximation noise or

$x_i = f(H_i) + v_i$ where

H_i is the state or context at t_i and v_i is the approximation noise

For example, $H_i = \sum_{j=1}^{K} a_j x(t_{i-j}) = \sum_{j=1}^{K} a_j x_{i-j}$

Let us use notation $f_i(\lambda)$ for either $f(t_i)$ or $f(H_i)$

We need to define a criterion for optimization in order to obtain the values of the parameter set λ

$$\text{Error}: \; E = \sum_i d(x_i, f_i(\lambda)) \qquad \text{Likelihood}: \; L = \prod_i l(x_i, f_i(\lambda))$$

Objective: find $\arg\min_{\lambda} E = \arg\min_{\lambda} \sum_i d(x_i, f_i(\lambda))$

Or find $\arg\max_{\lambda} L = \arg\max_{\lambda} \prod_i l(X_i, f_i(\lambda))$

Choice of Criteria

- Squared error

$$E = \sum_i (X_i - f_i(\lambda))^2$$

- Weighted square error

$$E = \sum_i w_i (X_i - f_i(\lambda))^2$$

- Likelihood (log-concave or elliptically symmetric functions)

$$L = K \exp\{-\beta E\} = K \exp\left\{-\beta \sum_i (X_i - f_i(\lambda))^2\right\}$$

$$L = K \exp\{-\beta E\} = K \exp\{-\beta (\mathbf{X} - \mathbf{F})' \mathbf{W} (\mathbf{X} - \mathbf{F})\}$$

$$\mathbf{F}^t = [f_1, f_2, \cdots f_I]$$

Key considerations:
- Number of parameters in λ and coding/quantization of such.
- Ease and consistency in encoding the residue/residual signal.

Examples

X= (1 2 2 3 4 5 4 5 6 7 8 9 8 9 10 11 12 13 14 15)

- Linear model

  $X_i = X(t_i) = ai + e_i \quad \text{where } t_i = i$

  Find a by minimizing $E = \sum_{i=1}^{20} e_i^2 = \sum_{i=1}^{20} (X_i - ai)^2$

- Linear recursive model

  $X_i = X_{i-1} + r_i$

  R=(1 1 0 1 1 1 -1 1 1 1 1 1 -1 1 1 1 1 1 1 1)

- Non-linear model?

  What about X= {(**1** 2)(**2** 3 4 5)(**4** 5 6 7 8 9)(**8** 9 …)}?

  How to transmit it?

Global vs. Local Approximation

- As in discussions of global vs. local source models, structural approximation can be global or local.
- Localized data fitting is similar to building an hierarchy for the representation of functions:
  - Find the locality;
  - Use local approximation functions
- Information transmitted needs to provide ways to retrieve the local structure of the information; again a tradeoff between the structural part and the random part of the information source.

Remarks: Function approximation software can be a term project.