Neural Coding and Decoding

Albert Parker
MATH586 Project
May 4, 2001

Goal of the research: Determine a Coding Scheme: How does neural ensemble activity represent information about sensory stimuli?

GOAL OF THIS TALK: Show how the *Ergodic Theorem*, and the *Shannon-McMillan-Breiman Theorem* justify our approach

Some **MATHEMATICAL CONCEPTS** needed to understand the methodology:

- A **quantizer** is a stochasic map, and a **reproduction space** is the image of a quantizer.

- **Entropy**
  In Information Theory, the entropy of a random variable X is

  $$H = E_x \log p(x)$$

  The concept of entropy was first introduced in thermodynamics to provide a statement of the second law of thermodynamics: the entropy of an isolated system is non-decreasing. In information theory, entropy is desribed as a measure of:

  1. The amount of information required on the average to describe a r.v. (Cover 18)
  2. The number of possible states of a r.v. (Reike 105)
  3. Variability (Reike 118), randomness (Durrett 61), uncertainty, or self information of a r.v. (Cover 12)

- **Joint** and **Conditional** entropy are defined as:

  $$\begin{aligned} H(X,Y) &= E_{x,y} \log p(x,y) \\ H(Y|X) &= E_{x,y} \log p(y|x) \end{aligned}$$

  so $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$

- **Mutual Information**

  $$\begin{aligned} I(X,Y) &= H(X) + H(Y) - H(X,Y) \\ &= \log E_{x,y} \frac{p(x,y)}{p(x)p(y)} \end{aligned}$$

  The amount of information that one r.v. contains about another r.v.. Also called the *relative entropy* or *Kullback Leibler distance* of $p(x,y)$ and $p(x)\dot{p}(y)$.

- The **Typical Sequences** $(x_1, x_2, x_3, ..., x_n)$ are those that one "typically observes" (with probability close to 1). Need SMB Theorem to formalize this.

- What are the associated σ-**algebra's** for X and Y?

SO WHAT?? So now we can formulate a model and describe our methodology:

- **Problem**: It would take an inordinate amount of data to determine the coding scheme between X and Y.
- **The Model**: Consider the problem of determining the coding scheme between X and Y_f, a quantization of Y, such that: Y_f preserves as much mutual information with X as possible and the entropy of $Y_f|Y$ is maximized.
- **Optimization Techniques**
- **Results** with cricket data

PROBABILITY THEORY

Why do we need it? Need justification for assuming things like:

- **Averaging over time** is equivalent to **averaging over experiments**
  Justification: **Ergodic Theorem** - No matter when one starts observing the sequence of r.v.'s, the resulting observation has the same probabilistic structure (Breiman 104).
- **Typical Sequences**
  Justification: **Shannon-McMillan-Breiman Theorem** - One can divide the set of all sequences into two sets, the typical set, where the sample entropy is close to the true entropy, and the non-typical set, which contains all other sequences (Cover 50).

First, some necessary **definitions**

- $\{X_i\}$ is **stationary** if for each n and k, $(X_0, ..., X_n)$ and $(X_k, ..., X_{k+n})$ have the same distribution.
- A measurable transformation $\varphi : \Omega \to \Omega$ is **measure preserving** if $P(\varphi^{-1}A) = P(A) \ \ \forall A \in \mathcal{F}$
- A set $A \in \mathcal{F}$ is **invariant** if $\varphi^{-1}A = A$. Let $\mathcal{I} = \{A|A \text{ is invariant}\}$
- A measurable transformation φ is **ergodic** if $\forall A \in \mathcal{I}, P(A) \in \{0,1\}$
- $X_i = X \circ \varphi^i$ is said to be ergodic if φ is ergodic

ERGODIC THEOREM (Birkhoff 1931)

- φ is a measure preserving transformation on $(\Omega, \mathcal{F}, P)$ and X a r.v. with $E(X) < \infty$. Then
  $$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} X(\varphi^i \omega) = E(X|\mathcal{I}) \text{ a.s.}$$
  (for proof, see Durrett 341-3, Breiman 113-115)

- **Corollary**: SLLN for ergodic processes - If φ is ergodic, then $E(X|\mathcal{I}) = E(X)$.

SHANNON-McMILLAN-BREIMAN THEOREM (1948, 1953, 1957)
If $X_n, n \in \mathbf{Z}$, is an ergodic stationary sequence taking values in a finite set S, then

$$\lim_{n\to\infty} -\frac{1}{n} \log p(X_0, X_1, ..., X_{n-1}) = H$$

where $H \equiv \lim_{n\to\infty} E(-\log p(X_n|X_{n-1}, ..., X_0))$ is the **entropy rate** of $\{X_i\}$.

Remark 1: The **entropy rate** is exactly the **entropy** when $\{X_i\}$ are independent

Remark 2: **Shannon's Theorem** (1948) If $\{X_i\}$ are i.i.d. then

$$\lim_{n\to\infty} -\frac{1}{n} \log p(X_1, X_2, ..., X_n) = H(\mathbf{X}) \text{ a.s.}$$

where $H \equiv E_x \log p(X)$.
proof: *SLLN*

Proof:

- Yes! *Three* more **definitions**. Let

$$\begin{aligned} H_k &\equiv E(-\log p(X_k|X_{k-1}, ..., X_0)) \\ &= E(-\log p(X_0|X_{-1}, X_{-2}, ..., X_{-k})) \text{ by stationarity} \\ H_\infty &\equiv E(-\log p(X_0|X_{-1}, X_{-2}, ...)) \end{aligned}$$

The k^{th} **Markovian approximation** for $k < n$ is

$$p^k(X_0, X_1, ..., X_{n-1}) \equiv p(X_0, ..., X_{k-1}) \prod_{i=k}^{n-1} p(X_i|X_{i-1}, ..., X_{i-k})$$

- **IDEA**: H gets sandwiched between H_k and H_∞.

- **Lemma 1** (Markov approximations):

$$\begin{aligned} \lim_{n\to\infty} -\frac{1}{n} \log p^k(X_0, X_1, ..., X_{n-1}) &= H_k \text{ a.s.} \\ \lim_{n\to\infty} -\frac{1}{n} \log p(X_0, X_1, ..., X_{n-1}|X_{-1}, X_{-2}, ...) &= H_\infty \text{ a.s.} \end{aligned}$$

proof: By the Ergodic Theorem and the definitions of H_k and H_∞

- **Lemma 2** (No gap): $H_k \searrow H_\infty$ and $H = H_\infty$
proof: Martingale Convergence Theorem and *LDCT*

- **Lemma 3** (Sandwich):

$$\begin{aligned} (i) \qquad & \limsup \frac{1}{n} \cdot \log \frac{p^k(X_0, X_1, ..., X_{n-1})}{p(X_0, X_1, ..., X_{n-1})} \le 0 \\ (ii) \qquad & \limsup \frac{1}{n} \cdot \log \frac{p(X_0, X_1, ..., X_{n-1})}{p(X_0, X_1, ..., X_{n-1}|X_{-1}, X_{-2}, ...)} \le 0 \end{aligned}$$

proof: Chebyshev's Inequality and Borel-Cantelli Lemma 1

Now we can formalize the concept of *typical sequences*:

- Given $n, \epsilon > 0$, the **typical set** A_ϵ^n with respect to X is defined as

$$A_\epsilon^n \equiv \{(x_1, x_2, ..., x_n) \in \mathcal{X}^n | 2^{-n(H(\mathbf{X})+\epsilon)} \leq p(x_1, x_2, ..., x_n) \leq 2^{-n(H(\mathbf{X})-\epsilon)}\}$$

  $(x_1, x_2, ..., x_n) \in A_\epsilon^n$ is called **typical sequence**.

- **Asymptotic Equipartition Property**, merely a reformulation of SMB Theorem, gives the following properties of typical sequences:

  1. If $(x_1, x_2, ..., x_n) \in A_\epsilon^n$ then $H(\mathbf{X}) - \epsilon \leq -\frac{1}{n}p(x_1, x_2, ..., x_n) \leq H(\mathbf{X}) + \epsilon$
  2. $P(A_\epsilon^n) > 1 - \epsilon$ for n sufficiently large
  3. $(1 - \epsilon) \cdot 2^{-n(H(\mathbf{X})-\epsilon)} \leq |A_\epsilon^n| \leq 2^{-n(H(\mathbf{X})+\epsilon)}\}$

  Thus the typical set has probability nearly 1, typical sequences are nearly equiprobable and the number of typical sequences is nearly $2^{nH\mathbf{X}}$

REFERENCES

1. L. Breiman. *Probability.* Addison Wesley, Menlo Park, CA, 1968.

2. T. Cover and J. Thomas. *Elements of Information Theory.* Wiley Series in Communication, New York, 1991.

3. A. G. Dimitrov, J.P. Miller, and Z. Aldworth. Non-uniform quantization of neural spike sequences through an information distortion measure. In J. Bower, editor, *Computational Nueral Science: Trends in Research*, 2001 (*to appear*)

4. R. Durrett. *Probability: Theory and Examples.* Duxbery Press, New York, 1997

5. F. Reike, D Warland, R.R. de Ruyter van Steveninck and W. Bialek. *Spikes: Exploring the Neural Code.* The MIT Press, Cambridge, MA, 1997