From Signal to Information Processing

Don H. Johnson
Computer and Information Technology Institute
Department of Electrical & Computer Engineering
Rice University
Houston, Texas
What’s the problem?

Signal processing has been concerned with form, not what the signal represents.
Not all signals are so easy to analyze.
Neural representation of information

Information represented by *when* spikes occur either in *single* neuron responses
Crayfish dorsal light reflex pathway

Photo-reception

Analog Neural Processing

Spike Generation

Motor Integration

Movement (behavior)

Light

Analog Spikes (events)

Head-Down Motoneuron

Brain

Muscle 13A

Lamina

Laminar Monopolar Neurons

Transmedullary Neurons

Sustaining Fiber

Brain

Head-Down Motoneuron

Muscle 13A
Beginnings of information processing

- Information is “in the eye of the beholder”
  - Cellular telephony example (interference to one is information to another)
  - Without interacting with information encoded by a signal, examining signals won’t reveal how well (or if) information is represented
- Signals convey information, but how effectively to they do so?
- Systems process information, selectively suppressing irrelevant information and accentuating important information by acting on signals (information filters)
- System design is usually signal-based, not information based. What effect does system design have on information processing?
Information is always represented—encoded—by signals.

Systems “process information” indirectly by acting on signals.

Result $Z$ is an action or a behavior (i.e., a measurable quantity).

Any viable information processing theory must encompass a variety of signals.

Here, all signals are assumed to be stochastic.
Signals represent information

Let $\square$ represent the **information** encoded in a signal $X(\square)$

Quantify how accurately information **changes** $\square_0 \rightarrow \square_1$ are represented by signals with a *distance measure* $d_X(\square_0, \square_1)$

Diagram:

- Input: $\square_0, \square_1$
- Encoder: $X(\square_0), X(\square_1)$
- Output: $d_X(\square_0, \square_1)$
How to choose a distance?

- Calculate distance between the probability distributions $p_X(x; a_0)$, $p_X(x; a_1)$ characterizing the signal

- Because $p_X(x; \cdot)$ maps the signal domain to the real-line, we can calculate distances regardless of the kind of signal

- Information extraction systems—determining $a$ from $X(a)$—fall into two categories
  - **Classification**: Which of several values of $a$ occurred
    - Optimal classifier is the likelihood ratio test
    - No general formula for performance is known
  - **Estimation**: Determine $a$ from a continuum of values
    - Mean-squared error a frequently used performance measure
Distances and optimal processing

The optimal classifier that tries to determine whether $\square_0$ or $\square_1$ was encoded will have an error probability of the form

$$P_e \sim 2^{d_X(\square_0, \square_1)}$$

Cramér-Rao lower bound on the mean-square error incurred by any (unbiased) estimator

$$E[\hat{\theta}^2] \geq \frac{1}{F(\theta)} \quad \text{(scalar \: \theta)} \quad E[\hat{\theta}] \geq [F(\theta)]^{-1} \quad \text{(vector \: \theta)}$$

$$[F(\theta)]_{ij} = E \left[ \frac{\partial \ln p_X(x; \theta)}{\partial \theta_i} \frac{\partial \ln p_X(x; \theta)}{\partial \theta_j} \right]$$ \quad \text{Fisher information matrix}

Fisher information matrix related to distance induced by small information changes (locally Gaussian property)

$$d_X(\square_0, \square_0 + \delta) \sim K \cdot \delta \cdot F(\theta)$$

With one distance, we can quantify how well information is represented from both classification and estimation viewpoints.
Information processing fundamental

Information-theoretic distance measures obey the Data Processing Theorem:

$$d_X(\square_0, \square_1) \geq d_Y(\square_0, \square_1)$$

Systems cannot increase how well information is represented by their inputs.
Choosing a distance measure

- Many information theoretic distances have the locally Gaussian property
- Only two are known to be related to optimal classifier performance
- We choose distance measures related to the Kullback-Leibler distance

\[ D_X(\theta_1 \| \theta_0) = \sum_x p_X(x; \theta_1) \log \frac{p_X(x; \theta_1)}{p_X(x; \theta_0)} \]

- Choose base-2 logarithms, which gives distance “units” of bits.
**Properties of K-L distance**

- \( D_X(\square_1 \parallel \square_0) \geq 0 \)  
  Equality only when \( p_X(x;\square_1) = p_X(x;\square_0) \)

- \( D_X(\square_1 \parallel \square_0) \neq D_X(\square_0 \parallel \square_1) \) (K-L “distance” is not necessarily symmetric)

- If \( X(\square) \) has statistically independent components,
  \[
  D_X(\square_1 \parallel \square_0) = \prod_{n} D_{X_n}(\square_1 \parallel \square_0)
  \]

- K-L distance is the “exponential rate” of Neyman-Pearson detector’s false-alarm probability
  \[
  P_F \sim 2^{ND_X(\square_1 \parallel \square_0)} \quad \text{for fixed } P_M
  \]

- Distance resulting from information perturbations is “proportional” to Fisher information
  \[
  D_X(\square_0 + \mathcal{N} \parallel \square_0) \propto \frac{\mathcal{F}(\square_0)/\mathcal{F}}{2 \ln 2}
  \]
Distance between LSO response patterns

cumulative KL distance
Analyzing system performance

- Quantify a system’s information processing performance with the information transfer ratio
  \[
  I_{X,Y}(\bar{a}_0, \bar{a}_1) = \frac{d_Y(\bar{a}_0, \bar{a}_1)}{d_X(\bar{a}_0, \bar{a}_1)}
  \]

-\( 0 \leq I_{X,Y}(\bar{a}_0, \bar{a}_1) \leq 1 \)
- If \( I_{X,Y}(\bar{a}_0, \bar{a}_1) = 1 \), the information change is well encoded in the output signal.
- If \( I_{X,Y}(\bar{a}_0, \bar{a}_1) \ll 1 \), the information change is poorly encoded in the output signal.
- Choose a reference \( \bar{a}_0 \); explore how \( \bar{a} \) varies about this point
- Information filtering
Information transfer across a synapse

- Raw, recorded data, contrast = 0.3
- Spikes and membrane potential separated by denoising, (contrast = 0.3)
- Spikes and membrane potential separated by denoising, (contrast = 0.6)
**Information filtering: Array processing**

\[ X = [X_0(t), X_1(t), X_2(t), X_3(t), X_4(t)] \]

\[ X = [X_0(t), X_1(t), X_2(t), X_3(t), X_4(t)] \]

\[ Y(t) \]

\[ Y(n) \]
System theory of information processing

Cascade of systems

\[ X \rightarrow Y \rightarrow Z \quad \mathbb{Q}_{X,Z} = \mathbb{Q}_{X,Y} \cdot \mathbb{Q}_{Y,Z} \]

Multiple input systems

If inputs are independent,

\[ \frac{1}{\mathbb{Q}_{X,Y}} = \prod_n \frac{1}{\mathbb{Q}_{X_n,Y}} \quad \mathbb{Q}_{X,Y} \equiv \min \{ \mathbb{Q}_{X_n,Y} \} \]

Multiple output systems (e.g., neural populations)

\[ \mathbb{Q}_{X,\{Y_1, \ldots, Y_N\}} = \mathbb{Q}_{X,Y_1} + \prod_{n=2}^N \mathbb{Q}_{X,\{Y_n|Y_1, \ldots, Y_{n-1}\}} \]
Non-cooperative populations

- The non-cooperative structure defines a baseline for multi-output systems

- The outputs are \textit{conditionally} independent, \textit{not} statistically independent

\[ p(Y_1, Y_2, \ldots, Y_N; \emptyset) = \prod p(Y_i|x) p(Y_{i+1} | x) \cdots p(Y_N | x) p_X(x; \emptyset) \, dx \]

- The outputs contain only input-induced dependence
Non-cooperative population theory

- Assume each system is not too noisy ($\mathcal{Q}_i \geq \mathcal{Q}_{\min} > 0$)
- As the population size $N$ increases, the population can represent the information expressed by its input without loss, regardless of the information representation

\[
\lim_{N \to \infty} \mathcal{I}_{X,Y}(N) = 1
\]

Continuous code

\[
\mathcal{I}_{X,Y}(N) \approx 1 - \frac{k}{N}
\]

or

Discrete code

\[
\mathcal{I}_{X,Y}(N) \approx 1 - k_1 e^{-k_2 N}
\]
Cooperative populations

If the cooperation among systems involves output feedback to a limited number of other systems, the asymptotics of noncooperative systems apply as well.
Population coding performance limits

Informationally ineffective cooperative structure

\[ \mathbb{E}_{X,Y}(N) \]

Non-cooperative baseline

\[ N \]
Distributed decision systems

What is the most effective way to integrate individual decisions into a global decision?

- Hierarchical
- Democratic
Results: Hierarchical structure
Results: Democratic structure

Democratic Decision System

Decision Probability vs Correct Decision Probability $P$

- $S=1$
- $S=2$
- $S=3$
- $S=4$
- $S=5$
Summary

- A **theory of information processing** must not depend on the nature of the signals representing information.
- The theory presented here uses information theoretic distances, particularly the Kullback-Leibler distance, as the primary tool.
- Data Processing Theorem is a *fundamental* result that can be widely applied.
- Information processing *structures* have fundamental properties regardless of...
  - the information being processed
  - the signals representing the information.
- We can assess signal encoding and system processing, hopefully leading to better designs that focus on the *information*, not the signal.
Collaborators

**Co-Investigators**
- Keith Baggerly
- Raymon Glantz

**Graduate students**
- Michael Lexa
- Chris Rozell
- Sinan Sinanovic

**Undergraduates**
- Michelle Lloyd

**Post-Docs**
- Charlotte Gruner