Information Processing in Dynamical Systems: Foundations of Harmony Theory

Author: P. Smolensky (1986)
Presenter: Dishan Gupta
Theory of Information Processing

- A body of mathematical results, exploiting powerful mathematical techniques, crucial to the development of cognitive science.
- Example TOC models in cognitive science:
  - Von Neumann computer metaphor
  - symbolic paradigm
  - Subsymbolic paradigm
  - Smolensky’s harmonium model (RBMs) based on the subsymbolic paradigm.
Subsymbolic paradigm

- Explores fundamental insights into cognition
- Involves “spread of activation”, relaxation and statistical correlation
- Mathematically based on:
  - Probability theory
  - Theory of dynamical systems – study of sets of variables that evolve in time in parallel and interact through differential equations
What is Harmony Theory?

- A mathematical framework for studying dynamical systems that perform cognitive tasks in accordance to the subsymbolic paradigm.

- Dynamical systems:
  - Human cognition
  - Artificial cognitive systems

- **Goal**: Develop a body of mathematical results for the theory of information processing that complements the classic theory (symbolic) of computation.
What is (NOT) Harmony Theory?

- A mathematical theory (e.g. group theory) providing a language for a class of scientific theories.
- A powerful language for expressing cognitive theories in the subsymbolic paradigm to describe cognition.
- Embodies the fundamental scientific claims of the paradigm, but no commitments on issues such as how knowledge is represented in particular cases.
- Example, regular expressions provide a language for writing compilers, but no scientific theory on how to build a particular one.
Some considerations

- A top-down theoretical strategy to harmony theory.
  - Central idea is that properties of the task are powerfully constraining on mechanism.

- Conceptualize processing in the abyss between
  - Logical reasoning – highest level
  - Sensory processing – lowest level

- Smolensky’s approach, use the task of perception as completing an internal representation of a static state of an external world.
Section Overview

   - Top-down presentation of the perceptual perspective on cognition leads to the basic features of harmony theory.

2. Harmony Theory.
   - Bottom-up presentation of harmony theory that starts with the primitives of knowledge representation.
   - Informal description of theorems.

3. An application of harmony theory
   - Elementary electrical circuit

• Appendix
  - A formal presentation of definitions and theorems

Not covered in this presentation
Section 1

Schema Theory and Self-Consistency
Example cognitive tasks

- The abstract task that Smolensky analyzes captures a common part of the tasks (such as):
  - Passing from an intensity pattern to a set of objects in 3D space,
  - From a sound pattern to a sequence of words,
  - From a sequence of words to a semantic description,
  - From a set of givens in a physics problem to a set of unknowns.

Each of these processes is viewed as completing an internal representation of a static state of an external world.
Suppose you have a headache at a restaurant, and you decide to ask the waitress for aspirin.

How did you create this plan, if you have never had a headache in a restaurant before?

Approach 1: Integrate the two scripts
Knowledge Atoms

- What kind of cognitive system affords this flexibility?
- Suppose there are no scripts:
  - The knowledge base consists of a set of knowledge atoms that configure themselves dynamically in each context to form a set of tailor-made scripts.
  - Schemata are knowledge structures that embody our knowledge of objects, words and other concepts.

*At the time of inference, stored knowledge atoms are dynamically assembled into context-sensitive schemata*
Letter Perception Model

- McClelland and Rumelhart (1981)
- Facilitates letter perception for letters embedded in words.
- Does **NOT** consider schemata as dynamically created entities typically.
- Does not work for orthographically irregular nonwords.

\[ \text{MAVE} \xrightarrow{\text{activates}} \text{MAKE, WAVE, HAVE} \]

- Dynamically created pseudoword schema
Important Points

- Schemata are coherent assemblies of knowledge atoms, only these can support inference.

- The harmony principle. The cognitive system is an engine for activating coherent assemblies of atoms and drawing inferences that are consistent with the knowledge represented by the activated atoms.
Micro and Macro Levels

- Harmony theory exists on two distinct levels of description
  - Microlevel – Involving knowledge atoms, their interaction, development etc.
  - Macrolevel – Involving schemata, their interaction, development etc.
- Analogous to micro (quantum physics) and macro-theories (classical physics) in physics

Formal isomorphism with physics - Understand macroscopic behavior using the universally valid microscopic theory.
The Nature of Knowledge

- Knowledge can be represented in many ways:
  - Words can be single atoms (letter-perception model)
  - Or a combination of activated digraph units
Basic Structure of a Harmony Model

- Two layers: (1) Knowledge atoms, (2) Domain features

Knowledge atoms are fragments of representation that accumulate with experience.

Mathematically, each atom is a vector of $+$ (on), $-$ (off) and 0 (irrelevant).
**The Completion Task**

- A general inferential task in harmony theory, for example:
  - Perception – low-level features of environment given, task is to infer (complete) the remaining features.

![A perceptual completion task](image)

**A procedure for performing the completion task**

<table>
<thead>
<tr>
<th>Input:</th>
<th>Assign values to some features in the representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activation:</td>
<td>Activate atoms that are <em>consistent</em> with the representation</td>
</tr>
<tr>
<td>Inference:</td>
<td>Assign values to unknown features of representation that are <em>consistent</em> with the active knowledge</td>
</tr>
</tbody>
</table>
The Completion Task: Self-Consistency

Assembly of schemata (activation of atoms) and inference (completing missing parts of the representation) are both achieved by finding maximally self-consistent states of the system that are consistent with the input.

High consistency because wherever an active atom is connected to a representational feature by a + (respectively -), connection that feature has the value on (respectively, off).
The Harmony Function

- The self-consistency of a possible state of the cognitive system can be assigned a quantitative value by the harmony function, \( H \).

- A simple harmony function can be one which computer \(+1\) for every agreement and \(-1\) for disagreement for an active atom.

\[
\text{harmony}_{\text{knowledge}} \left( \text{representational feature vector, activations} \right) = \\
\sum_{\text{atoms}} \left( \text{strength of atom } \alpha \right) \left( \begin{array}{l}
0 \text{ if atom } \alpha \text{ inactive; } \\
1 \text{ if active}
\end{array} \right) \text{ similarity of atom } \alpha \text{ to feature vector}
\]

A schematic representation of a harmony function
Probablistic Formulation of Schema Theory

- Take, for example, a child’s birthday party schema.
  - Variables: birthday cake, guest of honor, other guests, gifts, location etc.
  - Schema: May specify, for example, default values, value restrictions, and dependency information for the variables.
- Probability theory to the rescue!
  - Default values: Mode of the marginal probability distribution
  - Value restrictions: value with nonzero probability
  - Dependency: Joint distribution

Each schema encodes the statistical relations among a few representational features. During inference, the probabilistic information in many active schemata are dynamically folded together to find the most probable state of the environment.
How do various schemata become active?

- Go back to micro-level theory: knowledge atoms

- When many schemata become active at once, it means that knowledge atoms that comprise them are simultaneously active.

When atoms, not schemata, are the elements of computation, the problem of coordinating many schemata gets subsumed in the problem of activating the appropriate atoms.
Section 2

Harmony Theory
Knowledge Representation: Representation Vector ($r$)

- A *representational* state of the cognitive system is determined by a collection of values for all the *representational* variables.

- This collection can be designated by a list or vector of ‘+’ or ‘-’s, the *representation* vector ($r$).

- Binary (present +1, absent -1) values contain a tremendous amount of representational power, so it is not a great sacrifice to accept the conceptual and technical simplification they afford.
Knowledge Representation: Activation Vector ($k_\alpha$)

- **Knowledge vector:** Each knowledge atom is characterized by a knowledge vector ($k_\alpha$), which is a list of +1, -1 and 0 values, one for each of representation variable $r_i$.

- **Activation vector:** The list of \{0,1\} values for the activations comprises the activation vector ($a_\alpha$).

- **Strengths:** Knowledge atoms encode subpatterns of feature values that occur in the environment. The different frequencies with which such patterns occur is encoded in the set of strengths \{\sigma_\alpha\} of the atoms.
Harmony Networks: Hierarchy

- The nodes in the representation layer support representation of the environment at all levels of abstraction.
Harmony and Probability: The Harmony Function

- \((r, a)\) describe the state of a cognitive system.
- \(H_k(r, a)\) assigns a real number to the state, using the knowledge base \(K = \{(k_\alpha, \sigma_\alpha)\}\) as parameters.

\[
H(r, a) = H(r_1, a_1) + H(r_2, a_2)
\]

The harmony function is decomposable.
Harmony and Probability: The Harmony Function

- Smolensky’s Harmony Function:

\[ H_K(r, a) = \sum_{\alpha} \sigma_{\alpha} \cdot a_{\alpha} \cdot h_k(r, k_{\alpha}) \]

where,\( h_k(r, k_{\alpha}) = r \cdot \frac{k_{\alpha}}{|k_{\alpha}|} - \kappa \) and \(-1 \leq \kappa \leq 1\)

- When \( \kappa = 0 \), harmony function = \# agreements - \# disagreements (Section 1), i.e., if over 50% of \( k_{\alpha} \) agrees with \( r \), harmony is raised.

- In general, the criterial fraction = \( \frac{1+\kappa}{2} \), for raising harmony.
Estimating Probabilities with H

- The estimated probability of a set of variables:
  \[ p \propto e^{\frac{H}{T}} \]

  Similar to Gibb’s/Boltzmann’s law in physics

where, H = harmony and T = computational temperature

- Isomorphism with statistical physics, where H = Hamiltonian function measuring the energy of a system, and T = temperature of the system.

The lower the computational temperature, the more the estimated probabilities are weighted towards completions of highest harmony.
Derivation of \( p, H \) relation

- According to unconnectedness, the knowledge used in inference process does not relate features in two disconnected networks to each other.

- Thus, \( P(AB) = P(A) \cdot P(B) \). In other words, \textit{adding the harmonies should correspond to multiplying the probabilities}.

Mathematical fact: The only continuous functions that can achieve this are \textit{exponential functions}.

\[
f(x + y) = f(x) + f(y) \\
f(x) = a^x \text{ or } f(x) = e^{\frac{x}{T}} \text{ where } T = \frac{1}{\ln(a)}
\]
Significance of $T$

- $T > 0$, otherwise greater harmony would correspond to smaller probability.
- The magnitude of $T$ is only meaningful once a specific scale has been set for $H$.
  - Indistinguishable systems $a$ and $b$: $\frac{H_a}{T_a} = \frac{H_b}{T_b}$
- $T$ sets the scale for those differences in harmony that correspond to significant differences in probability.

$$\frac{\text{prob}(s_1)}{\text{prob}(s_2)} = \left[ e^{H(s_1) - H(s_2)} \right]^{\frac{1}{T}}$$

*The smaller the value of $T$, the smaller will be the harmony differences that will correspond to significant likelihood ratios.*
Core Theorems

- Three theorems that form the core of harmony theory are informally described:
  - The Competence Theorem
  - The Realizability Theorem
  - The Learnability Theorem
Suppose a cognitive system is capable of observing the frequencies with which each pattern in some pre-existing set \( \{k_\alpha\} \) occurs in the system’s environment.

There will be many environmental distributions consistent with the known pattern frequencies. How to select from these?

For example, two environmental variables \((r_1, r_2)\)

- We know that \(r_1 = +\) occurs 80% of the time
- There are 4 environmental events \((r_1, r_2) \in \{(+, +), (+, -), (-, +), (-, -)\}\)
Principal of Maximal Missing Information

- Two kinds of distributions emerge:
  - Homogeneous – $p(+, -) = 0.4; p(+, +) = 0.4$
  - Inhomogeneous – $p(+, -) = 0.7; p(+, +) = 0.1$
- Use Shannon’s entropy formula $- \sum_x P(x) \ln(P(x))$, 
  - Homogeneous – 0.73
  - Inhomogeneous – 0.48

Of all possible distributions that are consistent with the known frequencies, select the distribution with the maximum homogeneity; the one that supposes the environment to have no more inhomogeneity than is needed to account for the known information.
The Competence Theorem

- Suppose a cognitive system can observe the frequency of the patterns \( \{k_\alpha\} \) in its environment. The probability distribution with the most Shannon missing information that is consistent with the observation is:

\[
\pi(r) \propto e^{U(r)}.
\]

- The maximum likelihood completions of this distribution are the same as those of:

\[
p(r, a) \propto e^{H(r,a)}
\]

where, \( U(r) = \sum_\alpha \lambda_\alpha \xi_\alpha(r) \) and \( \sigma_\alpha = \frac{\lambda_\alpha}{1 - \kappa} \)
The Harmonium Model: Procedure

- **Realizability Theorem**: In the graphical representation of a harmony system (see figure) let each feature node takes values \{-1, +1\} and each knowledge atom take values \{0, 1\} (activation). Let the input to a completion problem be specified by assigning the given feature nodes their correct values (these are fixed throughout the computation). All other nodes repeatedly update their values during computation. The feature not specified in the input are assigned random initial values, and the knowledge atoms initially all have value 0. Let each node stochastically update its value according to the rule:

\[
prob(value = 1) = \frac{1}{1 + e^{-\frac{I}{T}}}
\]

where \(T\) is a global system parameter and \(I\) is the “input” to the node from the other nodes attached to it (see figure).
The Harmonium Model: Procedure (cont.d)

- All the nodes in the upper layer update in parallel
- All the nodes in the lower layer update in parallel, and the process alternates throughout the computation.
- During the update $T$ starts at some positive value and is gradually lowered.
- If $T$ is lowered to 0 sufficiently slowly, then asymptotically, with probability 1, the system state forms the best completion (the one that maximizes the harmony).
The Harmonium Model: Update equations

- The input to each node $I$:
  - Feature Node ($i$):
    \[ I_i = 2 \sum_{\alpha} W_{i\alpha} a_\alpha \]
  - Knowledge Atoms ($\alpha$):
    \[ I_\alpha = \sum_i W_{i\alpha} r_i - \kappa \]
  - where,
    \[ W_{i\alpha} = \frac{(k_\alpha)_i \sigma_\alpha}{|k_\alpha|} \] is the weight between feature node $i$ and knowledge atom $\alpha$. 
Understanding the Stochastic Decision Rule

- If the input to the node is large and positive (i.e. selecting the value 1 would produce much greater system harmony), then it will almost certainly choose the value 1.

- If the input to the node is large and negative (i.e. selection the value 1 would produce much lower system harmony, then it will almost certainly not choose the value 1.

- If the input to the node is 0, then it will choose the value 1 with probability 0.5
The Learnability Theorem

- Suppose states of the environment are selected according to the probability distribution defining that environment, and each state is presented to a cognitive system. Then there is a procedure for gradually modifying the strengths of the knowledge atoms that will converge to the values required by Competence Theorem.
The summary so far ...

- The contribution of harmony theory is not so much in the search procedure for finding the maxima of $H$, but *rather the function $H$ itself*.

- **Importance of Theorem 1:** It gives a high, functional-level characterization of the performance of the system – says what the machine does – and introduces the concept of harmony.

- **Importance of Theorem 2:** It describes a statistical dynamical system that performs completions; it gives an implementation-level description of a kind of completion machine.

- **Importance of Theorem 3:** The harmony function can be tuned with experience.
Conclusions

- This chapter takes a cognitive perspective into the formation of one the most renowned neural network models, the Restricted Boltzmann Machine (RBM), also called the harmonium model in the chapter.
- Even with just two layers, the RBM is able to perform the functions of multi-layer hierarchical networks.
- It is able to learn a rich structures in a manner which is understandable to the average person due to its simplicity. (typically not the case with large and complicated neural networks)