Information Processing in Dynamical Systems: Foundations of Harmony Theory

P. Smolensky

1986
Main Product of the Paper: The Harmonium

- Freund and Hassler (1992) discover that the harmonium is actually the simplest form of a Product of Experts model

- The Harmonium developed in this work is what is later known as the Restricted Boltzmann Machine
Section Overview

1. Schema Theory and Self-Consistency
   – Presentation of how the perceptual perspective on cognition leads to the basic features of harmony theory

2. Harmony Theory
   – How the theorems came about
   – Provide descriptions of the theorems

3. An Application of Harmony Theory
   – Elementary electric circuit

• Appendix
  – A more formal presentation of the math

Not covered in this presentation
Schema Theory and Self-Consistency
What is Harmony Theory?

- A **mathematical framework** for studying a class of dynamical systems that perform *cognitive tasks*

  **Cognition:** A group of mental processes that includes attention, memory, producing and understanding language, learning, reasoning, problem solving, and decision making

  *In cognitive psychology, cognition is typically assumed to be information processing in a an operator’s mind or brain.*

- **Goal:** To provide a powerful language for expressing **cognitive theories**
  - *i.e.*, provides a language for stating alternative hypotheses and techniques and for studying their consequences
Example Cognitive Tasks

- The abstract task the author analyzes captures a common part of the tasks of (for example):
  - *Passing from an intensity pattern to a set of objects in 3D space*
  - *From a sound pattern to a sequence of words*
  - *From a set of patient symptoms to a set of disease states*

*Each of these processes is viewed as completing an internal representation of a static state of an external world.*
Basic Problem

While at a restaurant, you get a headache.

You decide to ask the waitress for some aspirin.

• How did you create this plan? You’ve never had a headache in a restaurant before

Supports the planning and inferences required to reach the usual goal: getting a meal

Need to integrate knowledge that is usually in two separate scripts

'reask waitress for aspirin'
Knowledge Atoms

• What kind of cognitive system is capable of this degree of flexibility?

• Suppose that the knowledge base of the system does not consist of a set of scripts
  – Suppose instead that the knowledge base is a set of knowledge atoms that configure themselves dynamically in each context to form tailor-made scripts

Note: This is the fundamental idea in harmony theory
Letter-Perception Model

MAVE

• In this model, as this stimulus is processed, several word units become and stay quite active, including *MAKE*, *WAVE*, and *HAVE*.

• In this case, the perception of the stimulus is the result of an inference process that is supported by the collection of activated units.
  – This collection is a *dynamically-created pseudoword*.
Representing the Letter-Perception Model

- We can represent words graphically using digraph units at the upper level, *e.g.*:
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.

Note: This principle focuses on the notion of coherency or consistency (i.e., harmony)
The Completion Task

In general, in the completion task, some features of an environmental state are given as input, and the cognitive system must complete that input by assigning likely values to unspecified features.

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 consistent with the representation</td>
</tr>
<tr>
<td>Inference:</td>
<td>Assign values to unknown features of representation that are consistent with the active knowledge</td>
</tr>
</tbody>
</table>
There is high self-consistency here because whenever an active atom is connected to a representational feature by a +/- connection, that feature has a value on/off
An Important Point

The 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 also consistent with the input.
The Harmony Function

- The previous point asserts that a central cognitive process is the construction of cognitive states that are "maximally self-consistent".
  - Therefore, we only need to measure that self-consistency.

- Definitions:
  - The *state of the system* is defined by the set of atoms which are *active* and the vector of all values for all representational features.
  - The *harmony* of the state is the sum of terms, one for each active atom, weighted by the *strength* of that atom.
  - The *self-consistency* is the similarity between the vector and the features defining the atom (the vector of its connections) and the representational feature vector.

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

• To perform a completion, the most probable set of values must be assigned the unknown variables

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*
Harmony Theory

• *Schemata are collections of knowledge atoms that become active in order to maximize harmony*
  – Inferences are also drawn to maximize harmony

• This suggests that the *probability of a possible state of the environment is estimated by computing its harmony*

  The higher the harmony, the greater the probability
Harmony Theory
Knowledge Representation (1)

• **Representation Vector**
  – We have a set of representational features \( r_1, r_2, \ldots \)
    • *e.g.*, (in the environment of visual perception) pixels, edges, depths of surface elements
  – These features constitute the cognitive system’s representation of possible states of the environment
    • Take on binary values: present/absent, true/false

• **Representational State**
  – Determined by a collection of values for all representational variables \( \{r_i\} \)
  – Designated by a vector \( r \) of +’s and –’s
Knowledge Representation (2)

• Each knowledge atom \( \alpha \) is characterized by a knowledge vector \( k_{\alpha} \):
  – A list of +1, -1, and 0 values, one for each representational feature \( r_i \)
  – This is static

• Activation Vector
  – Associated with each knowledge atom \( \alpha \) is its activation variable \( a_{\alpha} \)
    • 1 denotes active, 0, inactive
  – \( \{a_{\alpha}\} \) comprises the activation vector \( a \)
  – This may change depending on the stimulus (dynamic)

• Knowledge atoms encode subpatterns of feature values that occur in the environment
  – The different frequencies with which various such patterns occur is encoded in the set of strengths, \( \{\sigma_{\alpha}\} \) of the atoms

\((k_{\alpha}, \sigma_{\alpha}, a_{\alpha})\) define the knowledge atom \( \alpha \)
Example

Note: All atoms are assumed to have unit strength, but in general, different atoms will have different strengths (which would be indicated above the atom in the drawing)

+/-/0 denote on/off/irrelevant

- Indicates whether there is an excitatory connection, inhibitory connection, or no connection between the corresponding nodes

knowledge vectors

\[
\begin{align*}
k_{W_1A_2} &= (- - + + - 0 0) \\
k_{M_1A_2} &= (- + - + - 0 0) \\
k_{A_2K_3} &= (0 0 0 + - + -)
\end{align*}
\]
Harmony Network Architecture (1)

• A harmony network *always* contains two layers of nodes
  - The nodes in the upper layer correspond to patterns of values in the lower layer

• The nodes in the representation layer support representations of the environment at all levels of abstractness

[Diagram showing the structure of the harmony network with nodes labeled for letter recognition and word recognition.]
Harmony Network Architecture (2)

We can redraw the previous figure:

...alternating representation and knowledge nodes, restoring the image of a series of layers
Why Two Layers?

• The nodes that encode patterns are *not* part of the representation
  – There is a *firm distinction* between representation and knowledge nodes

• The advantages of a two-layer scheme come from simplicity and uniformity
  – There are no connections within layers, only between layers
  – Also simplifies mathematical analysis

• The hope is that by placing all the knowledge in the patterns encoded by knowledge atoms, we will be better able to *understand* the function and structure of the models
Wouldn’t More Stuff be Better?

If the goal is instead to get the most “intelligent” performance out of the fewest number of nodes and connections, it is obviously wiser to allow arbitrary connectivity patterns, weights, and thresholds, as in the Boltzmann machine.

There are theoretical disadvantages to having so many degrees of freedom:

– Too many free parameters in a psychological model make it too theoretically unconstrained and therefore insufficiently instructive

The only way to reasonably expect to make progress is by chaining together many such small steps
Recap

In harmony theory, a cognitive system’s knowledge is encoded in its knowledge atoms. Each atom represents a pattern of values for a few features describing environmental states, values that sometimes co-occur in the system’s environment.

The strengths of the atoms encode the frequencies with which the different patterns occur in the environment.

The atoms are used to estimate the probabilities of events in the environment.
The Harmony Function Revisited (1)

• **Recall**: A state of the cognitive system is determined by the values of the lower and upper level nodes
  – Such a state is determined by a pair \((r, a)\)

• A harmony function assigns a real number \(H_k(r, a)\) to each such state
  – The harmony function has as parameters the set of knowledge vectors and their strengths: \(\{(k_\alpha, \sigma_\alpha)\}\)
    (call this the *knowledge base* \(K\))
The Harmony Function Revisited (2)

• The basic requirement of the harmony function $H$ is that it be **additive under decompositions of the system**
  
  - If the network can be partitioned into two unconnected networks, the harmony of the whole network is the sum of harmonies of the parts:
  
  $$H(r, a) = H(r_1, a_1) + H(r_2, a_2)$$

• Recall the earlier proposed harmony function:

$$H_K(r, a) = \sum_{\alpha} \sigma_\alpha a_\alpha h_\kappa(r, k_\alpha)$$

**Sum over each knowledge atom**
(satisfies the additivity requirement)

**harmony contributed by activating atom $\alpha$, given the current representation $r$**
The Harmony Function Revisited (3)

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

Let  
\[ h_\kappa(r, k_\alpha) = \frac{r \cdot k_\alpha}{|k_\alpha|} - \kappa \]

- The contribution of active atom \( \alpha \) is the product of its strength and the consistency between its knowledge vector \( k_\alpha \) and the representation vector \( r \) (measured by \( h_\kappa(r, k_\alpha) \))

- The parameter \( \kappa \) is always in the interval \([-1, 1]\)
Comments on the Parameter $\kappa$

$$h_\kappa(r, k_\alpha) = \frac{r \cdot k_\alpha}{|k_\alpha|} - \kappa$$

- When $\kappa = 0$, $h_\kappa(r, k_\alpha)$ is the number of representational features whose values agree with the corresponding value in the knowledge vector minus the number that disagree
  - This is the simplest harmony function

- **Problem**: If over 50% of the knowledge vector $k_\alpha$ agrees with $r$, the harmony is raised by activating atom $\alpha$
  - This is a weak criterion for matching, and sometimes it is important to have a more stringent criterion than 50%
  - As $\kappa$ goes from -1 to 0 to 1, the criterion goes from 0% to 50% to 100%
Estimating Probabilities with the Harmony Function

- What we know about harmony functions is that they are additive under network decomposition
  - What, then, is required of the probability assigned to the state? .... *Should be the product of the probabilities assigned to the states of the component networks*

- If adding the harmonies should correspond to multiplying the probabilities of the components’ states \( f(x+y) = f(x) * f(y) \), what function can we use?

\[
\text{Exponential} \quad \Rightarrow \quad f(x) = a^x \rightarrow f(x) = e^{x/T}
\]
Remarks about $T$

• Therefore, let the probability $\propto e^{H/T}$

• **Note:** The sign of $T$ must be positive, for otherwise **greater** harmony would correspond to **smaller** probability

• Consider the following likelihood ratio of two states $s_1$ and $s_2$:

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

**Observation:** $T$ sets the scale for those differences in harmony that correspond to significant differences in probability

The smaller the value of $T$, the smaller the harmony differences that will correspond to significant likelihood ratios. **Thus, once a scale of $H$ has been fixed, decreasing the value of $T$ makes the probability more sharply peaked.**
Environmental Probability Distribution

• Suppose then that a particular cognitive system is capable of observing the frequency with which each pattern in some pre-existing set \( \{k_\alpha\} \) occurs in the system’s environment
  – *Given the frequencies of these patterns, how should the system estimate the probabilities of environmental events? What probability distribution should the system guess for the environment?*

• There will generally be many possible environmental distributions that are consistent with the known pattern frequencies
  – *How to select from all these possibilities?*
Example

- Suppose there are only two environmental features, \( r_1 \) and \( r_2 \)
  - We only know that the pattern \( r_1 = + \) occurs 80% of the time

- There are four environmental events:
  \[
  (r_1, r_2) = \{(+,+), (+,-), (-,+), (-,-)\}
  \]

- For example, we know nothing about the relative likelihood of the two events \((+,+)\) and \((+,-)\); all we know is that together their probability is 0.8
Homogeneity (1)

• One respect in which possible probability distributions differ is in the “degree of homogeneity”
  – A distribution P in which P(+,+) = 0.7 and P(+,-) = 0.1 is less homogeneous than one for which both have probability 0.4

• If the second, more homogeneous distribution applies, then at any given moment there is a greater amount of missing information about the current state of the environment than there is if the more inhomogeneous distribution applies

• Shannon’s formula for missing information of a probability distribution P: 
  \[ -\sum_{x} P(x) \ln P(x) \]

\[
\begin{array}{c|c|c}
& \text{Inhomogenous} & \text{Homogeneous} \\
\hline
\text{Value} & 0.48 & 0.73 \\
\end{array}
\]
Homogeneity (2)

• So one could select the distribution that is most homogenous
  – Principle of maximal missing information

• From the previous example, we would then choose: \( P(+,+)= (+,-) = 0.4, \ P(-,+)=(-,-) = 0.1 \)
  – The justification for choosing this distribution if that there is not enough given information to justify selecting any other distribution with less missing information
Resulting Probability Distribution

- The probability distribution of an environmental state is then given by:

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

where

\[ U(r) = \sum_{\alpha} \lambda_{\alpha} \chi_{\alpha}(r) \]

constrained by the known pattern frequencies =1 when the environmental state \( r \) includes the pattern \( k_{\alpha} \) defining atom \( \alpha \), and =0 otherwise
Back to the Completion Task

• We now have a formula for estimating the probability of an environmental state, so we can in principle perform the completion task

• **Task input:** A set of values for some of the features

• **Best completion:** Formed by assigning values to the unknown features so that the resulting vector $r$ represents the most probable environment state, as estimated by $\pi$
Performing Completions

- It turns out that completions performed by finding the most probable environmental state is exactly the same as those that would be formed using the same procedure with a different distribution:

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

- In passing \( \pi(r) \) to \( p(r, a) \), new variables have been introduced: the activations “a”

- Now to perform the completion the cognitive system must find those values of the unknown \( r \) and \( a \) that maximize the harmony \( H(r, a) \) (and thereby maximize the estimated probability \( p(r, a) \))

This discussion will be summarized in the first theorem
And Finally.... Some Theorems

Theorem 1: *Competence.*

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 observations is:

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

with \( U \) as defined earlier.

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

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

with the harmony function as defined earlier.

Notes: This theorem describes how a cognitive system *should* perform completions, according to some mathematical principles for statistical extrapolation and inference. In this sense, it is a competence theorem.
Theorem 2. Realizability

In the graphical representation of the harmony system let each node denote a processor. Each feature node processor can have a value of +1 or -1, and each knowledge atom a value of 1 or 0 (its 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 the computation. The features 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 following rule:

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

where T is a global system parameter and I is the “input” to the node from the other nodes attached to it. All the nodes in the upper layer update in parallel, then all the nodes in the lower layer update in parallel, and so on alternately throughout the computation. During the update process, T starts out at some positive value and is gradually lowered. \textbf{If T is lowered to 0 sufficiently slowly, then asymptotically, with probability 1, the system state forms the best completion (or one of the best completions if there are more than one that maximize harmony).}
Understanding the Stochastic Decision Rule

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

- If the input to the system is near 0, it will choose the value 1 with probability near 0.5.

\[
\text{prob}(\text{value} = 1) = \frac{1}{1 + e^{-I/T}}
\]
Some More Theorems (Continued)

Theorem 3. *Learnability*

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 Theorem 1.
The summary thus far...

- The contribution of harmony theory is not so much the search procedure for finding 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
The Harmonium

• For a given harmony model, this machine is constructed as follows:
  – Every node in the network becomes a simple processor
  – Every link in the network becomes a communication link between two processors
  – Processors have two possible values (+1/-1 for the representational feature processors, and 1=active, 0=inactive for the knowledge atom processors).
  – The input to a completion problem is provided by fixing the values of some of the feature processors
  – Each of the other processors continually updates its value by making stochastic decisions based on the harmony associated at the current time
  – It is most likely to choose the value that corresponds to greater harmony, but with some probability, it will make the other choice
  – Each processor computes the harmony associated with its possible values by a numerical calculation that uses as the input the numerical values of all the other processors to which it is connected
  – Alternately, all the atom processors update in parallel, and then all the feature processors update in parallel
  – The process repeats many times (all the while lowering the temperature $T$ to zero)
  – It can be proved that the machine will eventually “freeze” into a completion that maximizes the harmony
Conclusions

• This chapter forms the foundation from a cognitive point-of-view of a particular type of neural network: the Restricted Boltzmann Machine

• More in the paper...
  – Demonstration of the general theory with an elementary electric circuit
  – A more formal presentation of the mathematics behind harmony theory
Backup
Definitions

• Schemata:
  – Knowledge structures that embody our knowledge of objects, words, and other concept of comparable complexity
  – Defining properties:
    1. Have conceptual interpretations
    2. Support inference

• Knowledge Atoms:
  – The elementary constituents of which the author assumes schemata to be composed

Note: At the time of inference, stored knowledge atoms are dynamically assembled into context-sensitive schemata.
Schemata - Details

• Schemata are coherent assemblies of knowledge atoms; only these can support inference

• Schemata are created simply by activating the appropriate atoms

• Subassemblies of activated atoms that tend to recur exactly or approximately are the schemata
Definitions for Theorem 2

- To define the input $I$ to each node, it is convenient to assign weights $W_{i\alpha}$ to the link in the graph between atom $\alpha$ and feature $I$, where:

$$W_{i\alpha} = (k_{\alpha})_i \frac{\sigma_{\alpha}}{|k_{\alpha}|}$$

- Using these weights, the input to a node is essentially the weighted sum of the value of the nodes connected to it:

  For feature nodes
  $$I_i = 2 \sum_{\alpha} W_{i\alpha} a_{\alpha}$$

  For knowledge atoms
  $$I_{\alpha} = \sum_i W_{i\alpha} r_i - \kappa$$