27-642/SDE 733
Artificial Neural Networks
Lecture Notes

Dr. Deborah A. Stacey

March, 1992

The Grossberg Models - Additive, Shunting and ART

Introduction to Unsupervised Learning and Feedback Recall

1. Additive (Grossberg) CAM Model
2. Shunting (Grossberg) CAM Model
3. Adaptive Resonance Theory

		McCulloch-Pitts (1943)
Cohen-Grossberg (1983)	Additive (1967)	Brain-State-in-a-Box (1977)
		Boltzmann Machine (1985)
		BAM (1987)
	Shunting (1973)	Masking Field (1978, 1986)

CAM Models in Decreasing Generality

Additive Grossberg

- introduced by Grossberg in 1968.
- single-layer, autoassociative, nearest-neighbour classifier that stores arbitrary analog spatial patterns using either signal Hebbian or competitive learning.
- learns online and operates in continuous time.
- utilizes positive recurrent connections and negative lateral connections.
- the input to each PE can be either positive or negative.

Encoding

- signal Hebbian learning (the passive decay LTM equation)

$$\dot{w}_{ij} = -\alpha w_{ij} + \beta S(a_i^k)S(a_j^k)$$

where w_{ij} is the symmetric ($w_{ij} = w_{ji}$) connection strength, $S()$ is a sigmoid function, α is a positive constant controlling the passive decay and β is a positive constant controlling the signal Hebbian learning term and $\dot{w}_{ij} = \frac{dw_{ij}}{dt}$.

- Grossberg refers to the w_{ij} connections as **long term memory (LTM)**.
- an alternative encoding procedure is the gated decay (competitive) LTM equation:

$$\dot{w}_{ij} = S(a_i^k)[-\alpha w_{ij} + \beta S(a_j^k)]$$

- this only allows changes to the LTM connections that have non-zero signals being emitted from a_i^k.

Recall

- F_A activations are competitive and described by the Additive STM Equation:

$$\dot{a}_i = -\mu a_i + \delta \sum_{j=1}^{n} S(a_j) w_{ji} + I_i$$

where a_i and a_j are activations, I_i is the i^{th} input value, μ is a positive constant controlling activation decay and δ is a positive constant controlling lateral feedback.

- Grossberg refers to the F_A PE activation values as **short term memory (STM)**.
- nearest-neighbour classification
  - the F_A PEs most closely resembling the presented input pattern will become maximally activated.
  - those least resembling it will be nullified.
- during recall
  - the winning F_A PE saturates to 1 and nullified F_A PEs saturate to 0.
  - if activations are processed long enough, only one PE, a_n, will remain active.

Strengths

- ability to classify data in an unsupervised fashion.
- its provisions for online adaptation.
- both features are important if one is unable to obtain a priori data patterns.

Limitations

- signal and noise are equally regarded and therefore both noise and signal are encoded $\rightarrow$ the noise saturation dilemma.
  - resolved by addition of an automatic gain control over the input signal that nullifies noise while abstracting and encoding the signal.
- later input patterns can be similar enough to encoded patterns that the input pattern will create similar activations resulting in re-encoding over the previously encoded pattern and irretrievable loss of that information $\rightarrow$ the stability-plasticity dilemma.
  - solved by adaptive resonance theory, 1976.

Notes

- has been proven globally stable using the Cohen-Grossberg theorem, 1983.
- capacity to store $m = n^3$ patterns where n is the number of F_A PEs.
- AG is stable when placed in VLSI despite the presence of unwanted spurious oscillations that readily occur in assemblies with a large number of amplifiers.
- inherent stability and fault-tolerance make hardware implementations very promising.

Shunting Grossberg

- introduced in 1973.
- single-layer competitive-cooperative, autoassociative, nearest-neighbour classifier of analog patterns.
- uses either signal Hebbian or competitive learning.
- learns online and operates in continuous time.
- there are separate positive and negative inputs to each F_A PE.
- difference between Additive and Shunting Grossberg:
  1. their activation dynamics.
  2. their delineation of positive and negative inputs.
  3. SG also contains automatic gain control recall dynamics that contrast enhances inputs and nullifies noise.

Encoding

- uses either the passive decay LTM or the gated decay LTM equation.

Recall

- the shunting STM recall equation:

$$\dot{a}_i = -\alpha a_i + (\beta - a_i)[S(a_i) + I_i] - (a_i + \mu)[\sum_{j=1}^{n} S(a_j) w_{ji} + J_i]$$

  where

  - I_i is the excitatory (positive) input
  - J_i is the inhibitory (negative) input
  - w_{ij} is the symmetric LTM connection from i to j PEs
  - α is a positive constant controlling passive decay
  - β is a positive constant controlling inhibitory input and lateral feedback.

- automatic gain control – the amplification of signal and nullification of noise – is implemented through the shunting terms $(\beta - a_i)$ and $(a_i + \mu)$ – these balance the input and feedback signals, allowing the PE's activation to remain sensitive to the input signal and shunt noise.

Strengths

- global stability.
- online adaptation.
- unsupervised learning.
- great deal of neurological support.

Limitations

- the stability-plasticity dilemma.
- limited storage.

Adaptive Resonance Theory

- Concepts
  - active regulation of self-organizing learning.
  - recognition by attention and expectation.
- Motivation
  - the human ability to learn recognition codes in real-time through a process of self-organization.

- Design Problems
  - a basic design problem of intelligent systems that autonomously adapt in real-time to unexpected environmental changes $\rightarrow$ the **stability-plasticity dilemma**.
  - how can a learning system be designed to remain plastic (adaptive) in response to significant events while remaining stable in response to irrelevant events?
  - how to preserve previously learned knowledge while continuing to learn new things.
- Key Computational Idea of ART
  - top down learned expectations focus attention upon bottom up information in a way that protects previously learned memories from being eliminated by new learning and enables new learning to be automatically incorporated into the total knowledge base of the system in a globally self-consistent way.

ART Architecture

- neural networks that self-organize stable recognition codes in real-time in response to arbitrary sequences of input patterns.
  - adaptive pattern recognition is a special case of the more general cognitive process of hypothesis discovery, testing, search, classification and learning.

Competitive Learning

- ART models grew out of an analysis of the competitive learning model.
- in response to certain input environments, the competitive learning model has very appealing properties:
  1. proven that, if not too many input patterns form not too many clusters relative to the number of coding nodes in the output layer, learning of the recognition code eventually stabilizes and the learning process elicits the best distribution of LTM traces that is consistent with the structure of the input environment.
  2. it has been shown that a competitive learning model does not always learn a temporally stable code in response to an arbitrary input environment,
     - as input patterns pass into the system through time, the response of the system to the same input pattern can be different on each successive presentation of that input pattern. The response to a given input pattern might never settle down as learning proceeds.
- unstable learning is due to
  1. learning that occurs in response to other intervening inputs.
  2. simple changes in an input environment.
     - changes in the probabilities of inputs.
     - not peculiar to competitive learning models.

ART 1 Architecture

- capable of stably learning a recognition code in response to an arbitrary sequence of binary input patterns until it utilizes its full memory capacity.
- the adaptive weights (LTM traces) oscillate at most once during learning in response to an arbitrary binary input sequence, yet do not get trapped in spurious memory states or local minima.
- encodes a new input pattern, in part by changing the adaptive weights of a **bottom up** adaptive filter, i.e. the pathways from a feature representation field F_1 to a category representation field F_2.
- a second, **top down** adaptive filter, contained in the pathways from F_2 to F_1, aids in self-stabilization (learned expectations).

  1. an input pattern I activates F_1.
  2. F_1 activates the code, v_{j1}, at F_2 (receives the largest signal from F_1).
  3. F_2 reads out its learned top down expectation to F_1 – the bottom up input pattern and top down learned expectation are matched across F_1.
  4. if these patterns are badly matched, F_1 triggers a reset burst to F_2 which shuts off node v_{j1} for the remainder of the coding cycle.
  5. F_1 reactivates the same bottom up signal pattern to F_2 as before.
  6. F_2 reinterprets this signal and another node v_{j2} is chosen.
  7. this parallel search (hypothesis testing) cycle – steps 1 to 6 – repeats itself automatically at a very fast rate until one of three possibilities occur:
     (a) a node v_{jm} is chosen whose top down expectation approximately matches I.
     (b) a previously uncommitted F_2 node is selected.
     (c) the full capacity of the system is used and cannot accommodate I.
  8. if the hypothesis testing cycle ends in an approximate match, then the bottom up input pattern and top down expectation deform the activity pattern $X = (x_1, x_2, \ldots, x_M)$ across F_1 into a net pattern that computes a fusion between the bottom up and top down information – this represents the **attentional focus** of the system.
     - when fusion occurs, the bottom up and top down signal patterns mutually reinforce each other via feedback and the system gets locked into a resonant state of STM activations – now the LTM traces can learn any new information about the input pattern that is represented within the fused activation pattern across F_1.
     - learning occurs only in the resonant state.
     - the system allows one of its prior learned codes to be altered only if an input pattern is sufficiently similar to what it already knows to risk a further refinement of its knowledge.

9. if the hypothesis testing cycle ends by selecting an uncommitted node in F_2, then the bottom up and top down adaptive filters that are linked to this node learn the F_1 activation pattern that is generated directly by the input.
10. if full capacity has been reached – learning is automatically inhibited.

Summary

- an ART 1 network either
  1. refines its already learned codes based on new information that can be safely accommodated into them via approximate matches, or
  2. selects new nodes for initiating learning of novel recognition categories, or
  3. it defends its fully committed memory against erasure by new input events (the **capacity catastrophe**).
- non-self-stabilizing learning systems are not capable of functioning autonomously in ill-controlled environments.
- learning in the approximate mode enables rapid and stable learning to occur while buffering the system's memory against external noise.
- hypothesis testing cycle replaces internal system noise as a scheme for discovering a globally correct solution and does not utilize an externally controlled temperature parameter or teacher.

Attentional Priming and Prediction

Matching by the $\frac{2}{3}$ Rule

- regulation of the hypothesis testing cycle and the self-stabilization of learning.
- necessary to assume that F_1 can distinguish between bottom up and top down signals.
- third F_1 input source $\rightarrow$ an attentional gain control channel.
- when it is activated, it excites each F_1 node equally.
- at least two out of three input sources are needed to **supraliminally** activate an F_1 node – a bottom up input, a top down input and a gain control input.
  - supraliminally activated – activated enough to generate output signals to other parts of the net and initiate the hypothesis cycle.
- in top down processing mode
  - each F_1 node receives a signal from at most one source – **subliminally activated**.
  - sensitize, prepare or attentionally prime F_1 for future input patterns that may or may not generate an approximate match with this expectation – enables ART to anticipate future events or if locked into place (high gain top down) automatically suppresses all inputs not in a category while amplifying those that are.

- in bottom up mode
  - each active bottom up pathway can turn on the gain control node.
  - then all F_1 nodes receive a gain control input but only those with bottom up input are supraliminally activated.
- when both bottom up and top down inputs reach F_1, the gain control is shut off, so that only those F_1 nodes which receive top down confirmation of the bottom up input are supraliminally activated.

The Orienting Subsystem

- ART 1 architecture
  1. attentional subsystem
  2. orienting subsystem
- orienting subsystem
  - generates an output signal only when a mismatch occurs at level F_1 of the attentional system.
  - functions as a **novelty detector**.
  - output signal is called an **STM reset wave** because it selectively inhibits node(s) at level F_2 of the attentional subsystem.
  - it disconfirms the F_2 hypothesis that led to the F_1 mismatch.

The Orienting Subsystem and the $\frac{2}{3}$ Rule

- when a bottom up input pattern is presented, each of the active input pathways to F_1 also sends a signal to the orienting subsystem where they are summed.
- when the input pattern activates F_1, each of the activated F_1 nodes sends an inhibitory signal to the orienting subsystem.
- the total inhibitory signal is larger than the excitatory signal.
- thus in bottom up mode – balance between active F_1 nodes and active input lines that prevents a reset wave from being triggered.
- the balance is upset when a top down expectation is read out that mismatches.
- the total output from F_1 then decreases by an amount that grows with the severity of the mismatch.
- if the attenuation is sufficiently great, then inhibition from F_1 to the orienting subsystem can no longer prevent the orienting subsystem from emitting a reset wave.

Vigilance

- a parameter ρ (**vigilance** parameter) determines how large a mismatch will be tolerated before the orienting subsystem emits a reset wave.
- high vigilance
  - searches for new categories in response to small differences between input and expectation.
  - large number of fine categories.
- low vigilance
  - enables the tolerance of large mismatches.
  - grouping according to a coarse measure of mutual similarity.

ART 2

- developed to handle analog patterns (1987).
- can autonomously classify arbitrary sequences of analog input patterns into categories of arbitrary coarseness while suppressing arbitrary levels of noise.
- different versions have been developed for
  - visual pattern recognition
  - speech perception
  - radar classification.

ART 2 Characteristics

- level F_1 is split into separate sublevels
  - for receiving bottom up patterns
  - for receiving top down patterns
  - for matching the bottom up and top down data.
- interfacing level of interneurons that matches the transformed bottom up and top down information and feeds the results back to the bottom and top F_1 levels.

Design Principles

Stability-Plasticity Tradeoff

See discussion in ART 1 section.

Search-Direct Access Tradeoff

- ART 2 carries out a parallel search in order to regulate the selection of appropriate recognition codes during learning, yet automatically disengages the search as an input pattern becomes familiar.
- thereafter the familiar input directly accesses its recognition code.

Match-Reset Tradeoff

- system should be able to recognize and react to arbitrary small differences between an active F_1 STM pattern and the LTM pattern being read out from an established category.
- also, when an uncommitted F_2 node becomes active for the first time, it should be able to remain active, without being reset, so that it can encode its first input exemplar (no bottom up top down match present).
- a combination of an appropriately chosen ART 2 reset rule and LTM initial values work together to satisfy both processing requirements.
- ART 2 parameters can be chosen so that learning increases the system's sensitivity to mismatches between bottom up and top down patterns.

STM Invariance under Readout of Matched LTM

- prevents readout of a perfectly matched LTM pattern from causing reset by preventing any change from occurring in the STM patterning at the lower F_1 levels.
- extra F_1 levels provide enough degrees of computational freedom to
  - both readout top down LTM and normalize the total STM pattern at the top F_1 level
  - before this normalized STM pattern can interact with the middle F_1 level at which top down and bottom up information are matched.
- also allows for the compensation for fluctuations in baseline activity levels $\rightarrow$ prevents spurious resets and destabilization of the search and learning processes.
- the $\frac{2}{3}$ rule of ART 1 is realized as part of the F_1 internal levels.
  - a superset bottom up input pattern cannot recode a subset top down expectation.

Noise Suppression

- a combination of normalization and nonlinear feedback processes within F_1 determines a noise criterion and enables the system to separate noise from signal.
  - contrast enhance the F_1 STM pattern and the learned LTM patterns.

- degree of contrast enhancement and noise suppression is determined by the degree of nonlinearity in the feedback signal functions at F_1.

- a nonlinear signal function operating on the sum of normalized bottom up and top down signals also correlates these signals.
  - helps to attenuate the total activation of F_1 in response to mismatched patterns, as well as to contrast enhance and noise suppress bottom up patterns.

Rapid Self-Stabilization

- learning in ART needs to be slow relative to the STM processing rate but no restrictions are placed on absolute rates.
- ART 2 is capable of stable learning in "fast learning mode"
  - LTM traces change so quickly that they can approach new equilibrium values on every trial.

Normalization

- achieved by
  1. nonspecific inhibitory neurons
     - each normalizer uses $O(M)$ connections where M is the number of nodes to be normalized.
  2. shunting on-centre off-surround network
     - uses $O(M^2)$ connections.

Local Computations

- STM and LTM computations use only information available locally and in real-time.
- no assumptions of weight transport (as in back-propagation).
- no assumptions of an a priori input probability distribution (as in simulated annealing).

The $\frac{2}{3}$ Rule

- ART 2 implements a weak version of the $\frac{2}{3}$ rule – during matching, an F_1 node can remain active only if it receives significant top down input.
  - a node receiving large top down input can remain stored in memory even if bottom up input to the node is absent on a given trail – it would be partially restored in STM.
  - the relative importance of the feature would decline but not necessarily disappear.
  - a feature consistently absent from most category exemplars would eventually be removed from the category expectation pattern.
  - this feature thereafter would be treated as noise.

Recognition, Reinforcement and Recall

- reinforcement (reward and punishment)
  - calibrate whether an action has satisfied an internal need.
  - can modify the formation of recognition codes and shift attention to focus upon those codes whose activation promises to satisfy internal needs based upon past experience.
- recall
  - can generate equivalent responses or actions to input events that are classified by different recognition codes.

Carpenter-Grossberg Experiments

- models of self-organizing biological systems wherein all the ingredients of recognition, reinforcement and recall are joined in a single integrated circuit.
- an ART 2 system self-organizes recognition categories in response to the preprocessed inputs, its categorical choices at the F_2 level self-stabilizing through time.
- this system's choices can be used as the first level of an ART 1 or another ART 2 architecture to produce another classifying level F_3.
- level F_3 can be used as a source of prewired priming inputs to F_2.
- after learning these primes, a particular prime can activate a learned $F_3 \rightarrow F_2$ top down expectation.
- the prime causes the architecture to pay attention only to expected sources of input information.
- the output pathways from level F_2 of ART 2 to the postprocessor can learn to recall any spatial pattern by applying theorems about associative learning.
- the architecture as a whole can stably self-organize an invariant recognition code and an associative map to an arbitrary format of output patterns.

Reference Material

1. Carpenter, Gail A. and Stephen Grossberg, "Self-Organizing Neural Network Architectures for Real-Time Adaptive Pattern Recognition" in **An Introduction to Neural and Electronic Networks** edited by S.F. Zornetzer, J.L. Davis and C. Lau, 1989 (pages 455-478).

2. Carpenter, Gail A. and Stephen Grossberg, "A Massively Parallel Architecture for a Self-Organizing Neural Pattern Recognition Machine", in **Neural Networks and Natural Intelligence** edited by Stephen Grossberg, 1988 (pages 251-315).

3. Simpson, Patrick K., **Artificial Neural Systems: Foundations, Paradigms, Applications, and Implementations**, 1990.

   - Chapter 5: pages 30-46.

4. Grossberg, Stephen, "Nonlinear Neural Networks: Principles, Mechanisms, and Architectures," Neural Networks, Volume 1, Number 1, 1988, pages 17-61.

5. Pao, Yoh-Han, **Adaptive Pattern Recognition and Neural Networks**, 1989, page 194.