Universal Architectures for Progressive Machine Learning: Model, Performance Evaluation and Applications

John S. Baras

(joint work with Christos N. Mavridis)

Institute for Systems Research

ECE, CS, ME, AE, BioEng., DOIT, AMSSC

University of Maryland College Park

USA

January 15, 2024

AUTH

Thessaloniki, Greece

Towards Intelligent Autonomous Systems

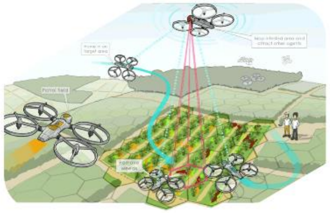
Robotics

Multi-Agent Systems

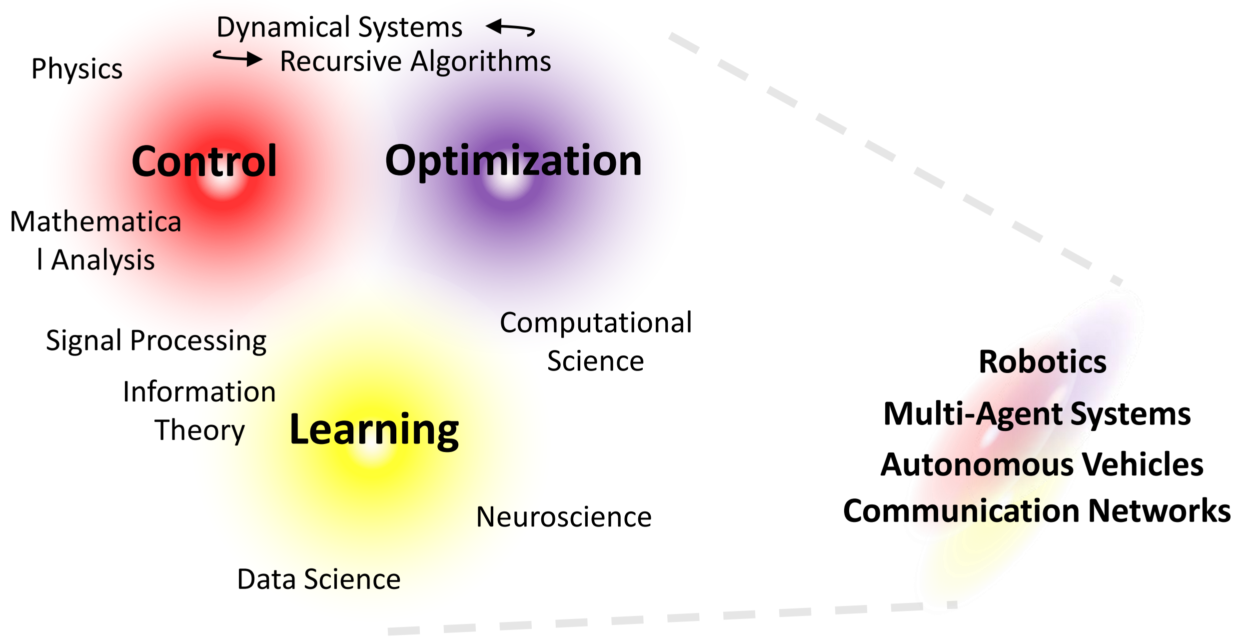
Autonomous Vehicles

Communication Networks

Healthcare

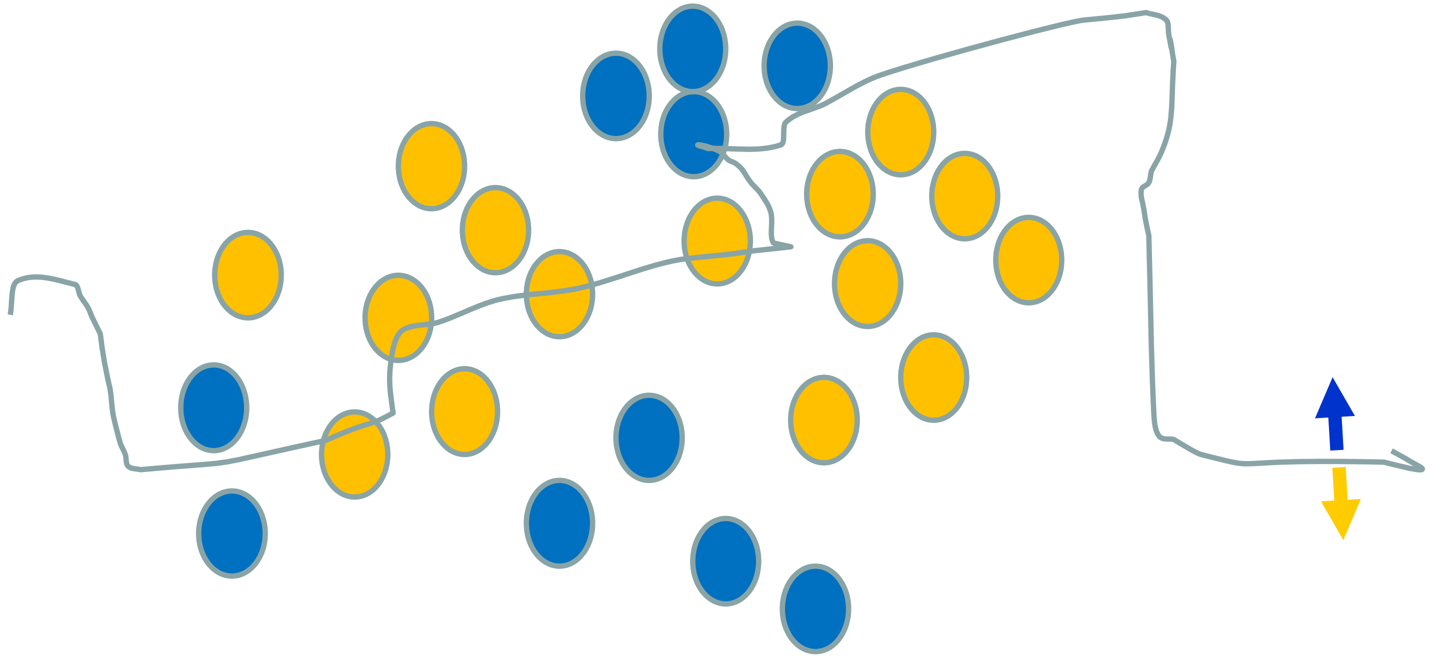


Control, Optimization, and Learning



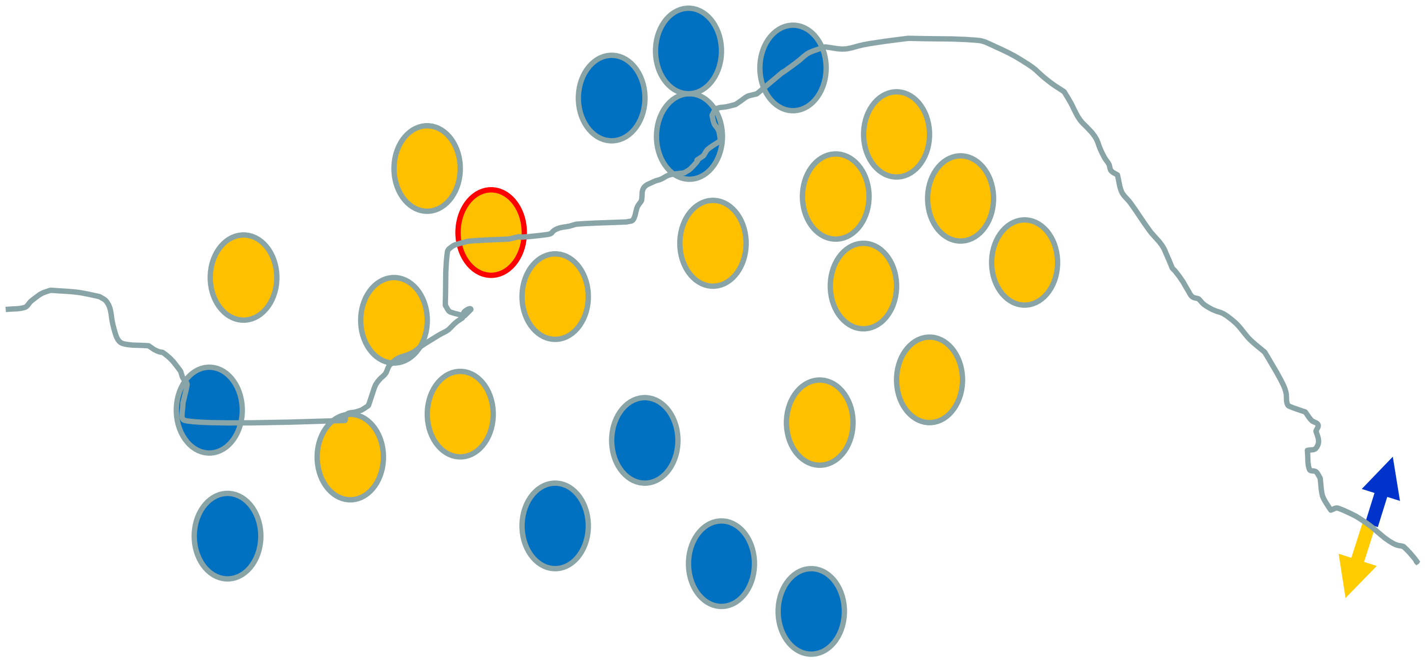
Progressive Decision Boundary

Initial random weights



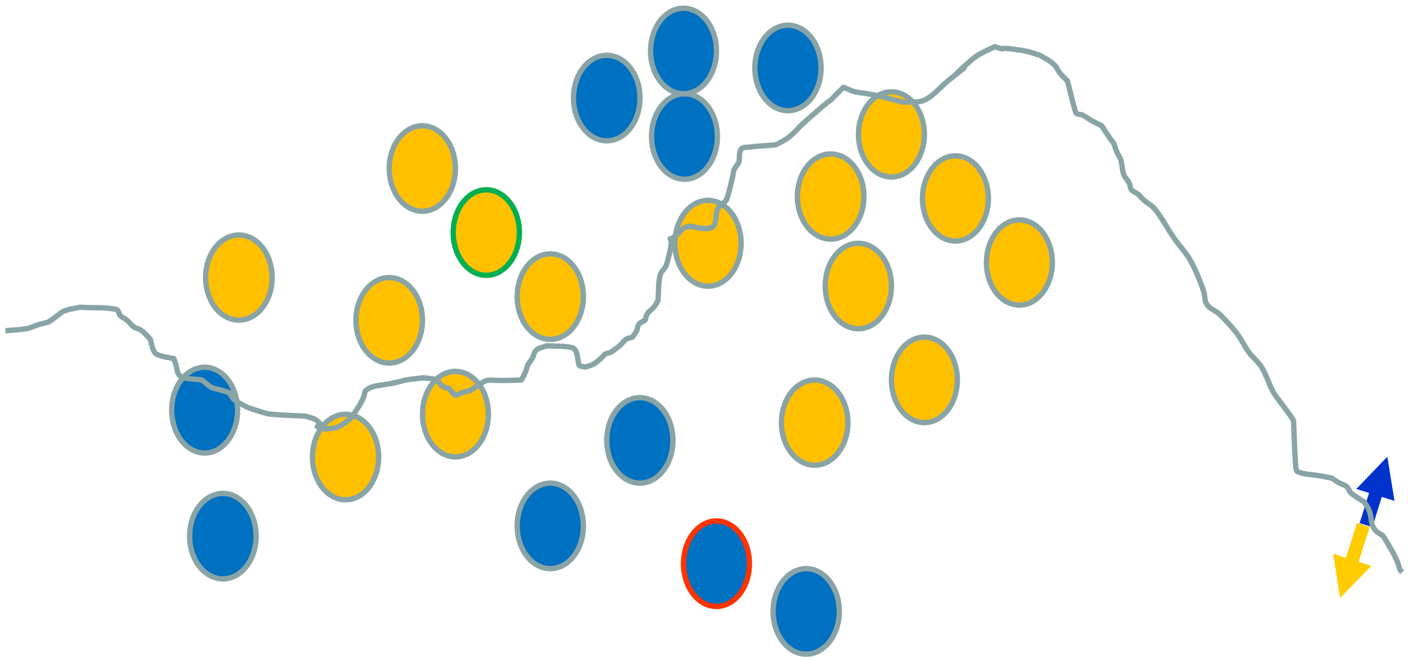
Decision Boundary (cont.)

Present a training instance / adjust the weights



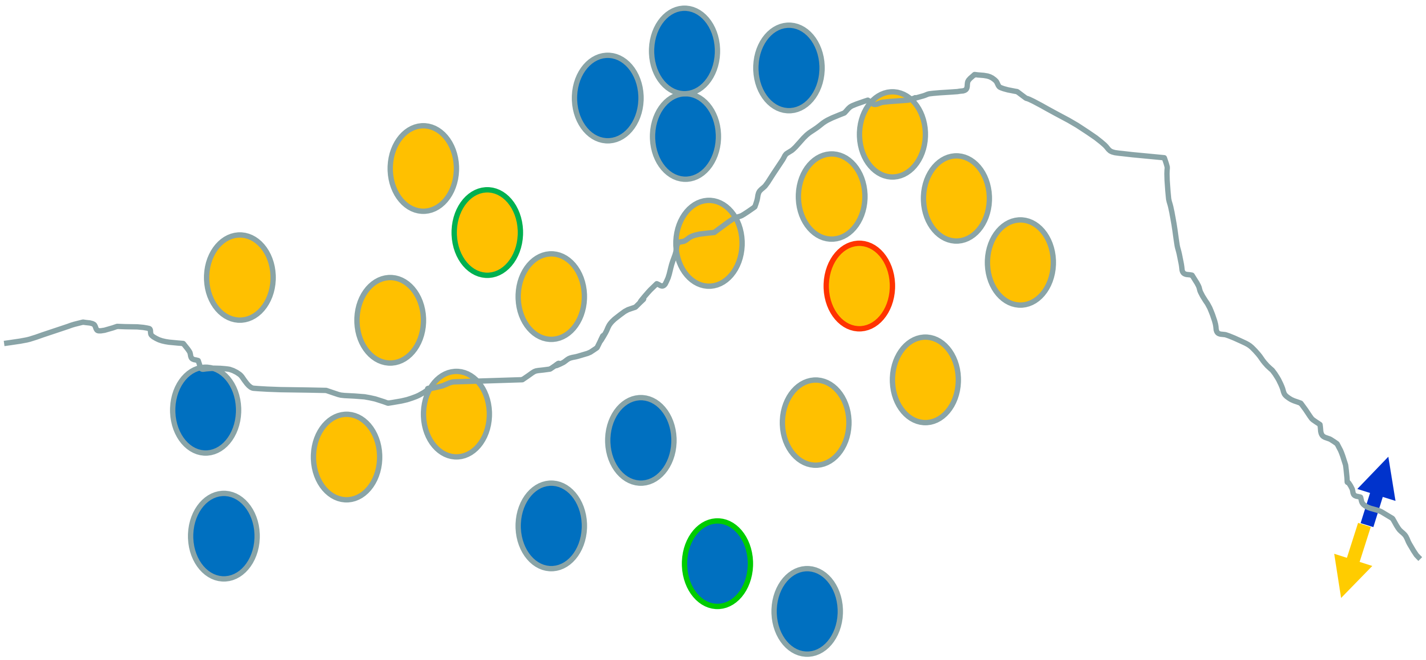
Decision Boundary (cont.)

Present a training instance / adjust the weights



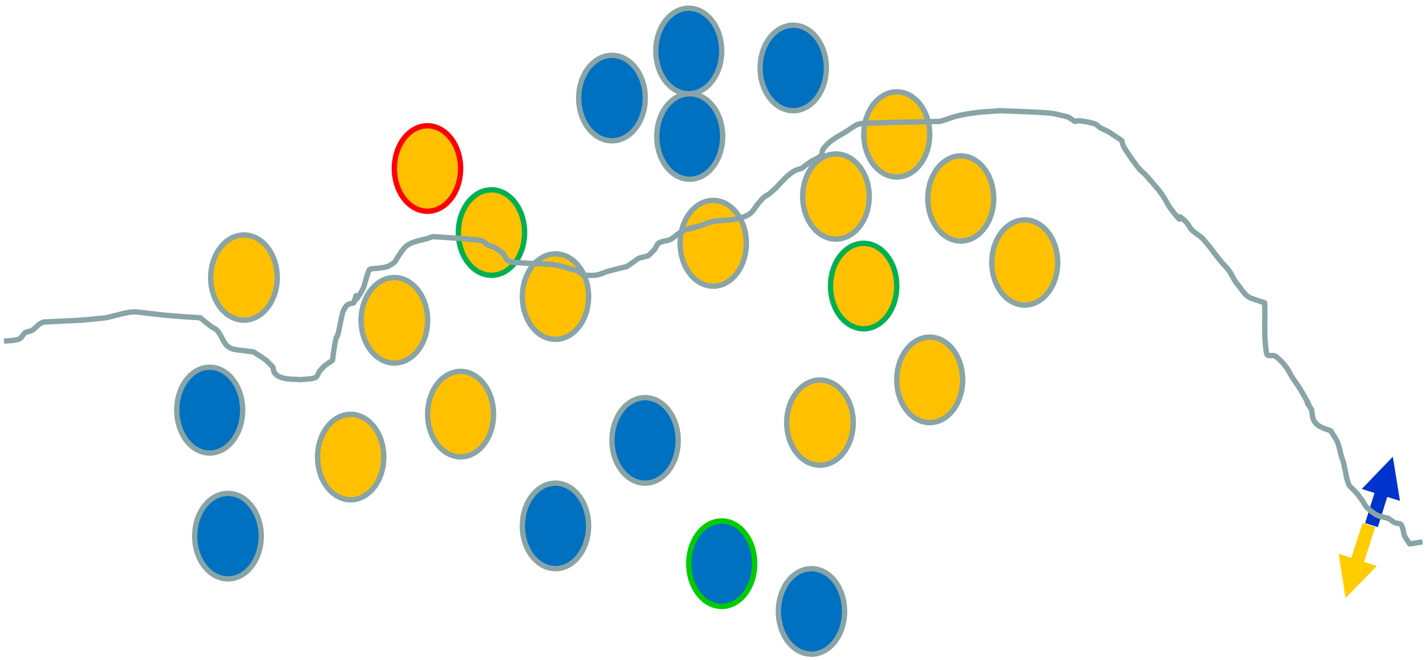
Decision Boundary (cont.)

Present a training instance / adjust the weights



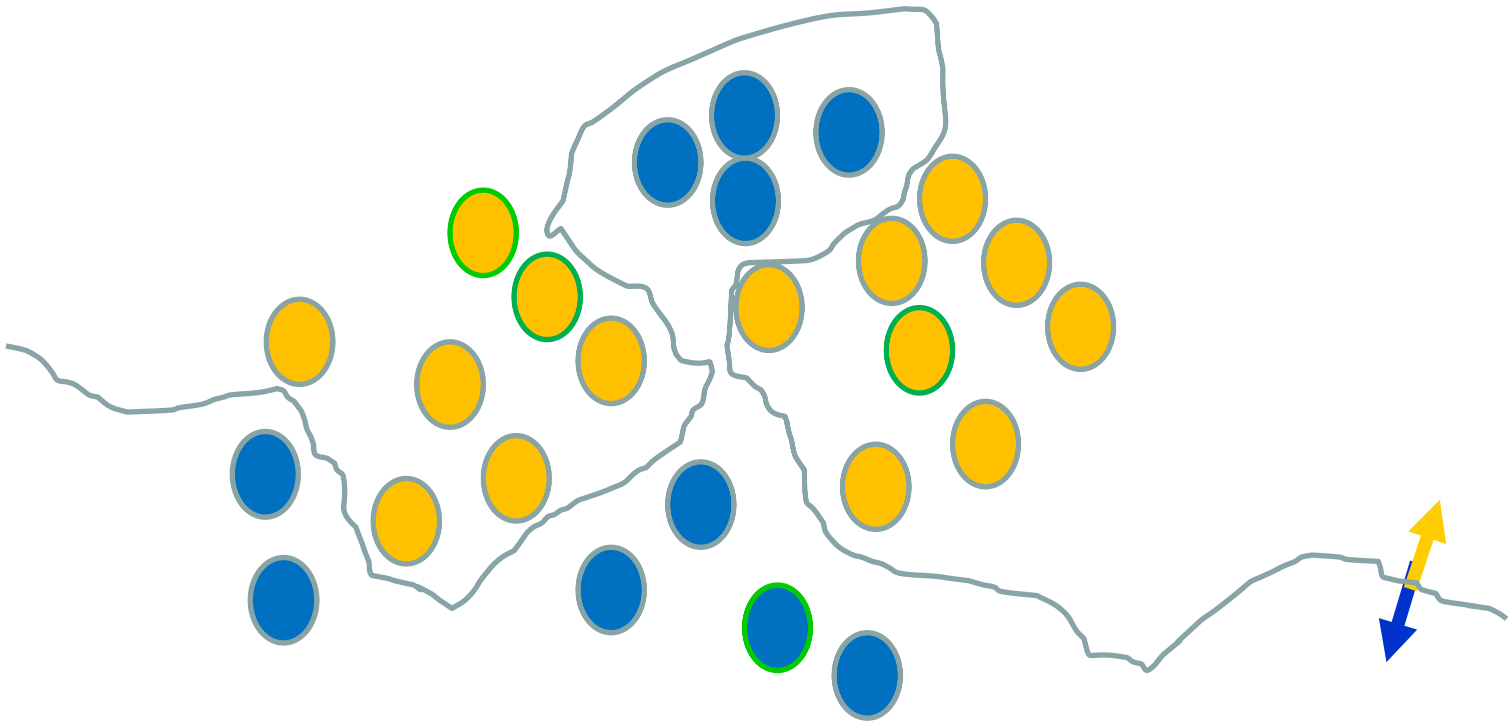
Decision Boundary (cont.)

Present a training instance / adjust the weights



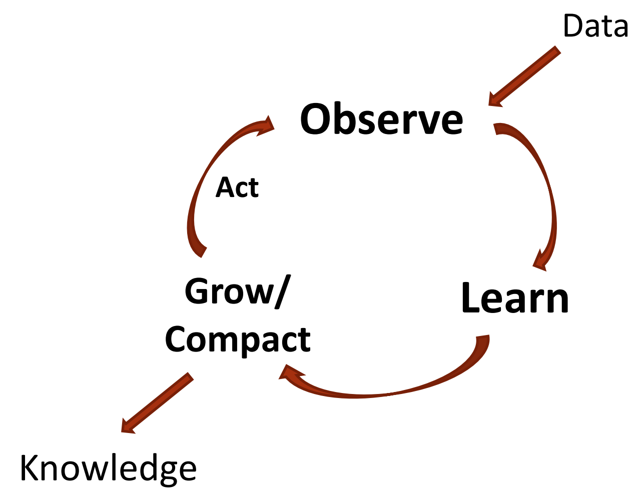
Decision Boundary (cont.)

Eventually ….



Learning Properties in Cyber-Physical Systems

- **Continuous/Dynamic/Adaptive Process**
- **Interpretation**
  - Why and when doesn't it work?
  - Knowledge Representation and Reasoning
- **Robustness**
  - Model uncertainty, overfitting, etc.
  - Transfer to real system?
- **Time and Memory Efficiency**
  - Real-time?
  - Processing/Communication bandwidth
  - Hyperparameter-tuning
  - Performance-Complexity Trade-off
  - **Progressive Learning?**



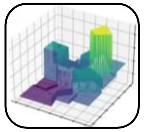
Key Take Away Points

- Formally Analyze **Learning as a Dynamic Process** of acquiring new understanding, knowledge, or skills
- Investigate **Learning with Progressively Growing Knowledge Representations** for Decision-Making Systems
- Towards a Neuroscience-inspired Universal Learning Algorithm: **Hierarchical, Memory-based, Progressive, Interpretable, Robust**
- Adaptive Space Aggregation for **Memory-Efficient Reinforcement Learning** in Robot Control
- Progressive **Graph Partitioning** and Image Segmentation

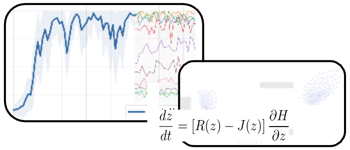
Outline

- **Learning as a Dynamical System**
- **Towards Universal Learning Architectures**
  - Multi-resolution-group invariance, local learning
- **Progressive Learning, On-Line**
  - Definition, Properties, Results
- **Applications to CPS**
  - Robust Reinforcement Learning
  - CPS Security
  - Robotics & Multi-Agent Systems
- **Future Research Directions**

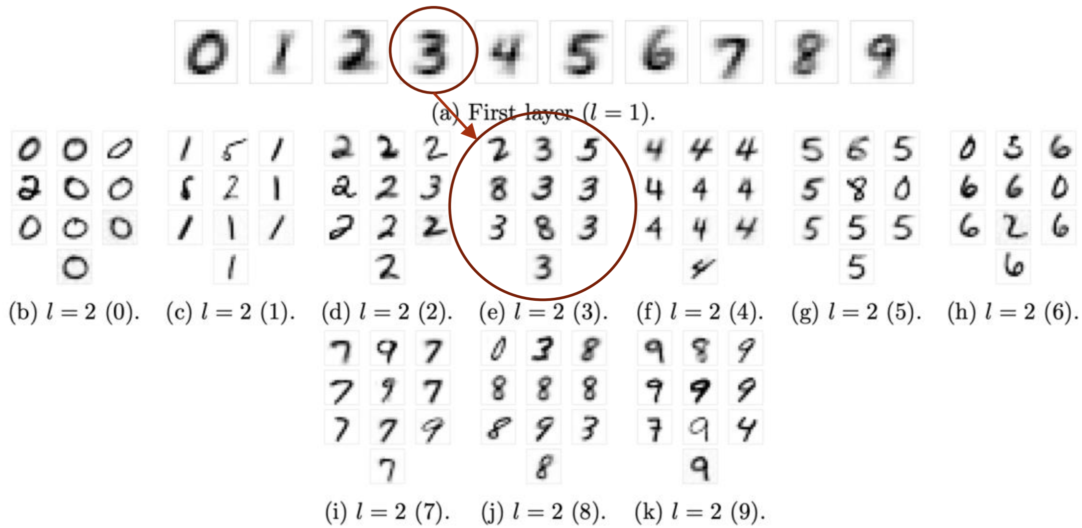
Hierarchical Learning
Stochastic Optimization
Knowledge Representation
Interpretable ML



Risk-Sensitive RL
Explainable RL
Swarm Dynamics
CPS Security
Community Detection
Influence Graphs
Human-Robot Collaboration

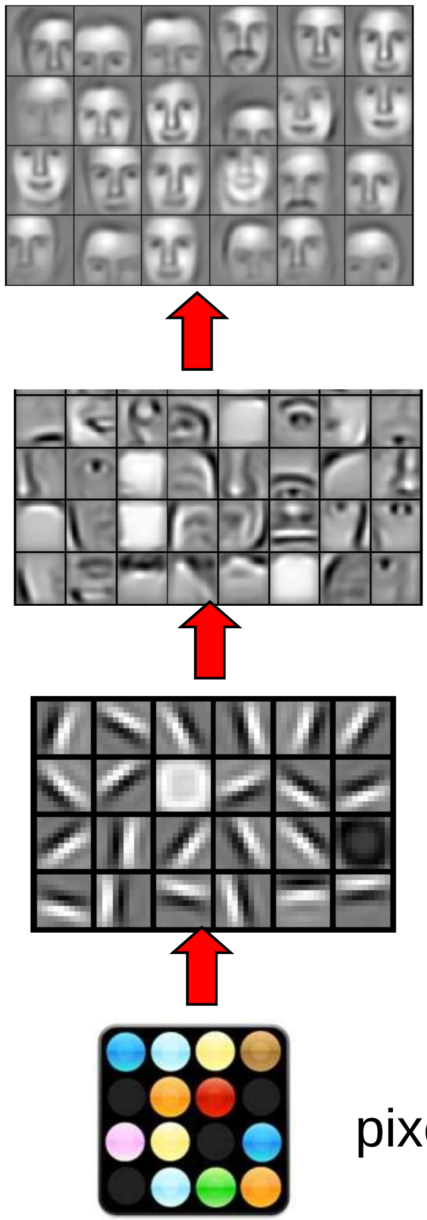


Multi-Resolution ODA (MNIST)



Representations generated by the first two layers of a multi-resolution ODA algorithm for the MNIST dataset.
Input: low-resolution images from wavelet analysis (14x14 pixels). Accuracy: 97.2% (can go up to 100% in training data).
The neurons represent different deformations of each digit. The relationship between them can lead to the identification of better features and invariances.

Feature Hierarchies

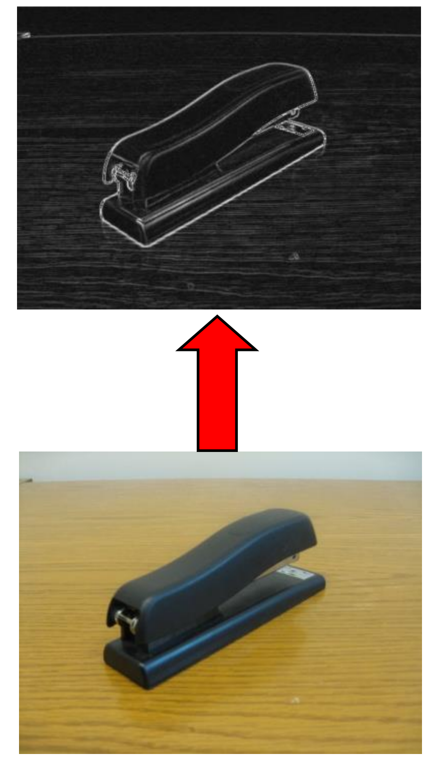


object models

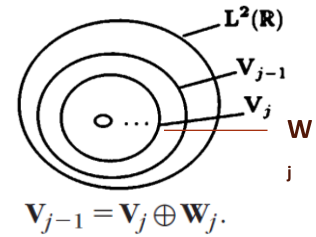
object parts
(combination
of edges)

edges

pixels



Group-Invariant Representations



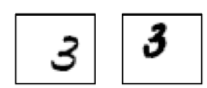
- **Wavelet Transform**
    - Multi-Resolution Analysis
    - Sparse, Stable, Translation Covariant

- **Convolution on Groups**

$$(f * g)(x) = \int_G f(y)g(y^{-1}x)d\lambda(y)$$

where for a Lie Group *G:* $g \in G \to g.f(x) := f(g^{-1}x)$

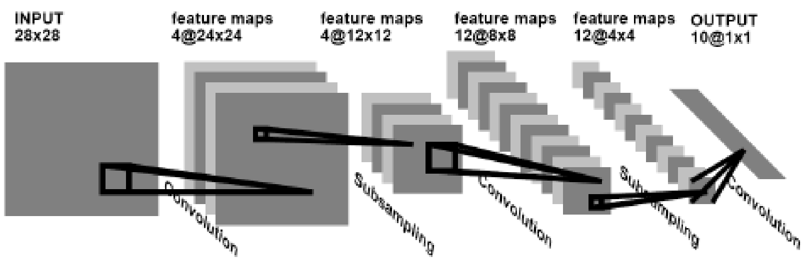
- **Locally Invariant Representations**



Repeat
- Build group-covariant representations (**wavelets**)
- Make them locally invariant (**non-linearity + averaging**)

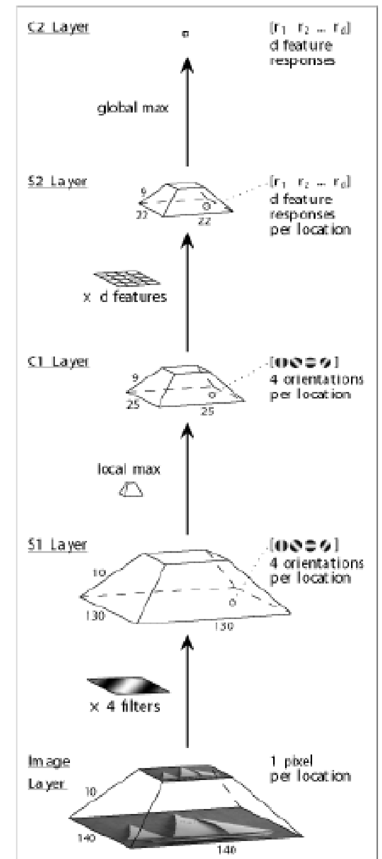
Some Current Deep Architectures

Convolutional Networks (Lecun)

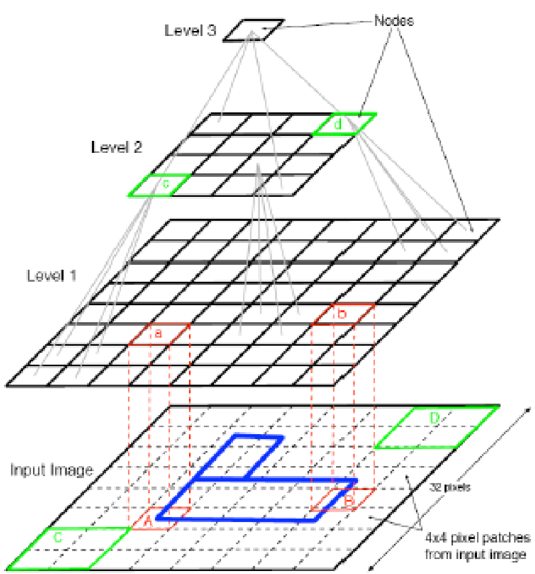


Lecun et al. (1995)

HMAX (Poggio)

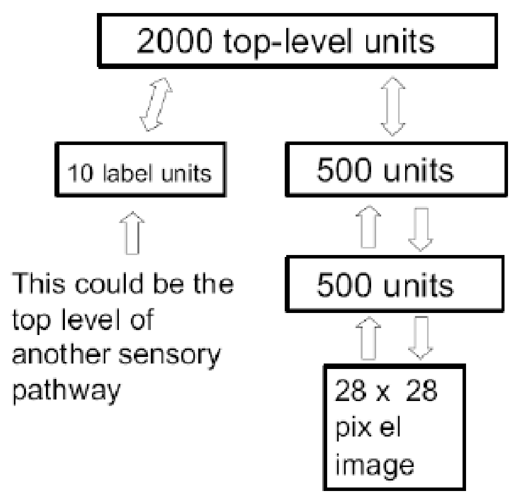


HTM (Hawkins)



Dileep George (2008)

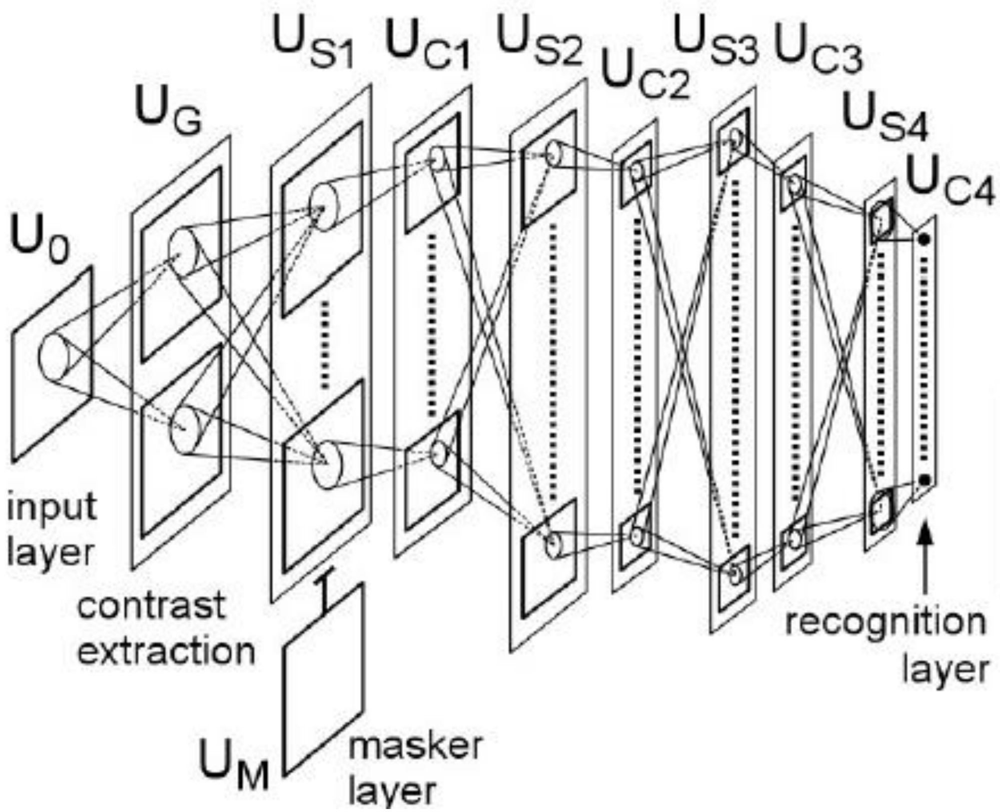
DBNs (Hinton)

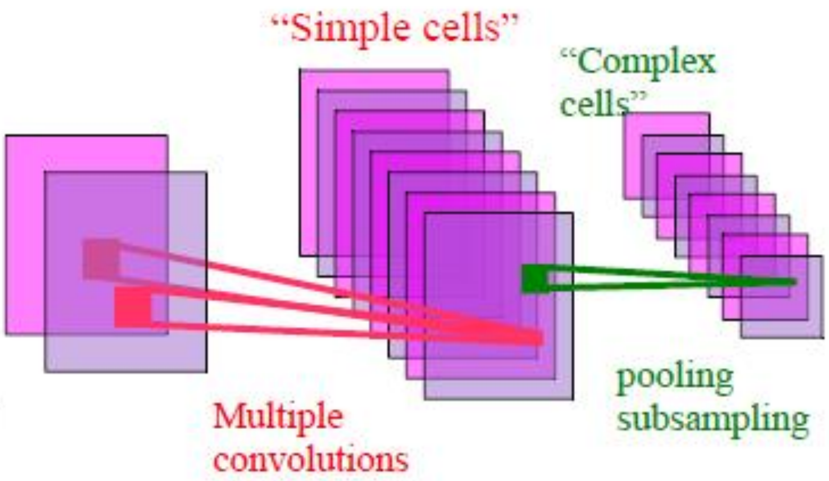


Hinton et al. (2006)

Early Hierarchical Feature Models for Vision

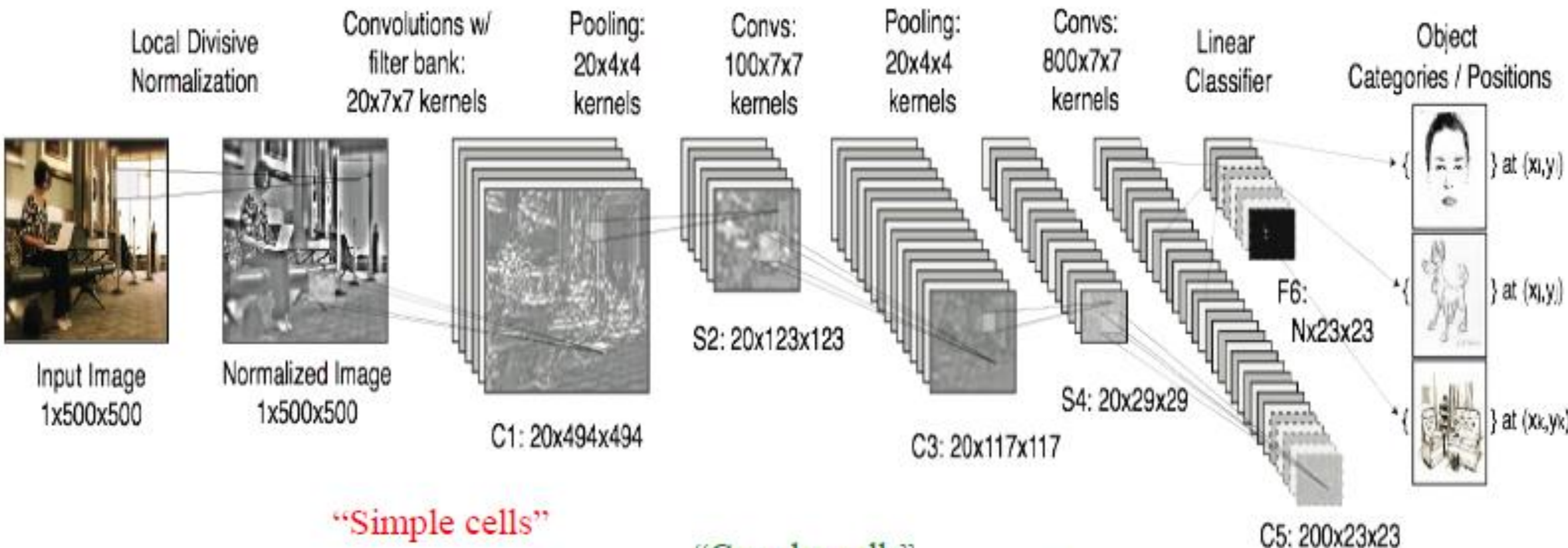
- [Hubel & Wiesel 1962]:
  - simple cells detect local features
  - complex cells "pool" the outputs of simple cells within a retinotopic neighborhood.

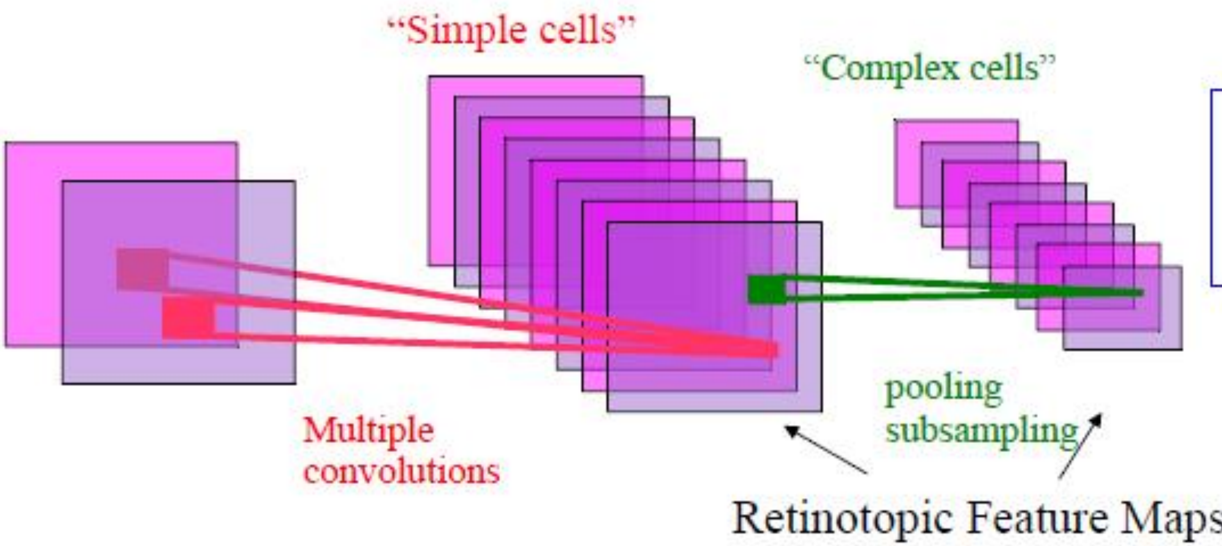




The Convolutional Net Model

Multistage Hubel-Wiesel System





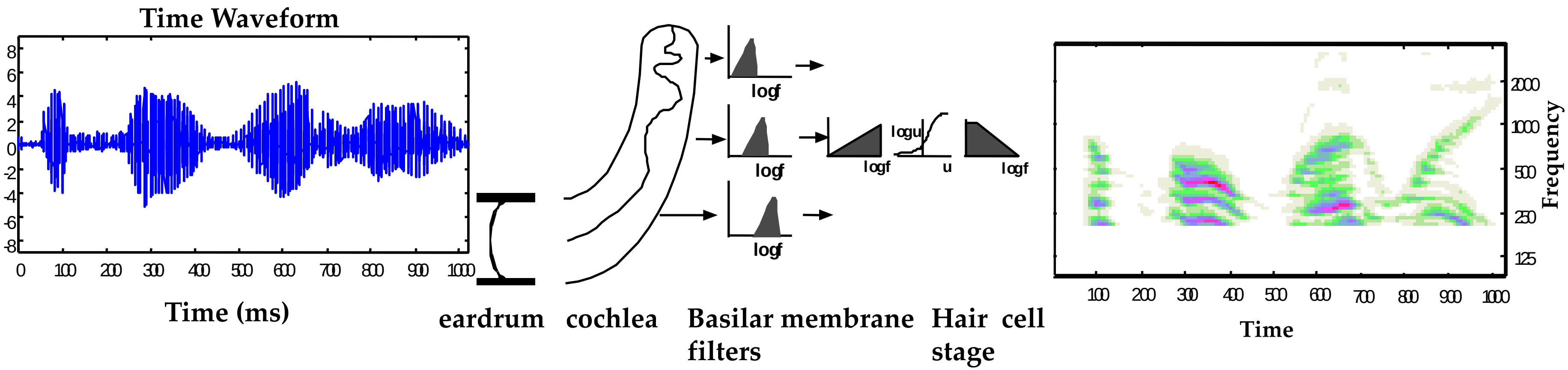
- Training is supervised
- With stochastic gradient descent

[LeCun et al. 89]

[LeCun et al. 98]

Multiresolution Preprocessor: Auditory Filtering (Shamma 2003)

Two auditory filters, motivated and designed according to acoustic physiology and acoustic cortex models, were used to compute the timbre spectrogram of one particular subframe in each frame



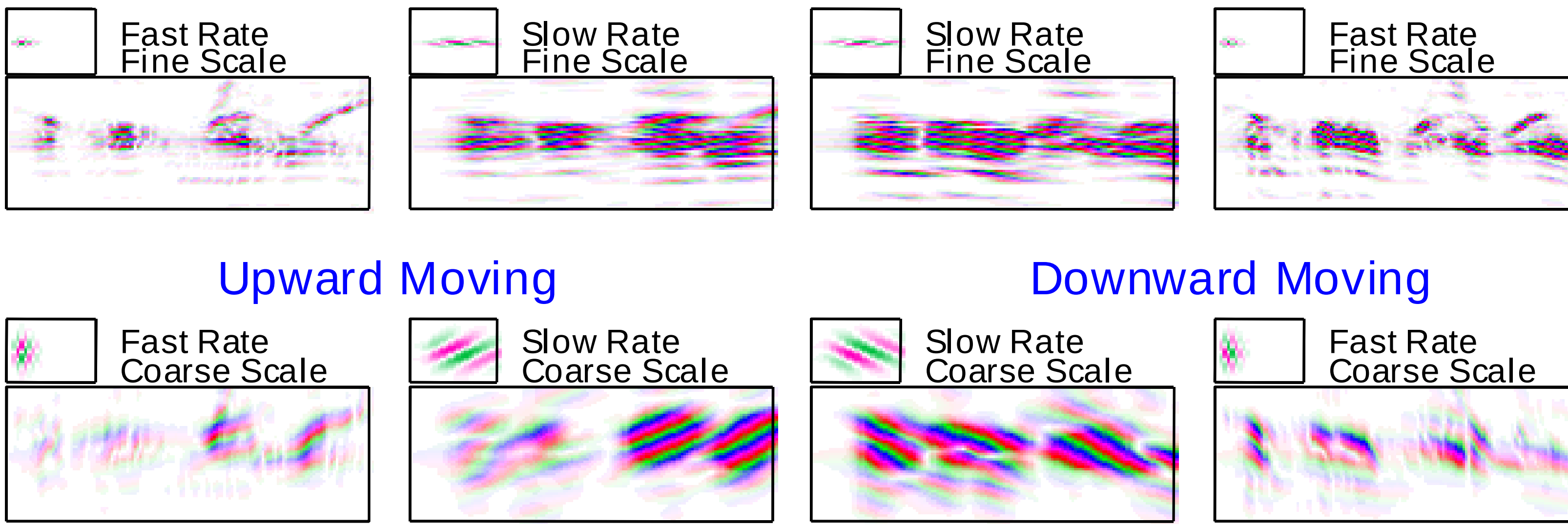
- The first filter mimics the action of the inner ear
- Computes the spectrogram of the sound sample, and performs various nonlinear operations, which models the nonlinear fluid-cilia couplings and ionic channels of conduction

(Wavelet Transform)

S. Shamma, IETE Journal of Research, Vol 49, No 2, March-April, 2003.

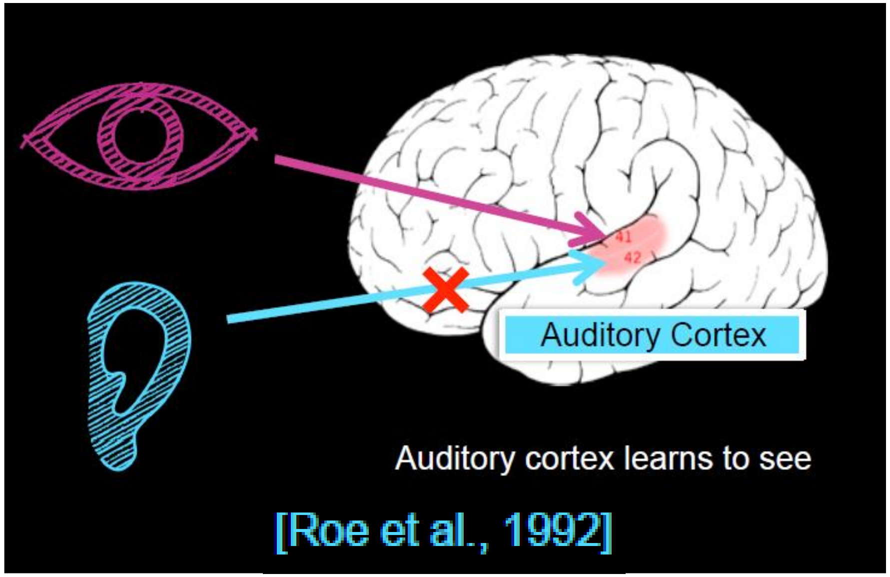
Spectro-temporal Processing: Multiresolution Preprocessor -- Auditory Filtering

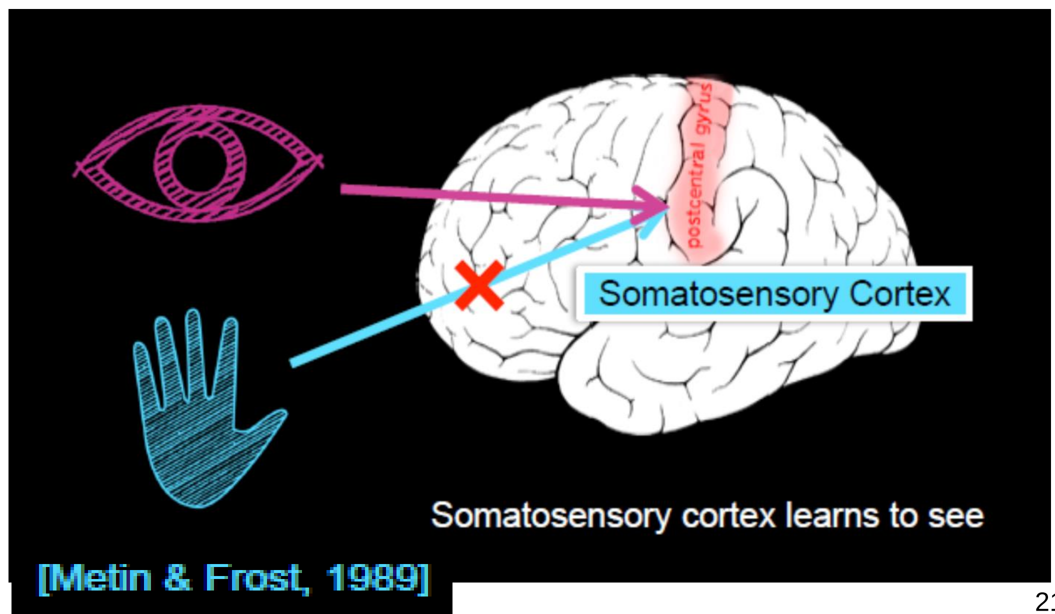
Multiresolution cortical filter outputs



- **The second filter models the multiscale processing of the signal that happens in the auditory cortex**
- **A Ripple Analysis Model, using a ripple filter bank, acts on the output of the inner ear to give multiscale spectra of the sound timbre (Wavelet Transform)**

"One Learning Algorithm" Hypothesis





Andrew Ng – Google Brain

The Man Behind the Google Brain: Andrew Ng and the Quest for the New AI

THERE'S A THEORY that human intelligence stems from a single algorithm.

The idea arises from experiments suggesting that the portion of your brain dedicated to processing sound from your ears could also handle sight for your eyes. This is possible only while your brain is in the earliest stages of development, but it implies that the brain is -- at its core -- a general-purpose machine that can be tuned to specific tasks.

About seven years ago, Stanford computer science professor Andrew Ng stumbled across this theory, and it changed the course of his career, reigniting a passion for artificial intelligence, or AI. "For the first time in my life," Ng says, "it made me feel like it might be possible to make some progress on a small part of the AI dream within our lifetime."

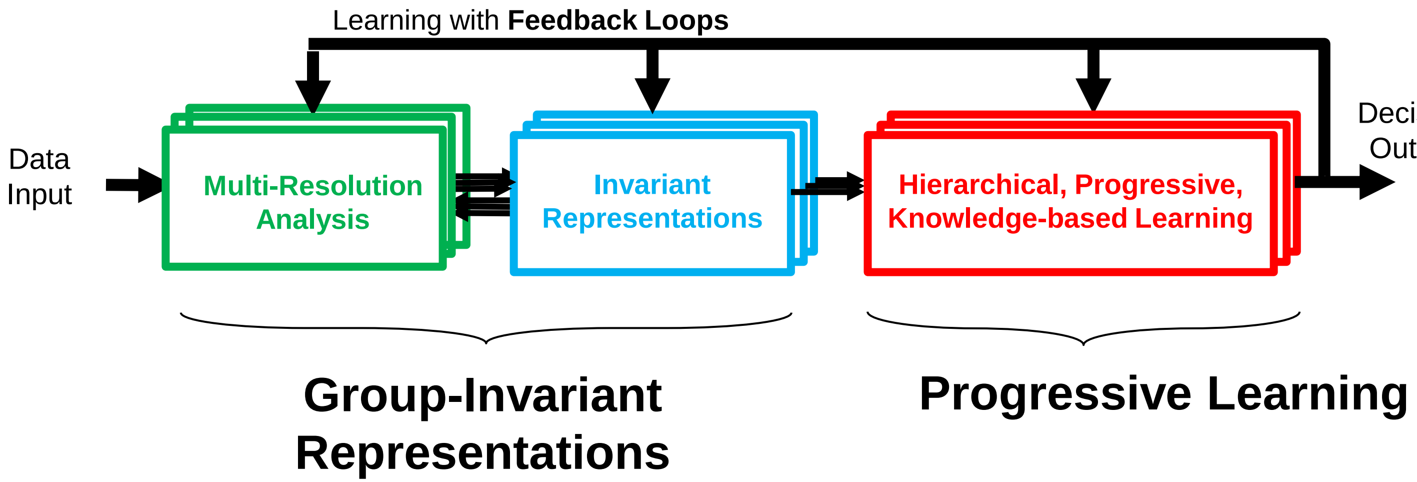
"one algorithm" hypothesis, popularized by Jeff Hawkins

Google Brain.

Towards a Universal Learning Architecture

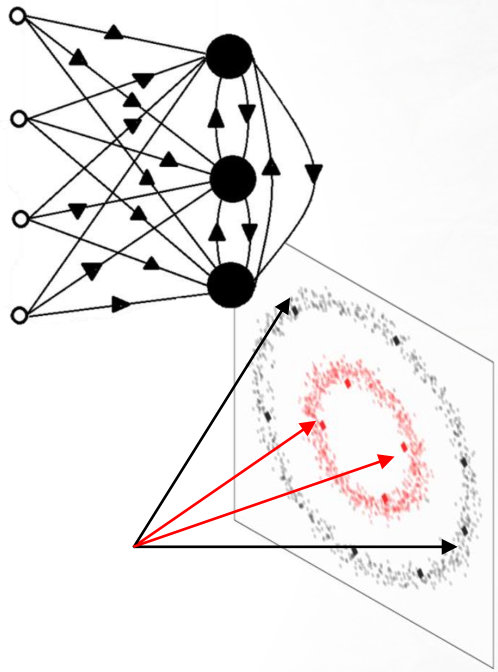
- A robust and interpretable alternative approach based on the same principles?

  a) multi-resolution analysis
  b) group-invariant representation
  c) hierarchical, knowledge-based decision-making



Towards a Universal Learning Architecture (cont.)

Dynamic Learning



I. Neurons live in the data space

- Interpretability
- Robustness w.r.t. perturbations and adversarial attacks
- Vector Quantization?

II. Progressively Growing

- Performance-Complexity Trade-off
- No over-fitting

III. Annealing Optimization

- Robustness w.r.t. initial conditions
- No poor local minima
- **Gradient-Free** Stochastic Approximation

Let's Go Back in Time

ISR INSTITUTE FOR SYSTEMS RESEARCH

Progressive Classification: Universal Algorithms and Applications

John S. Baras

Electrical Engineering Department
and Institute for System Research
University of Maryland College Park

Visiting EECS and LIDS, MIT

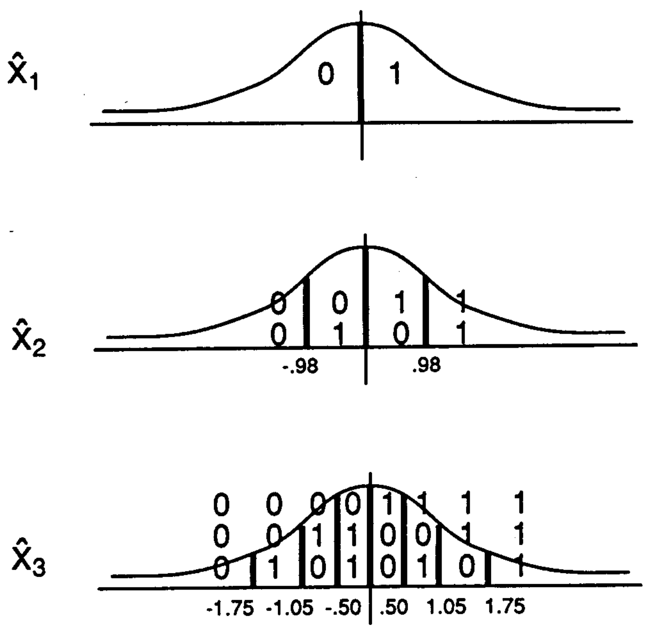
LIDS Colloquium
March 10, 1998

Progressive Classification

- Small amounts of information in the form of a coarse approximation of the signal, are used first to provide partial classification
- Progressively finer details are added until satisfactory performance is obtained
- Approach results in a scheme where:
  - Small amounts of computation are used initially (at coarse level)
  - Additional computations (more detailed) are performed as needed
- Approach leads to:
  - Faster classification algorithms (faster search)
  - Algorithms that preserve high fidelity in the search (the challenge)
  - Easily parallelizable algorithms

Successive Refinement of Information
(Equitz and Cover (1991))

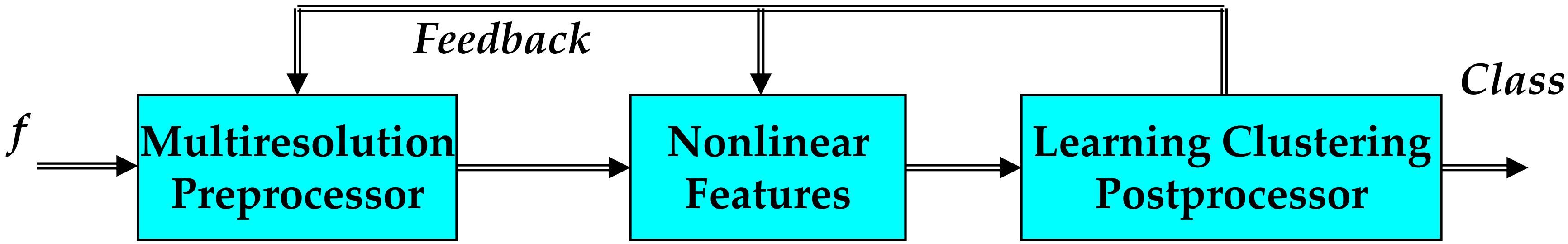
- **Successive refinement from a coarse description $\hat{X}_1$ with distortion D_1 to a finer description $\hat{X}_2$ with distortion D_2 can be achieved iff the conditional distributions $P(\hat{x}_1 | x)$ and $P(\hat{x}_2 | x)$, which achieve $I(X;\hat{X}_i) = R(D_i), \quad i = 1, 2,$ are Markov compatible: we can write $\hat{X}_1 \to \hat{X}_2 \to X$ as a Markov chain**



- **Conditions rarely satisfied; examples where they are satisfied:**
- **Gaussian signals with squared-error distortion**
- **Finite alphabet signals with Hamming distortion**
- **Laplacian signals with absolute-error distortion**

Minimize Average Squared Error from using a few bits to describe X~N(0,1)

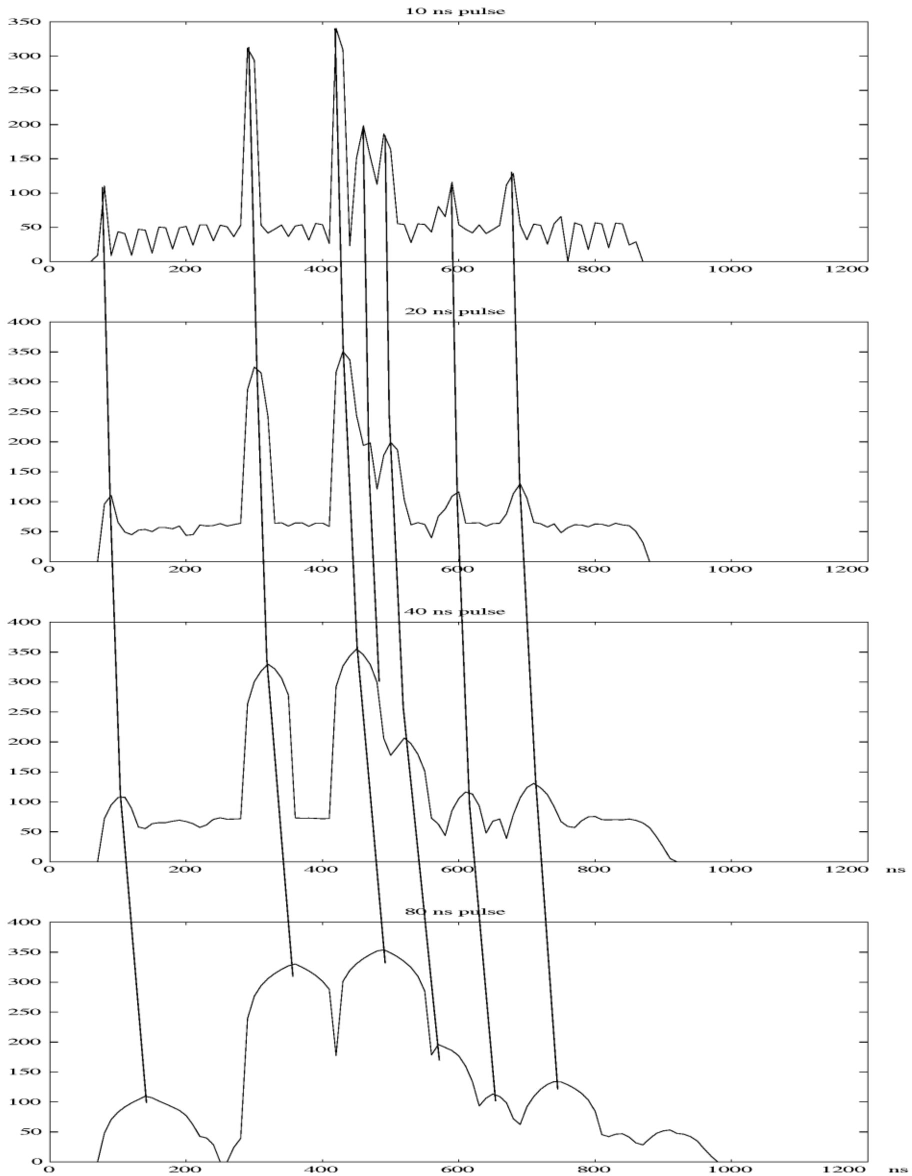
Multiresolution and Learning Clustering

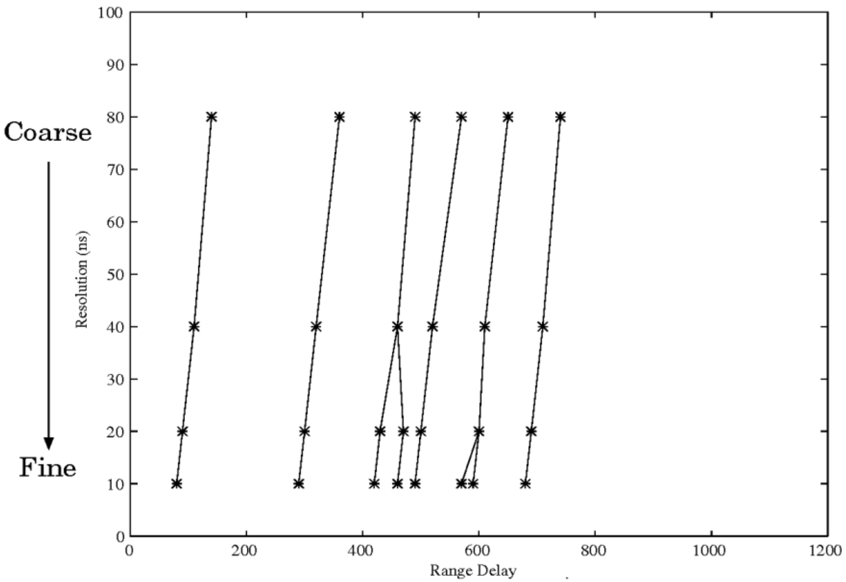


- Address both the hierarchical organization of signal databases and progressive classification:
  - Combine a multiresolution preprocessor with a learning clustering postprocessor
  - Feedback is also an option
- Resulting algorithms proved to have some "universal" qualities
- Found analogs of such algorithms in animals and humans:
  - Hearing and sound classification
  - Vision and identification of objects by humans
- Most promising mathematical formulation of the problem:

  combined compression and classification for general signals

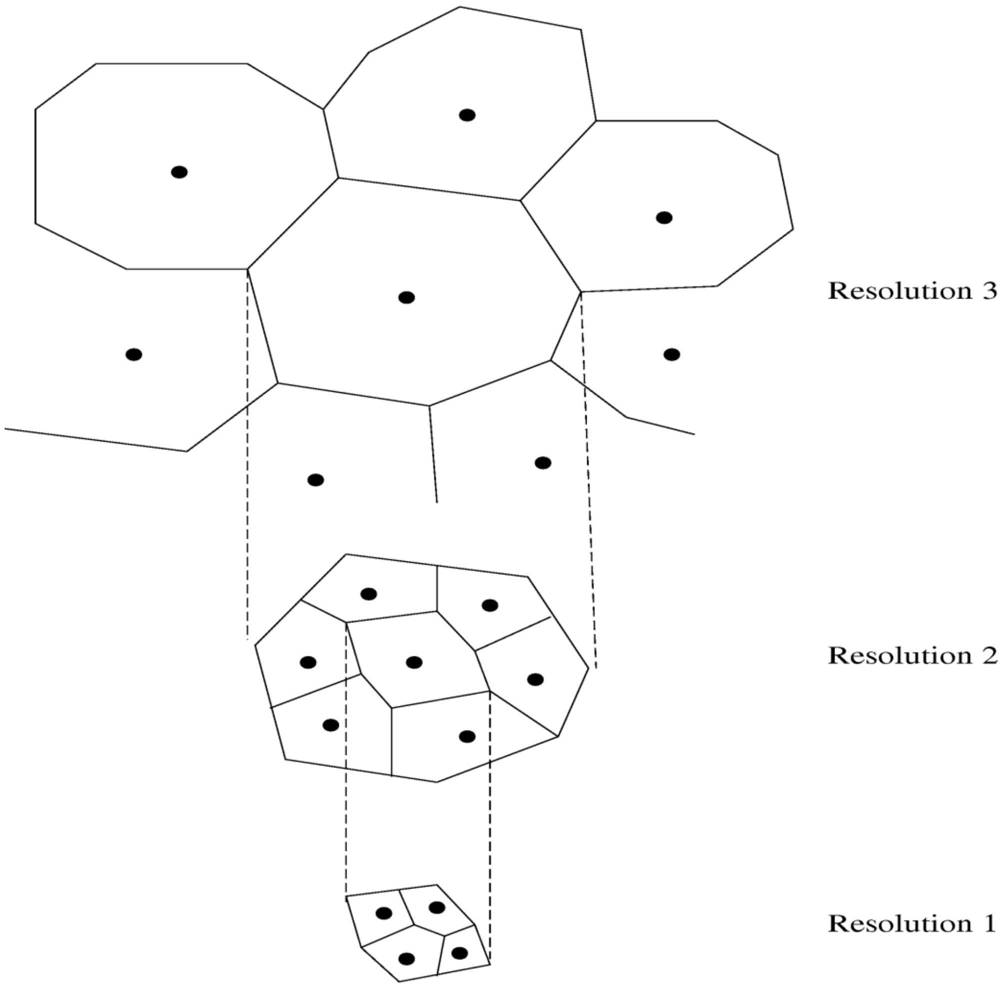
Scale-Space Diagrams of Radar Returns





- Uniform Localization
- A different "fingerprint" of the ship

Wavelet Tree-Structured Vector Quantization



First perform a multiresolution wavelet representation of the signals

Consider each signal f at different resolutions

$S^0 f,\ S^1 f,\ \ldots,\ S^{J^*} f$

Proceed by partitioning the signal space at various resolutions in progressively finer cells

Layer in tree $l = J^* - m$,

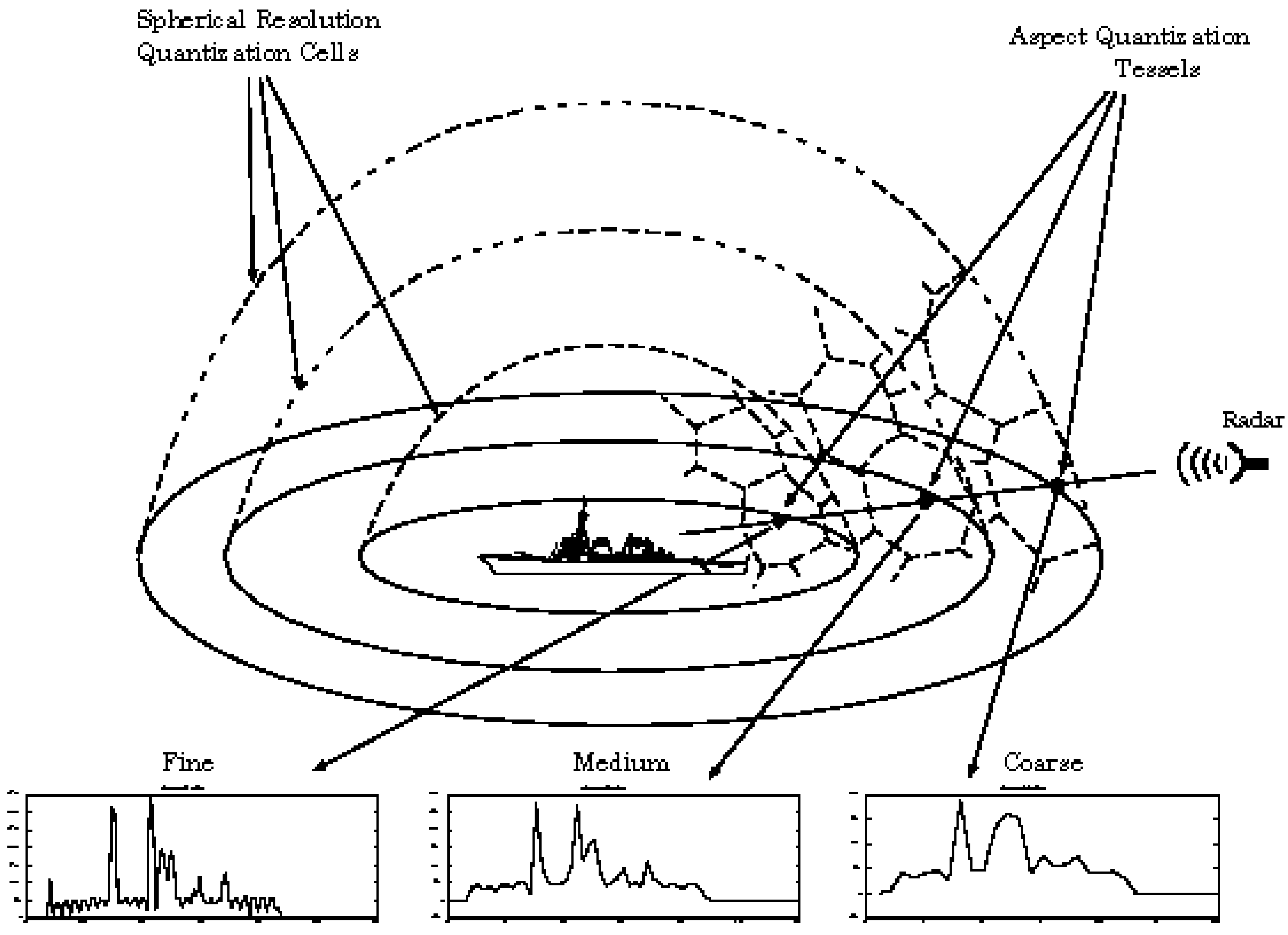
m the scale

(top layer 0: coarsest)

Cell labels: (layer, index)

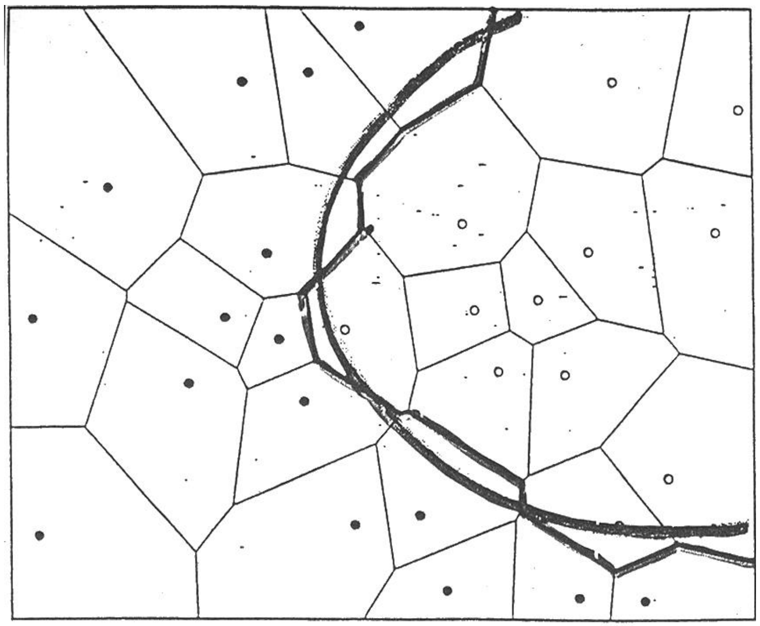
or (scale, index)

Multiresolution Aspect Graph: Radar Data



Learning Vector Quantization

- Data driven; uses past data directly in the classification scheme
- Does not assume any models for underlying data



- Estimates the decision regions directly
- Training phase and classification phase
- Training phase:

  Z = training data = $\{(y_n, d_{y_n})\}_{n=1}^{N}$

  Voronoi vectors = $\Theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$

  decisions = $\{d_{\theta_1}, d_{\theta_2}, \cdots, d_{\theta_k}\}$

- Pick $z_j=(y_j, d_{y_j})$ from Z and find ρ - closest vector θ_c
- Modify θ_c as follows $\theta_c(n+1) = \theta_c(n) - \alpha_n \nabla_\theta \rho(\theta_c(n), y_j)$ if $d_{y_j} = d_{\theta_c}$

  $\theta_c(n+1) = \theta_c(n) + \alpha_n \nabla_\theta \rho(\theta_c(n), y_j)$ if $d_{y_j} \neq d_{\theta_c}$
- Continue until convergence

LVQ

- **Classification phase: for new observation *x* declare**

$$d_x = d_{\theta_j} \text{ if } x \in V_{\theta_j}$$

- **LVQ adjustment has the general form**

$$\theta_i(n+1) = \theta_i(n) + \alpha_n \gamma(d_{y_n}, d_{\theta_i}(n), x_n, \Theta_n) \nabla_\theta \rho(\theta_i(n), y_n)$$

$$\gamma(d_{y_n}, d_{\theta_i}(n), y_n, \Theta_n) = -1_{\{y_n \in V_{\theta_i}\}}(1_{\{d_{y_n} = d_{\theta_i}\}} - 1_{\{d_{y_n} \neq d_{\theta_i}\}})$$

- $\Theta_{n+1} = \Theta_n + \alpha_n H(\Theta_n, z_n)$ **; stochastic approximation**

$$z_n = (y_n, d_{y_n})$$

- **For appropriate conditions on α_n, H, z_n, Θ_n approaches the solution of the ODE**

$$\frac{d}{dt}\overline{\Theta}(t) = h(\overline{\Theta}(t))$$

for appropriate *h*(Θ)

Stochastic Approximation

Theorem. *Almost surely, the sequence:*

$$x_{n+1} = x_n + \alpha(n)\left[h(x_n) + M_{n+1}\right], \; n \geq 0 \tag{1}$$

converges to a (possibly sample path dependent) compact, connected, internally chain transitive, invariant set of the o.d.e:

$$\dot{x}(t) = h\left(x(t)\right), \; t \geq 0, \tag{2}$$

provided that:

(A1) $h : \mathbb{R}^d \to \mathbb{R}^d$ *is Lipschitz.*

(A2) $\sum_n \alpha(n) = \infty$, *and* $\sum_n \alpha^2(n) < \infty$

(A3) $\{M_n\}$ *is a martingale difference sequence*

(A4) $\{x_n\}$ *remain bounded a.s.*

Examples:

$$h(x) = \begin{cases} -\nabla J(x), & \text{SGD} \\ F(x) - x, & \text{Fixed-Point Iter.} \end{cases}$$

*Borkar, Stochastic approximation: a dynamical systems viewpoint, Springer, 2009

Combined Compression and Classification

- Given a encoder-decoder pair γ, δ we associate the average distortion

$$D(\gamma,\delta) = E[\rho(x,\delta(\gamma(x)))]$$

- Associate the rate R(γ,δ) to a encoder-decoder pair γ,δ
- Given a classification rule *d*, the classification performance of the overall scheme can be measured by the Bayes risk

$$J_B(\gamma,d) = \sum_{i=1}^{L}\sum_{j=1}^{L} P(d(\gamma(x)) = H_j \mid x \in H_i) P(H_i) C_{ij}$$

- where C_{ij} is the relative cost assigned to the decision that $d(\gamma(x)) = H_j$, while the vector x comes from class H_i (typically $C_{ij} = 0$)
- Encoder δ does not affect the Bayes risk J_B
- Incorporate Bayes risk into the average distortion measure minimized by the design algorithm
- Resulting algorithm has complexity equivalent to that of an ordinary VQ algorithm

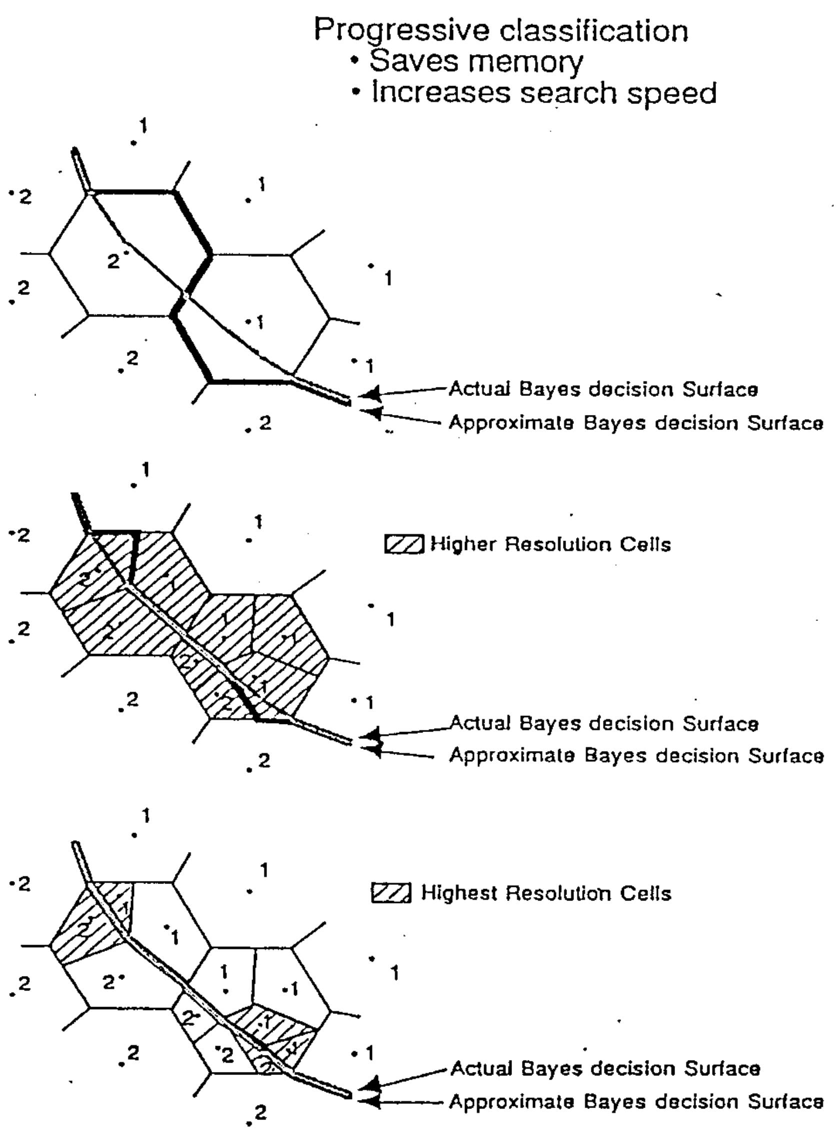
Analytical Framework

- Overall approach is non-parametric:
  probability distributions for the data are not needed
- Approach can be interpreted as using the training set to learn the empirical distributions of the vectors and use them as if they were true (like in LVQ)
- Combine the three criteria in one for some choice of the weights λ_R and λ_B

$$J_\lambda(\gamma,\delta,d)=D(\gamma,\delta)+\lambda_R\,R(\gamma,\delta)+\lambda_B\,J_B(\gamma,d),$$

- Three step iterative optimization:
  - Step 1 Choose $d^{(t+1)}$ to minimize J_λ $(\gamma^{(t)}, \delta^{(t)}, d^{(t+1)})$
  - Step 2 Choose $\delta^{(t+1)}$ to minimize J_λ $(\gamma^{(t)}, \delta^{(t+1)}, d^{(t+1)})$
  - Step 3 Choose $\gamma^{(t+1)}$ to minimize J_λ $(\gamma^{(t+1)}, \delta^{(t+1)}, d^{(t+1)})$
  - The iterations continue until the desired stoping level for J_λ is met

Extension to LTSVQ and Interpretation



Extension of the LVQ approach to Learning TSVQ

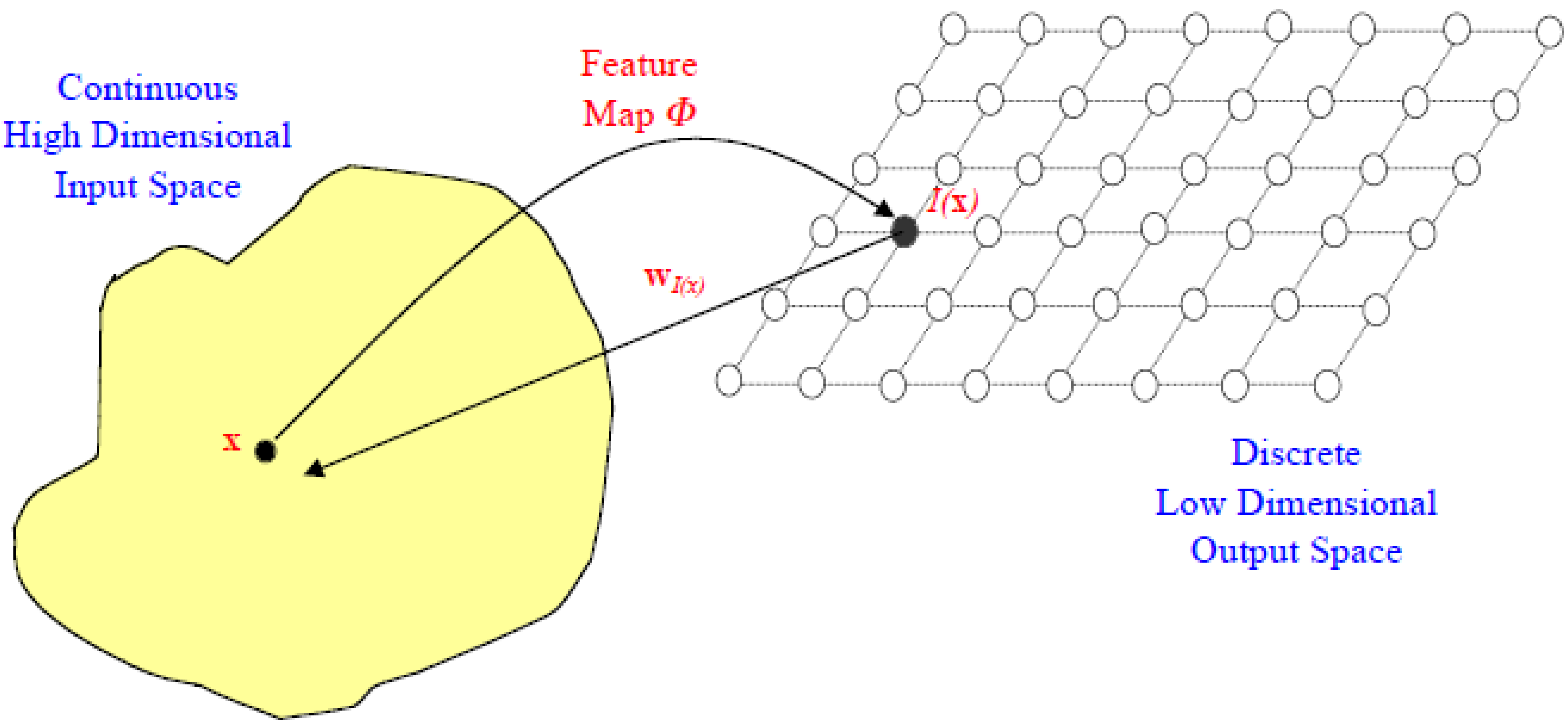
This step is needed for the full analysis of WTSVQ and its application in progressive classification within the framework of combined compression and classification

LTSVQ approximates directly the optimal Bayes surface with successive approximations and variable (along the surface) resolution

- Split cells where approximation is not very good using finer resolution information
- Akin to a multigrid numerical computation of the Bayes surface

SOM -- Kohonen Mapping

We have points **x** in the input space mapping to points $I(\mathbf{x})$ in the output space:

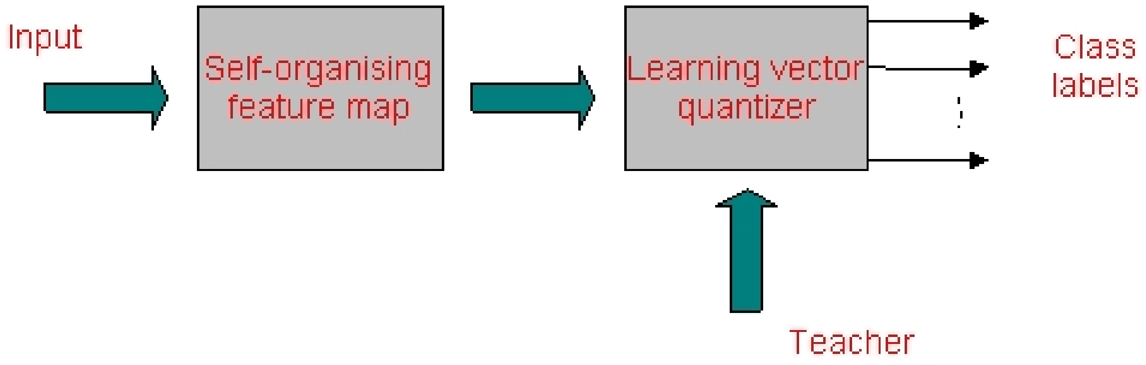


Each point I in the output space will map to a corresponding point $\mathbf{w}(I)$ in the input space.

SOM can find the manifold on “manifold-localized” data (e.g. data on a sphere, or circle)

SOM followed by LVQ

Computation of the feature map can be viewed as the first of two stages for adaptively solving a pattern classification problem as shown below. The second stage is provided by the learning vector quantization, which provides a method for fine tuning of a feature map. This is useful and typical for DNN



"Oldies but Goodies"

References https://johnbaras.com/

J. S. Baras and A. LaVigna, "Convergence of Kohonen's Learning Vector Quantization" *Proceedings of the 1990 IJCNN Conference on Neural Networks*, pp. 476-482, San Diego, CA, June 17-21, 1990

J.S. Baras and A. LaVigna, "Convergence of a Neural Network Classifier", *Proc. 29th IEEE CDC*, pp. 1735-40, 1990.

A. LaVigna, *Nonparametric Classification Using Learning Vector Quantization*, Ph.D. Thesis, EE, UMD, Fall 1989.

J.S. Baras and S.I. Wolk, "Hierarchical Wavelet Representations of Ship Radar Returns" NRL Report NRL/FR/5755-93-9593.

J.S. Baras and S.I. Wolk, "Efficient Organization of Large Ship Radar Databases Using Wavelets and Structured Vector Quanitzation" *Proc. 27th Annual Asilomar Conference*, Vol. 1, pp. 491-498, 1993, Pacific Grove, CA.

J.S. Baras and S.I. Wolk, "Model Based Automatic Target Recognition from High Range Resolution Radar Returns", *Proc. SPIE Intern. Symp. on Intelligent Information Systems*, Vol. 2234, pp. 57-66, Orlando, FL, April 5-8, 1994

J.S. Baras and S.I. Wolk, "Wavelet Based Progressive Classification of High Range Resolution Radar Returns", *Proc. SPIE Intern. Symp. on Intelligent Information Systems*, Vol. 2242, pp. 967-977, Orlando, FL, April 5-8, 1994

"Oldies but Goodies"
References https://johnbaras.com/

J.S. Baras and S.I. Wolk, "Wavelet Based Progressive Classification with Learning: Applications to Radar Signals" *Proc. SPIE 1995 Intern. Symp. Aerospace/ Defense*, Vol. 2491, Part 1 of 2, pp. 339-350, Orlando, Florida, 1995

J.S. Baras and S.I. Wolk, "Wavelet-Based Hierarchical Organization of Image Data: Applications to ISAR and Face Recognition", *Proc. Conf. Information Sciences and Systems*, Johns Hopkins Univ., Baltimore, MD, 1997

J.S. Baras and S.I. Wolk, "Wavelet-Based Hierarchical Organization of Large Image Databases: ISAR and Face Recognition" *Proc. SPIE 12th Intern. Symp. Aerospace, Defense*, Orlando, Florida, April 13-17, 1998.

J.S. Baras and S. Dey, "Combined Compression and Classification with Learning Vector Quantization", *IEEE Transactions on Information Theory*, Vol. 45, No. 6, pp. 1911-1920, September 1999.

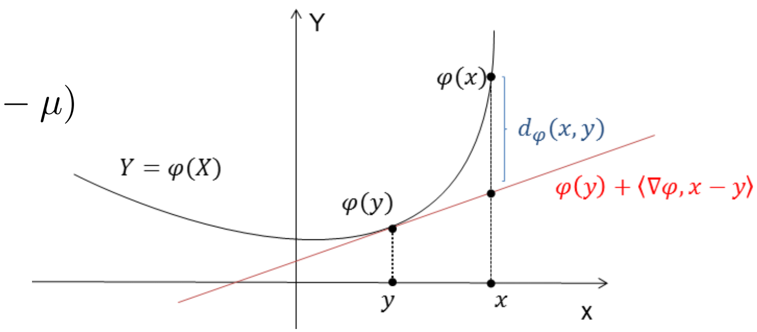
J.S. Baras and V.S. Borkar, "A Learning Algorithm for Markov Decision Processes with Adaptive State Aggregation," *Proceedings 39th IEEE CDC*, Sydney, 2000.

J.S. Baras and A.S. Baras, "A Novel Nonparametric Discrimination Measure for Analysis of Gene Expression Data," *SIAM Conference on the Life Sciences*, Portland, July 2004.

Back to Recent / Current Time

Dissimilarity Measures: Bregman Divergences

- $d_\phi(x,\mu) = \phi(x) - \phi(\mu) - \dfrac{\partial \phi}{\partial \mu}(\mu)(x-\mu)$
  - **Euclidean distance, KL divergence, …**



- **Theorem.** *Let $X:\Omega\to S$ be a random variable defined in the probability space $(\Omega,\mathcal{F},\mathbb{P})$ such that $\mathbb{E}[X]\in ri(S)$, and let a distortion measure $d: S\times ri(S)\to[0,\infty)$, where $ri(S)$ denotes the relative interior of S. Then*

$$\mu := \mathbb{E}[X] \in \arg\min_{s\in ri(S)} \mathbb{E}[d(X,s)]$$

is the unique minimizer of $\mathbb{E}[d(X,s)]$ in $ri(S)$, if and only if d is a Bregman divergence for any function ϕ that satisfies the definition.

Recent Proofs: Stochastic VQ

Problem 1. *Let $X : \Omega \to S$ be a random variable defined in the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and $d_\phi : S \times ri(S) \to [0, \infty)$ be a Bregman divergence with properly defined function ϕ. Let $V \triangleq \{S_h\}_{h=1}^k$ be a Voronoi partition of S with respect to d_ϕ and $M \triangleq \{\mu_h\}_{h=1}^k$, such that $\mu_h \in ri(S_h)$, $h \in K$, $K \triangleq \{1, \dots, k\}$, and define the quantizer $Q : S \to S$ such that $Q(X) = \sum_{h=1}^k \mu_h \mathbb{1}_{[X \in S_h]}$.*

Then the problem is formulated as

$$\min_{M,V} J(Q) \triangleq \mathbb{E}_X \left[d_\phi \left(X, Q(X) \right) \right]$$

$$\Leftrightarrow \min_{\{\mu_h\}_{h=1}^k} J(Q) \triangleq \sum_{h=1}^k \mathbb{E}_X \left[d_\phi \left(X, \mu_h \right) \mathbb{1}_{[X \in S_h]} \right],$$

$$\begin{aligned} \mu_h^{t+1} &= \mu_h^t + \alpha(t) \left(-\mathbb{1}_{[X_{t+1} \in S_h^{t+1}]} \right) \nabla_{\mu_h} d_\phi \left(X_{t+1}, \mu_h^t \right) \\ S_h^{t+1} &= \left\{ X \in S : h = \arg\min_{\tau = 1, \dots, k} d_\phi \left(X, \mu_\tau^t \right) \right\}, \ h \in K \end{aligned} \tag{3}$$

$$\begin{aligned} \theta_h(\mu) &= \lim_{t \to \infty} \mathbb{E} \left[\Theta_h(\mu^t, X_{t+1}) | \mu_h^t \right] \\ &= -\mathbb{E}_X \left[\mathbb{1}_{[X \in S_h]} \nabla_{\mu_h} d_\phi \left(X, \mu_h \right) \right] \end{aligned}$$

$$\dot{\mu}(t) = \theta\left(\mu(t)\right), \ t \geq 0,$$

$$\theta(\mu) = -\nabla_\mu J(\mu)$$

Theorem 3. *The sequence $\{\mu^t\}$ generated by the stochastic vector quantization algorithm (3) converges almost surely to a local solution μ^* of Problem 1, as long as the function ϕ satisfies Assumption 1, the stepsizes satisfy $\sum_t \alpha(t) = \infty$, $\sum_t \alpha^2(t) < \infty$, and μ^t visits a compact subset of the domain of attraction D^* of μ^* infinitely often, $\mu^0 \in D^*$.*

C. N. Mavridis and J. S. Baras, Convergence of Stochastic Vector Quantization and Learning Vector Quant. with Bregman Divergences, 2019

Recent Proofs: LVQ

Problem 2. *Let* $\{X,c\} \in S \times \{0,1\}$ *defined in a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$, $X : \Omega \to S$ *be a random variable, and* $c : S \to \{0,1\}$ *its associated decision variable, such that c represents the actual class of X. Let* $V \triangleq \{S_h\}_{h=1}^k$ *be a Voronoi partition of S with respect to* d_ϕ *and* $M \triangleq \{\mu_h\}_{h=1}^k$, $\mu_h \in ri(S_h)$, *and define* $C_\mu \triangleq \{c_{\mu_h}\}_{h=1}^k, c_{\mu_h} \in \{0,1\}, h \in K,\ K = \{1,\dots,k\}$, *such that* c_{μ_h} *represents the class of* μ_h *for all* $h \in K$. *Define the quantizer* $Q : S \to \{0,1\}$ *such that* $Q(X) = \sum_{h=1}^k c_{\mu_h} \mathbb{1}_{[X \in S_h]}$.

The minimum-error classification problem is then formulated as

$$\begin{aligned}\min_{\{\mu_h\}_{h=1}^k} \ J_B(Q) &\triangleq \pi_1 \sum_{H_0} \mathbb{P}_1 [X \in S_h] + \pi_0 \sum_{H_1} \mathbb{P}_0 [X \in S_h] \\ &= \pi_i + \sum_{H_i} \left(\pi_j \mathbb{P}_j [X \in S_h] - \pi_i \mathbb{P}_i [X \in S_h]\right)\end{aligned}$$

where $\pi_i = \mathbb{P}[c=i], \mathbb{P}_i\{\cdot\} = \mathbb{P}\{\cdot | c=i\}$, *and* H_i *is defined as* $H_i = \{h \in \{1,\dots,k\} : c_{\mu_h} = i\}$, $i,j \in \{0,1\}, i \neq j$.

$$\begin{aligned}\mu_h^{t+1} &= \mu_h^t + \alpha(t)\Theta_h(\mu^t, C_\mu^t, X_{t+1}, c_{t+1}) \nabla_{\mu_h} d_\phi \left(X_{t+1}, \mu_h^t\right) \\ S_h^{t+1} &= \left\{ X \in S : h = \underset{\tau=1,\dots,k}{\arg\min}\ d_\phi(X, \mu_\tau^t) \right\},\ h = 1,\dots,k\end{aligned} \qquad (13)$$

$$\Theta_h(\mu, C_\mu, X, c) = \left(-\mathbb{1}_{[X \in S_h]}\right)\left(\mathbb{1}_{[c=c_{\mu_h}]} - \mathbb{1}_{[c \neq c_{\mu_h}]}\right) \nabla_{\mu_h} d_\phi\left(X, \mu_h\right)$$

C. N. Mavridis and J. S. Baras, Convergence of Stochastic Vector Quantization and Learning Vector Quant. with Bregman Divergences, 2019

Recent Proofs: LVQ

$$\dot{\mu}(t) = \theta(\mu(t)),\ t \geq 0,$$

$$\theta(\mu) = -\nabla_{\mu} J_L(\mu)$$

and $J_L(\mu) \triangleq \sum_{h=1}^{k} J_h(\mu)$, where it is easy to show that

$$\begin{aligned} J_L &= \sum_{h=1}^{k} \delta_{\mu_h} \left(\pi_0 \mathbb{E}_0 \left[\mathbb{1}_{[X\in S_h]} d_\phi(X,\mu_h)\right] - \pi_1 \mathbb{E}_1 \left[\mathbb{1}_{[X\in S_h]} d_\phi(X,\mu_h)\right]\right) \\ &= J(Q) - 2J_{d_\phi}(Q) \end{aligned}$$

with $J(Q) = \sum_{h=1}^{k} \mathbb{E}\left[d_\phi(X,\mu_h) \mathbb{1}_{[X\in S_h]}\right]$ being the quantization error, and

$$J_{d_\phi}(Q) = \pi_1 \sum_{H_0} \mathbb{E}_1 \left[d_\phi(X,\mu_h) \mathbb{1}_{[X\in S_h]}\right] + \pi_0 \sum_{H_1} \mathbb{E}_0 \left[d_\phi(X,\mu_h) \mathbb{1}_{[X\in S_h]}\right]$$

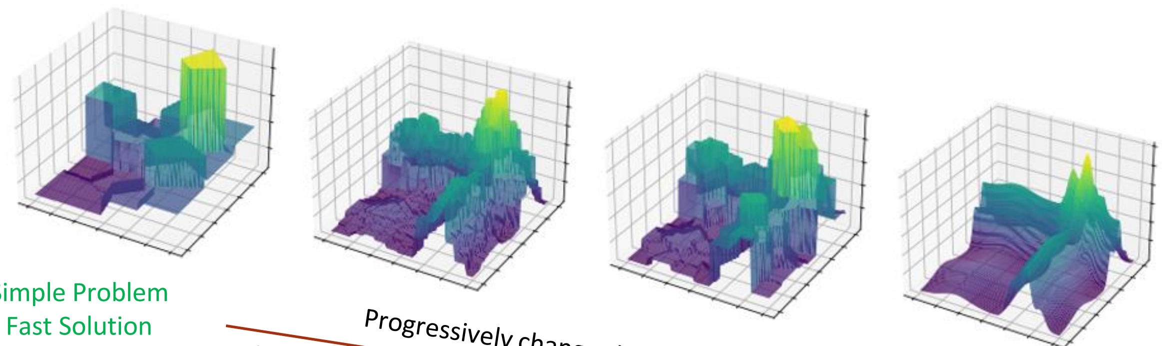
being the minimum risk error associated with the risk function

Theorem 4. *The sequence $\{\mu^t\}$ generated by the learning vector quantization algorithm (13) converges almost surely to a solution μ^* of Problem 2, as $k = k_t \to \infty$, provided that* $\lim_{t\to\infty} k_t^2 \frac{\log t}{t} \to 0$, $\sum_t \alpha(t) = \infty$, $\sum_t \alpha^2(t) < \infty$, *$\mu^t$ visits a compact subset of the domain of attraction D^* of μ^* infinitely often, $\mu^0 \in D^*$,* $\sup_t \|\mu^t\| < \infty$ *a.s., and the function ϕ satisfies Assumption 1.*

Progressive Learning for Cyber-Physical Systems

- **Goal: Hierarchically Approximate Optimal Solutions**

- optimal control
- motion planning
- function approximation
- reinforcement learning
- game policies
- clustering/classification



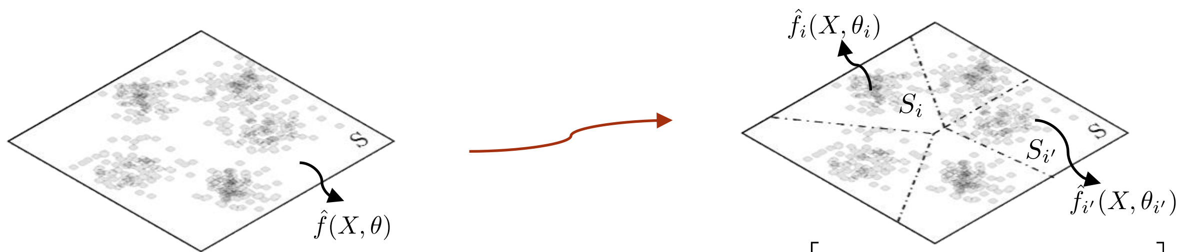
Simple Problem
Fast Solution
Coarse Approximation

Progressively change the problem
Move along the Performance-Complexity Pareto Curve

Complex Problem
Slow to Solve
Detailed Approximation

Progressive Learning for Cyber-Physical Systems

- **Divide and Conquer**
  - Partition the space and use local models

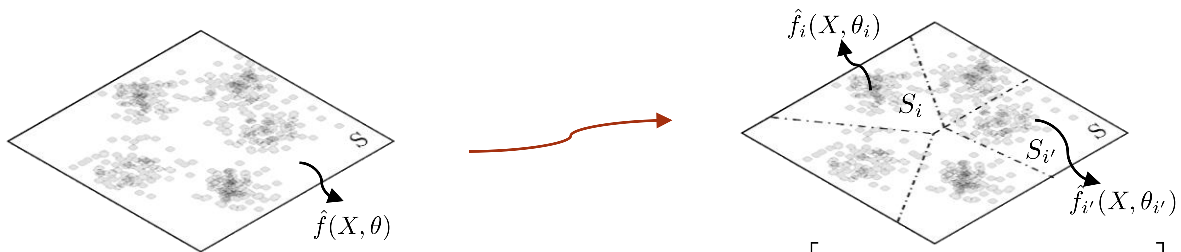


$$\min_{\theta}\ \mathbb{E}\left[d\left(f(X),\hat{f}(X,\theta)\right)\right]$$

$$\min_{\{S_i,\theta_i\}}\ \mathbb{E}\left[\sum_i \mathbb{1}_{[X\in S_i]}d\left(f(X),\hat{f}_i(X,\theta_i)\right)\right]$$

Progressive Learning for Cyber-Physical Systems

- **Divide and Conquer**
  - Partition the space and use local models



$$\min_{\theta} \; \mathbb{E}\left[d\left(f(X), \hat{f}(X,\theta)\right)\right]$$

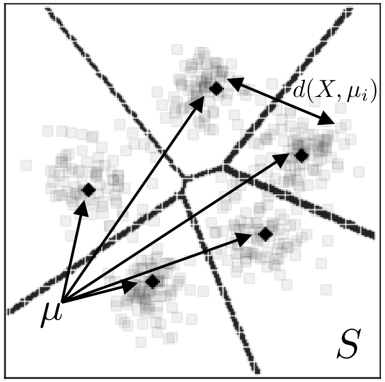
$$\min_{\{S_i,\theta_i\}} \; \mathbb{E}\left[\sum_i \mathbb{1}_{[X \in S_i]} d\left(f(X), \hat{f}_i(X,\theta_i)\right)\right]$$

- **Questions?**
  - **How many regions?**
    - Start with few and add as needed?
  - **Optimal parameters?**
    - Local minima? Gradients?
  - **Simultaneously learn local models?**

Online Deterministic Annealing

- *Observations:* $X^N := \{x_i\}_{i=1}^N,\ x_i \in S$ *realizations of a r.v.* $X \in S$
- *Codevectors:* $\mu = \{\mu_i\}_{i=1}^M,\ \mu_i \in S$ *domain of a r.v.* $Q \in S$

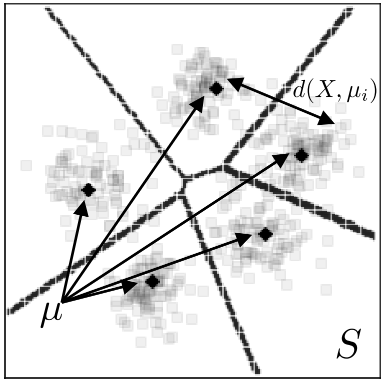
  defined by: $p(\mu_i|x) = \mathbb{P}\left[Q = \mu_i | X = x\right]$
- *Dissimilarity:* $d : S \times S \to [0, \infty)$



Online Deterministic Annealing

- *Observations:* $X^N := \{x_i\}_{i=1}^N,\ x_i \in S$ *realizations of a r.v.* $X \in S$
- *Codevectors:* $\mu = \{\mu_i\}_{i=1}^M,\ \mu_i \in S$ *domain of a r.v.* $Q \in S$

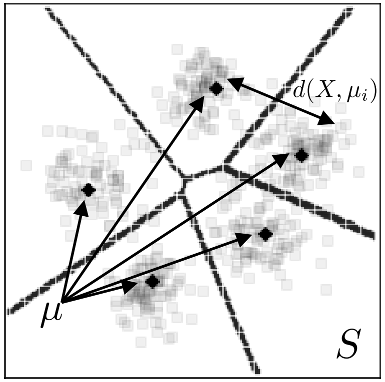
  defined by: $p(\mu_i|x) = \mathbb{P}\left[Q = \mu_i | X = x\right]$
- *Dissimilarity:* $d : S \times S \to [0, \infty)$



Online Deterministic Annealing

- *Observations:* $X^N := \{x_i\}_{i=1}^N,\ x_i \in S$ *realizations of a r.v.* $X \in S$
- *Codevectors:* $\mu = \{\mu_i\}_{i=1}^M,\ \mu_i \in S$ *domain of a r.v.* $Q \in S$

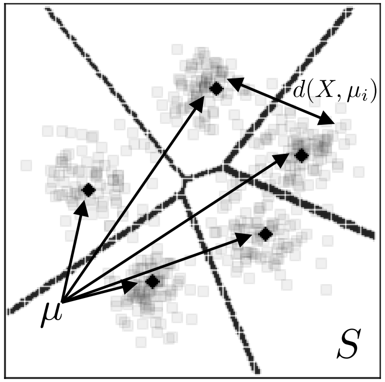
  defined by: $p(\mu_i|x) = \mathbb{P}\left[Q = \mu_i | X = x\right]$
- *Dissimilarity:* $d : S \times S \to [0, \infty)$



Online Deterministic Annealing

- *Observations:* $X^N := \{x_i\}_{i=1}^N,\ x_i \in S$ *realizations of a r.v.* $X \in S$
- *Codevectors:* $\mu = \{\mu_i\}_{i=1}^M,\ \mu_i \in S$ *domain of a r.v.* $Q \in S$

  defined by: $p(\mu_i|x) = \mathbb{P}\left[Q = \mu_i | X = x\right]$
- *Dissimilarity:* $d : S \times S \to [0, \infty)$



Problem Formulation

Solve: $\min_{\mu} \ F_T := D - TH$ for decreasing values of T.

where Distortion: $D(X,Q) := \mathbb{E}\left[d\left(X,Q\right)\right] = \int p(x) \sum_i p(\mu_i|x) d_\phi(x,\mu_i)\ \mathrm{d}x$

Entropy: $H(X,Q) := \mathbb{E}\left[-\log P(X,Q)\right] = H(X) - \int p(x) \sum_i p(\mu_i|x) \log p(\mu_i|x)\ \mathrm{d}x$

Lagrange Coefficient: T Controls Tradeoff
Simulates Annealing Optimization
Triggers Bifurcation (finds number of codevectors)

Mavridis, Baras, Online Deterministic Annealing for Classification and Clustering, IEEE TNNLS 2022.
Mavridis, Baras, Annealing Optimization for Progressive Learning with Stochastic Approximation, IEEE TAC 2022.

Online Deterministic Annealing (II)

Solving the Optimization Problem $\min F_T := D - TH$

- **Lemma.** *The solution to* $F^*(\mu) := \min_{\{p(\mu_i|x)\}} F(\mu)$ *s.t.* $\sum_i p(\mu_i|x) = 1$, *is given by the Gibbs distributions*
$$p^*(\mu_i|x) = \frac{e^{-\frac{d(x,\mu_i)}{T}}}{\sum_j e^{-\frac{d(x,\mu_j)}{T}}}, \ \forall x \in S.$$

- **Theorem.** *The solution to* $\min_\mu F^*(\mu)$ *is given by*
$$\mu_i^* = \mathbb{E}\left[X|\mu_i\right] = \frac{\int xp(x)p^*(\mu_i|x)\ dx}{p^*(\mu_i)}$$
if $d := d_\phi$ *is a* Bregman divergence. (sufficient condition)

e.g., squared Euclidean distance, KL divergence, …

Online Deterministic Annealing (III)

Solving the Optimization Problem $\min F_T := D - TH$

- **Theorem.** *The recursive training rule*

$$\begin{cases} \rho_i(n+1) & = \rho_i(n) + \alpha(n)\left[\hat{p}(\mu_i|x_n) - \rho_i(n)\right] \\ \sigma_i(n+1) & = \sigma_i(n) + \alpha(n)\left[x_n \hat{p}(\mu_i|x_n) - \sigma_i(n)\right] \end{cases}$$

where the quantities $\hat{p}(\mu_i|x_n)$ *and* $\mu_i(n)$ *are recursively updated as follows:*

$$\hat{p}(\mu_i|x_n) = \frac{\rho_i(n) e^{-\frac{d(x_n,\mu_i(n))}{T}}}{\sum_i \rho_i(n) e^{-\frac{d(x_n,\mu_i(n))}{T}}}$$

$$\mu_i(n) = \frac{\sigma_i(n)}{\rho_i(n)},$$

converges almost surely to a possibly sample path dependent solution of the optimization $\min_\mu F^*(\mu)$*, as* $n \to \infty$.

$\sigma_n \to \mathbb{E}\left[\mathbb{1}_{[\mu]} X\right]$
$\rho_n \to \mathbb{P}[\mu]$

Stochastic Approximation: Gradient-Free !

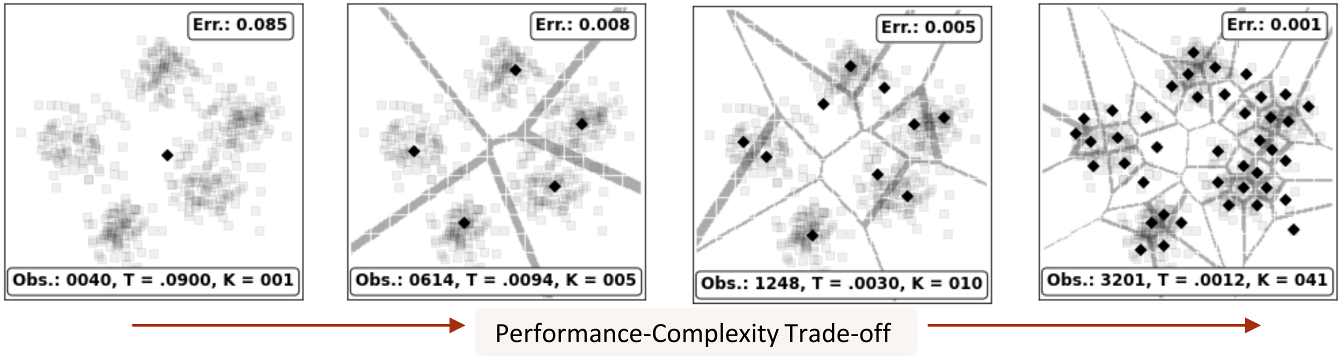
Online Deterministic Annealing (IV)

Bifurcation and the number of codevectors

- Sequentially solve: $\min F_{T_\infty} := D - T_\infty H$

  ...

  $\min F_{T_0} := D - T_0 H$, $T_i < T_{i+1}$: Decreasing Temperature
- **Remark.** *As $T \to \infty$, we get $\mu_i = \mathbb{E}\left[f(X)\right]$, $\forall i$, i.e., one unique pseudo-input.*
- **Remark.** *As T is lowered below a critical value, a bifurcation phenomenon occurs, and the number of pseudo-inputs increases.*



Online Deterministic Annealing (V)

Algorithmic Implementation

Algorithm 1 Online Deterministic Annealing

Initialize
while Termination Criterion **do**
  Perturb $\mu^i \leftarrow \{\mu^i + \delta, \mu^i - \delta\}, \forall i$
  repeat
    Observe (x, c)
    for $i = 1, \ldots, K$ **do**
      $s^i = \mathbb{1}_{[c_{\mu^i} = c]}$
      Update:

$$p(\mu^i|x) \leftarrow \frac{p(\mu^i)e^{-\frac{d_\phi(x,\mu^i)}{T}}}{\sum_i p(\mu^i)e^{-\frac{d_\phi(x,\mu^i)}{T}}}$$

$$p(\mu^i) \leftarrow p(\mu^i) + \alpha_n \left[s^i p(\mu^i|x) - p(\mu^i)\right]$$

$$\sigma(\mu^i) \leftarrow \sigma(\mu^i) + \alpha_n \left[s^i x p(\mu^i|x) - \sigma(\mu^i)\right]$$

$$\mu^i \leftarrow \frac{\sigma(\mu^i)}{p(\mu^i)}$$

    end for
  until Convergence
  Keep effective codevectors
  Remove idle codevectors
  Lower temperature $T \leftarrow \gamma T$
end while

Fix T

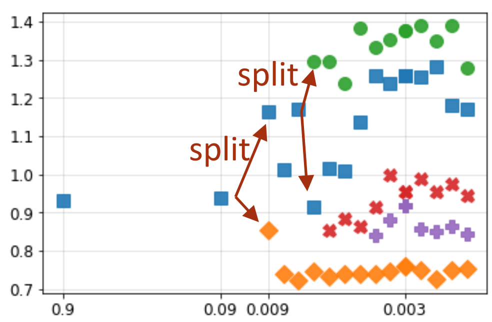
Perturb $\{\mu_i\}$

Observe $f(x)$

Update all μ_i

When Converged:
  Detect Bifurcation

  Lower T



Detect Bifurcation by perturbing the codevectors
- Will merge or separate → Critical Temperatures

Online Deterministic Annealing (VI)

Training Local Models: Two-Timescale Stochastic Approximation

Algorithm 1 Online Deterministic Annealing

Initialize
while Termination Criterion **do**
  Perturb $\mu^i \leftarrow \{\mu^i + \delta, \mu^i - \delta\}, \forall i$
  repeat
    Observe (x, c)
    for $i = 1, \ldots, K$ **do**
      $s^i = \mathbb{1}_{[c_{\mu^i} = c]}$
      Update:

$$p(\mu^i|x) \leftarrow \frac{p(\mu^i) e^{-\frac{d_\phi(x,\mu^i)}{T}}}{\sum_i p(\mu^i) e^{-\frac{d_\phi(x,\mu^i)}{T}}}$$

$$p(\mu^i) \leftarrow p(\mu^i) + \alpha_n \left[s^i p(\mu^i|x) - p(\mu^i) \right]$$

$$\sigma(\mu^i) \leftarrow \sigma(\mu^i) + \alpha_n \left[s^i x p(\mu^i|x) - \sigma(\mu^i) \right]$$

$$\mu^i \leftarrow \frac{\sigma(\mu^i)}{p(\mu^i)}$$

    end for
  until Convergence
  Keep effective codevectors
  Remove idle codevectors
  Lower temperature $T \leftarrow \gamma T$
end while

Slow SA

Change your model

Fix T

Perturb $\{\mu_i\}$

Observe $f(x)$

Update all μ_i

When Converged:
  Detect Bifurcation

Lower T

$$\theta_{n+1} = \theta_n + \alpha(n) \left[f(\theta_n, \mu_n) + M^{(\theta)}_{n+1} \right]$$

$$\mu_{n+1} = \mu_n + \beta(n) \left[g(\theta_n, \mu_n) + M^{(\mu)}_{n+1} \right]$$

$$\frac{\beta(n)}{\alpha(n)} \to 0$$

Fast SA $\left\{ \begin{array}{l} \text{Function Approximation} \\ \text{Q-Learning} \end{array} \right.$

$$\Delta\theta = -\beta_n \nabla_\theta g(x, \theta, \mu_i)$$

Train your model

Mavridis, Baras, et al., Gaussian Process Regression using Progressively Growing Learning Representations, IEEE CDC 2022.
Mavridis, Baras, Annealing Optimization for Progressive Learning with Stochastic Approximation, IEEE TAC 2022.

The Institute for Systems Research

Online Deterministic Annealing (VI)

Training Local Models: Two-Timescale Stochastic Approximation

Algorithm 1 Online Deterministic Annealing

```
Initialize
while Termination Criterion do
  Perturb μ^i ← {μ^i + δ, μ^i − δ}, ∀i
  repeat
    Observe (x, c)
    for i = 1, ..., K do
      s^i = 1[c_{μ^i} = c]
      Update:
    end for
  until Convergence
  Keep effective codevectors
  Remove idle codevectors
  Lower temperature T ← γT
end while
```

$$p(\mu^i|x) \leftarrow \frac{p(\mu^i)e^{-\frac{d_\phi(x,\mu^i)}{T}}}{\sum_i p(\mu^i)e^{-\frac{d_\phi(x,\mu^i)}{T}}}$$

$$p(\mu^i) \leftarrow p(\mu^i) + \alpha_n\left[s^i p(\mu^i|x) - p(\mu^i)\right]$$

$$\sigma(\mu^i) \leftarrow \sigma(\mu^i) + \alpha_n\left[s^i x p(\mu^i|x) - \sigma(\mu^i)\right]$$

$$\mu^i \leftarrow \frac{\sigma(\mu^i)}{p(\mu^i)}$$

Slow SA

Change your model

Fix T

Perturb $\{\mu_i\}$

Observe $f(x)$

Update all μ_i

When Converged:
Detect Bifurcation

Lower T

$$\theta_{n+1} = \theta_n + \alpha(n)\left[f(\theta_n,\mu_n) + M^{(\theta)}_{n+1}\right]$$

$$\mu_{n+1} = \mu_n + \beta(n)\left[g(\theta_n,\mu_n) + M^{(\mu)}_{n+1}\right]$$

$$\frac{\beta(n)}{\alpha(n)} \to 0$$

Fast SA $\left\{ \begin{array}{l} \text{Function Approximation} \\ \text{Q-Learning} \end{array} \right.$

$$\Delta\theta = -\beta_n \nabla_\theta g(x,\theta,\mu_i)$$

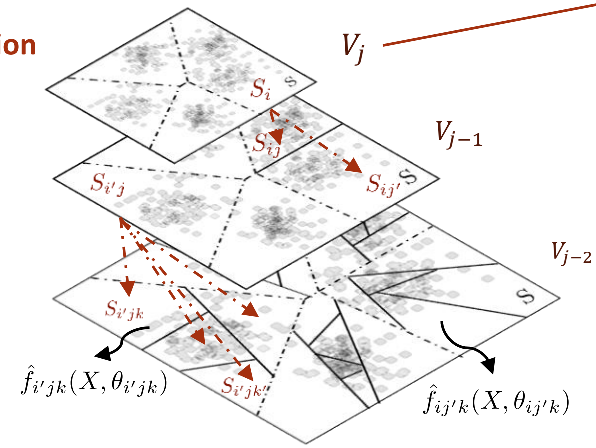
Train your model

- https://github.com/MavridisChristos/OnlineDeterministicAnnealing

Online Deterministic Annealing (VII)

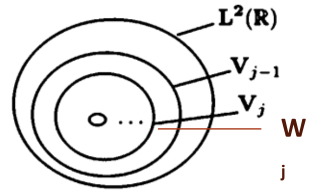
Multi-Resolution Hierarchical Learning

Low Resolution



High Resolution

Example: Group-convolution Wavelets



$$\mathbf{V}_{j-1} = \mathbf{V}_j \oplus \mathbf{W}_j.$$

- **Constructive (Structured Representation)**
- **Provably Consistent**
- **Localization**
  - Emphasis on regions with high error
- **Asynchronous/Parallel Computation**
- **Reduced Complexity**

Mavridis, Baras, Multi-Resolution Online Deterministic Annealing: A Hierarchical and Progressive Learning Architecture [under review].
Mavridis, Baras, Towards the One Learning Algorithm Hypothesis: A System-theoretic Approach [under review].

Bifurcation and the number of Codevectors

- **Theorem.** *Bifurcation occurs under the following condition*

$$\exists y_n \ s.t.\ p(y_n) > 0 \ and \ \det\left[I - T\frac{\partial^2 \phi(y_n)}{\partial y_n^2} C_{x|y_n}\right] = 0$$

where $C_{x|y_n} := \mathbb{E}\left[(x-y_n)(x-y_n)^{\mathrm{T}}|y_n\right]$.

Proof. From variational calculus and the second order condition:

$$\frac{d^2}{d\epsilon^2} F^*(\{\mu + \epsilon\psi\})|_{\epsilon=0} \geq 0$$

$\square$

- T_c depends on:
  - The Bregman divergence
  - The data space

The Institute for Systems Research

Online Deterministic Annealing (ODA)

Algorithmic Implementation & Open-Source Code

Algorithm 1 Online Deterministic Annealing

```
Initialize
while Termination Criterion do
    Perturb μ^i ← {μ^i + δ, μ^i − δ}, ∀i
    repeat
        Observe (x, c)
        for i = 1, ..., K do
            s^i = 1[c_{μ^i} = c]
            Update:
```

$$p(\mu^i|x) \leftarrow \frac{p(\mu^i)e^{-\frac{d_\phi(x,\mu^i)}{T}}}{\sum_i p(\mu^i)e^{-\frac{d_\phi(x,\mu^i)}{T}}}$$

$$p(\mu^i) \leftarrow p(\mu^i) + \alpha_n \left[s^i p(\mu^i|x) - p(\mu^i)\right]$$

$$\sigma(\mu^i) \leftarrow \sigma(\mu^i) + \alpha_n \left[s^i x p(\mu^i|x) - \sigma(\mu^i)\right]$$

$$\mu^i \leftarrow \frac{\sigma(\mu^i)}{p(\mu^i)}$$

```
        end for
    until Convergence
    Keep effective codevectors
    Remove idle codevectors
    Lower temperature T ← γT
end while
```

Detect Bifurcation by perturbing the set of models

- Will merge or separate → Critical Temperatures

➢ https://github.com/MavridisChristos/OnlineDeterministicAnnealing

Why Maximum Entropy?

- **Bifurcation:** Progressively grow set of models
- **Jayne's Maximum Entropy Principle**
  - Most "Unbiased" estimator: each sub-problem induces "good" initial conditions for the next
  - Duality (Legendre-type) and Regularization:

$$\frac{1}{\beta}\log \mathbb{E}_{P_\mu}\left[e^{\beta Z}\right] = \inf_{P_\nu \in \mathcal{P}(\Omega)} \left\{ \mathbb{E}_{P_\nu}\left[Z\right] - \frac{1}{\beta} D_{KL}(P_\nu, P_\mu) \right\},\ \beta < 0$$

$$\min F_T \ \simeq\ \min \frac{1}{\beta}\log \mathbb{E}\left[e^{\beta D}\right],\ \beta = -\frac{1}{T}$$

Risk-Sensitivity

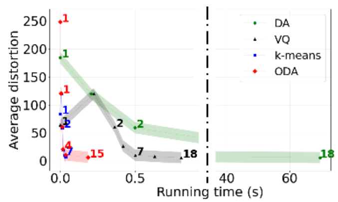
$$\frac{1}{\beta}\log \mathbb{E}\left[e^{\beta J}\right] = \mathbb{E}\left[J\right] + \frac{\beta}{2}\mathrm{Var}\left[J\right] + O(\beta^2)$$

  - Robustness w.r.t. initial conditions, input perturbations.

Mavridis, Baras, et al., Risk Sensitivity and Entropy Regularization in Prototype-based Learning, IEEE MED 2022.

Online Deterministic Annealing (ODA)

Clustering



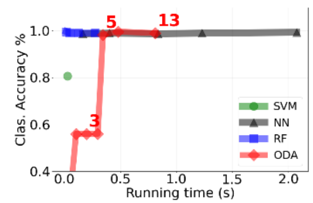
- **“Progressively” finds number of clusters**
- Much fewer samples than k-means
- **Online!**

❖ **Complexity**

$$O(N_c(2\bar{K})^2 d)$$

$$N_T \le \bar{K} \le \min\left\{ \sum_{n=0}^{N_T-1} 2^n, \sum_{n=0}^{\log_2 K_{max}} 2^n \right\} < N_T K_{max}$$

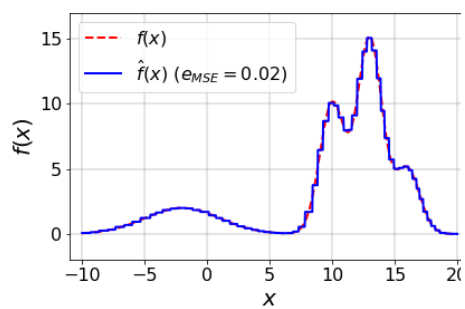
Classification



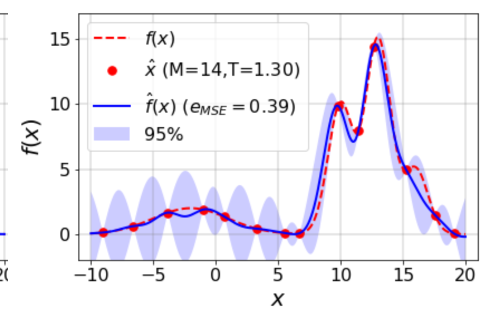
Data set	ODA	SVM	NN	RF
Gaussian	98.9 ± 0.0	79.5 ± 0.0	98.6 ± 0.0	98.7 ± 0.0
WBCD	90.7 ± 0.0	85.6 ± 0.0	92.7 ± 0.0	94.6 ± 0.0
Credit (F1)	95.6 ± 0.0	69.1 ± 0.2	58.9 ± 0.1	62.8 ± 0.1
PIMA	70.5 ± 0.0	62.9 ± 0.0	76.3 ± 0.0	74.4 ± 0.0

Unbalanced Dataset
Other models cannot generalize

Regression

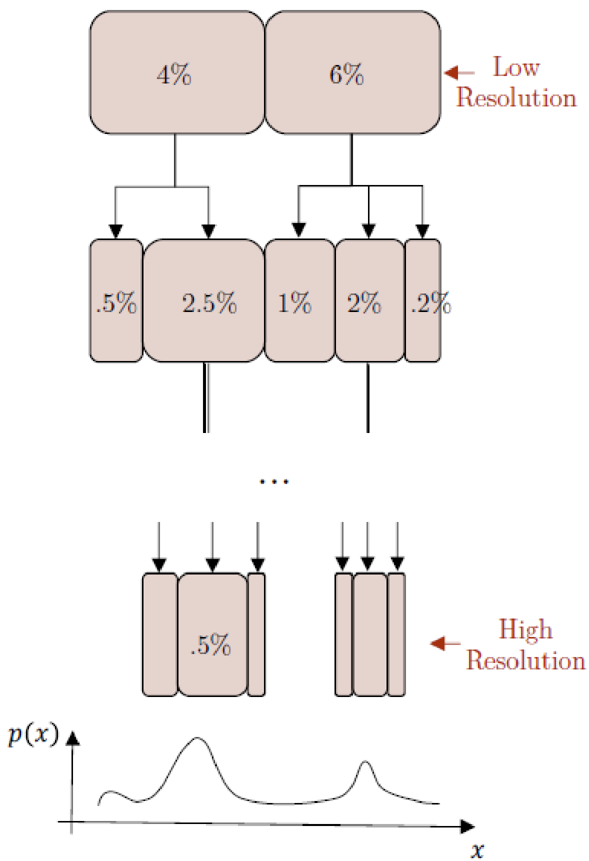


Piece-wise constant approximation



Sparse Gaussian Processes

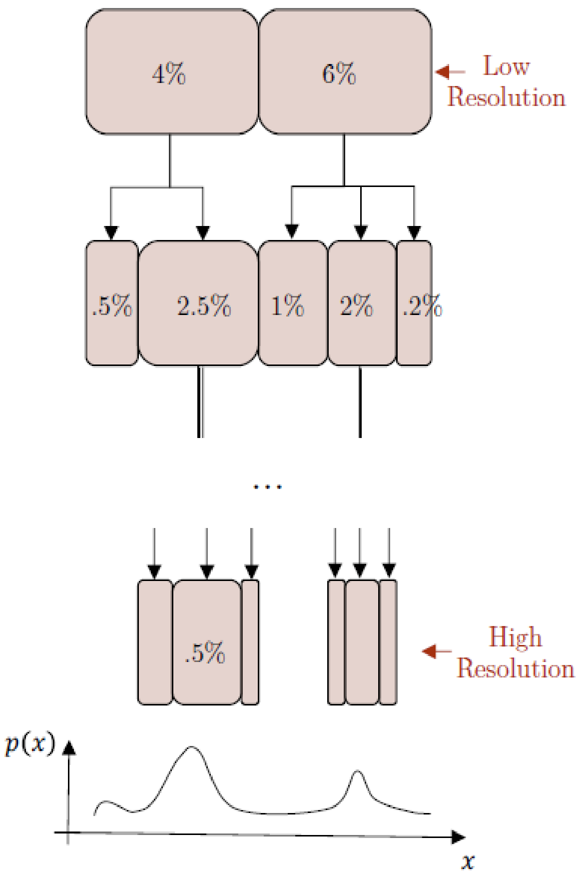
Hierarchical Multi-Resolution Learning

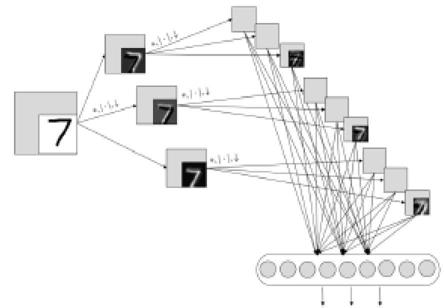


- Constructive
- Provably Consistent
  - Speed depends on probability density
- Localization
  - Emphasis on regions with high error
- Asynchronous/Parallel Computation
- Complexity: $O(|C|^2 + |C| \log_{|C|} K) \ll O(K^2)$
- Non-binary Tree: $|C| > 2 \rightarrow \log_{|C|} K < \log_2 K$
- Online Observations!

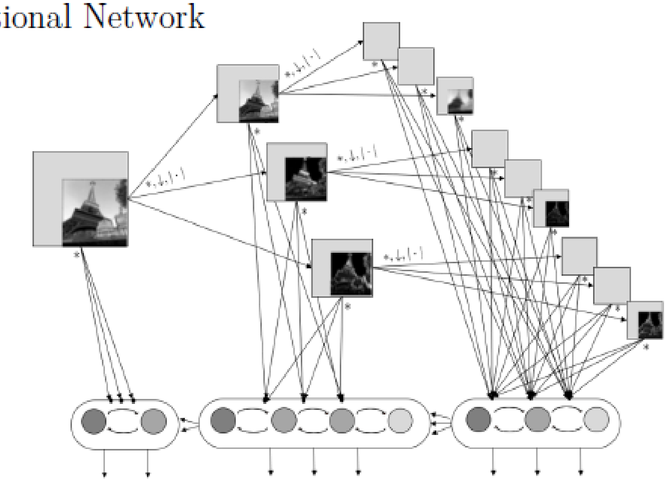
- Mavridis, Baras, Multi-Resolution Online Deterministic Annealing: A Hierarchical and Progressive Learning Architecture [under review].

Hierarchical Multi-Resolution Learning





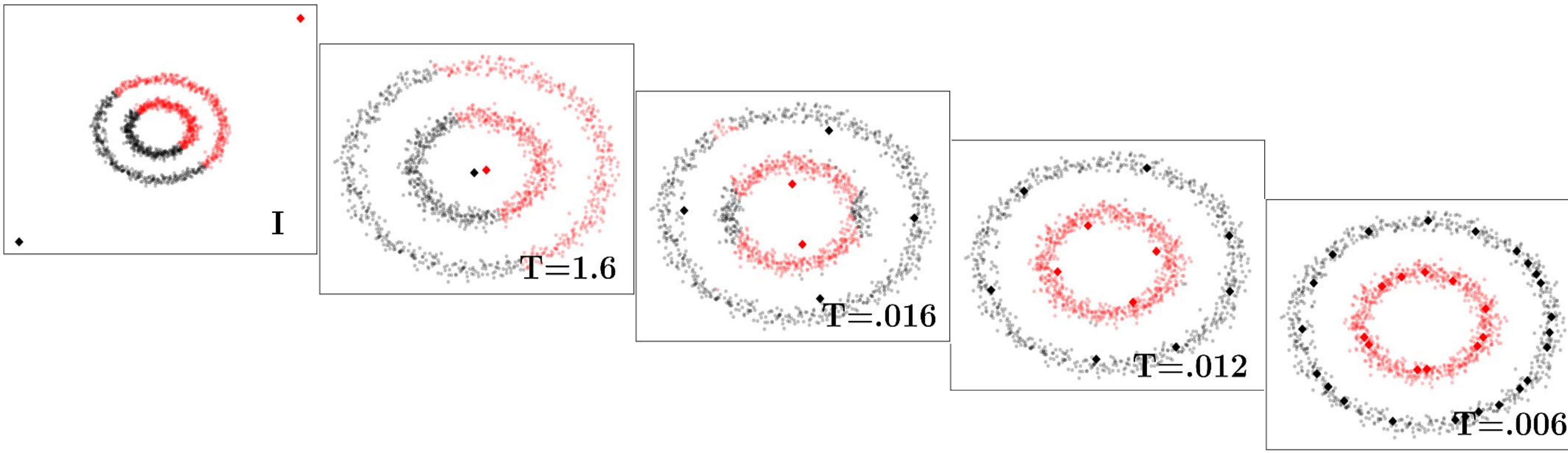
Deep Convolutional Network



This Approach
with Group-Convolutional Wavelets

- Mavridis, Baras, Towards the One Learning Algorithm Hypothesis: A System-theoretic Approach [arXiv:2112.02256]

Online Deterministic Annealing



Toy Example. Evolution in 2D.

Data set	ODA	SVM	NN	RF
Gaussian	$98.9_{\pm 0.0}$	$79.5_{\pm 0.0}$	$98.6_{\pm 0.0}$	$98.7_{\pm 0.0}$
WBCD	$90.7_{\pm 0.0}$	$85.6_{\pm 0.0}$	$92.7_{\pm 0.0}$	$94.6_{\pm 0.0}$
Credit (F1)	$95.6_{\pm 0.0}$	$69.1_{\pm 0.2}$	$58.9_{\pm 0.1}$	$62.8_{\pm 0.1}$
PIMA	$70.5_{\pm 0.0}$	$62.9_{\pm 0.0}$	$76.3_{\pm 0.0}$	$74.4_{\pm 0.0}$

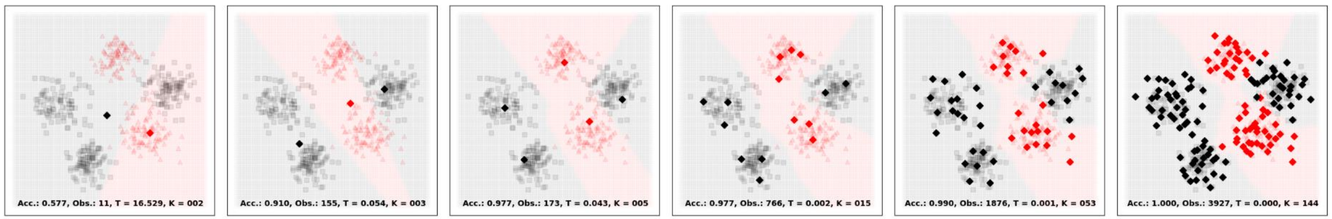
Unbalanced Dataset
Other models cannot generalize

Classification accuracies in 5-fold cross-validation for 4 datasets*.

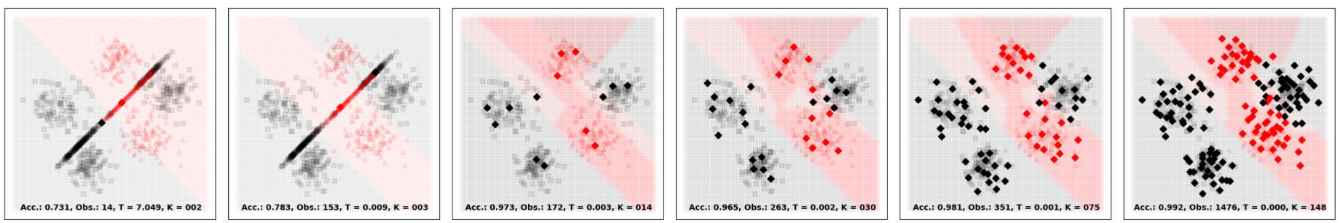
*Mavridis and Baras, Online Deterministic Annealing for Classification and Clustering, arXiv 2021 (IEEE TNNLS+)

Multi-Resolution ODA in 2D

- Toy Examples



Evolution of ODA in 2D.



Evolution of multi-resolution ODA in 1D (first principal component) and 2D.

Algorithm 2 Multi-Resolution ODA Algorithm

Set temperature schedule: $\overline{T} = \{\overline{T}_{\tilde{l}}, \overline{T}_{\tilde{l}-1}, \ldots, \overline{T}_0\}$, $\underline{T} = \{\underline{T}_{\tilde{l}}, \underline{T}_{\tilde{l}-1}, \ldots, \underline{T}_0\}$
Initialize $\nu_0^{(0)}$, M_{ν_0}, V_{ν_0}.
repeat
  Observe data point (X, c)
  $w = \nu_0$, $l = \tilde{l}$, $x = X_{\tilde{l}}$
  while $C(w) \neq \emptyset$ **do**
    $w = v \in C(w)$ such that $x \in S_v$
    $l = l - 1$
    $x = X_l$
  end while
  Update M_w using Alg. 1 in S_w with $(T_{max} = \overline{T}_l, T_{min} = \underline{T}_l)$
  if ODA in S_w converged **and** $l < \tilde{l}$ **then**
    Split w to $C(w)$ with respect to V_w
  end if
until Convergence

Multi Resolution ODA

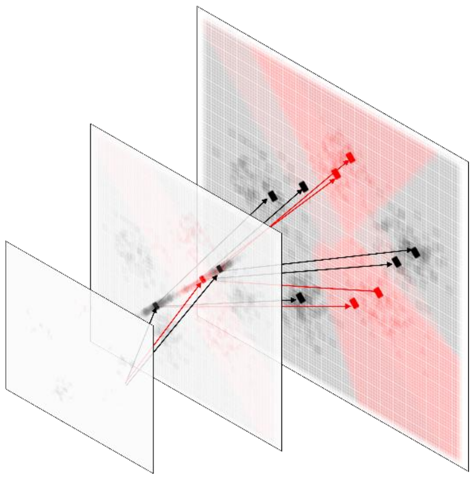
Vector Quantization
- Competitive-Learning NN
- Interpretable
- Robust
- Topology-Preserving
- Sparse in Memory

Deterministic Annealing
- Progressively Growing in Size
- Performance/Complexity Trade-off
- Avoids Poor Local Minima

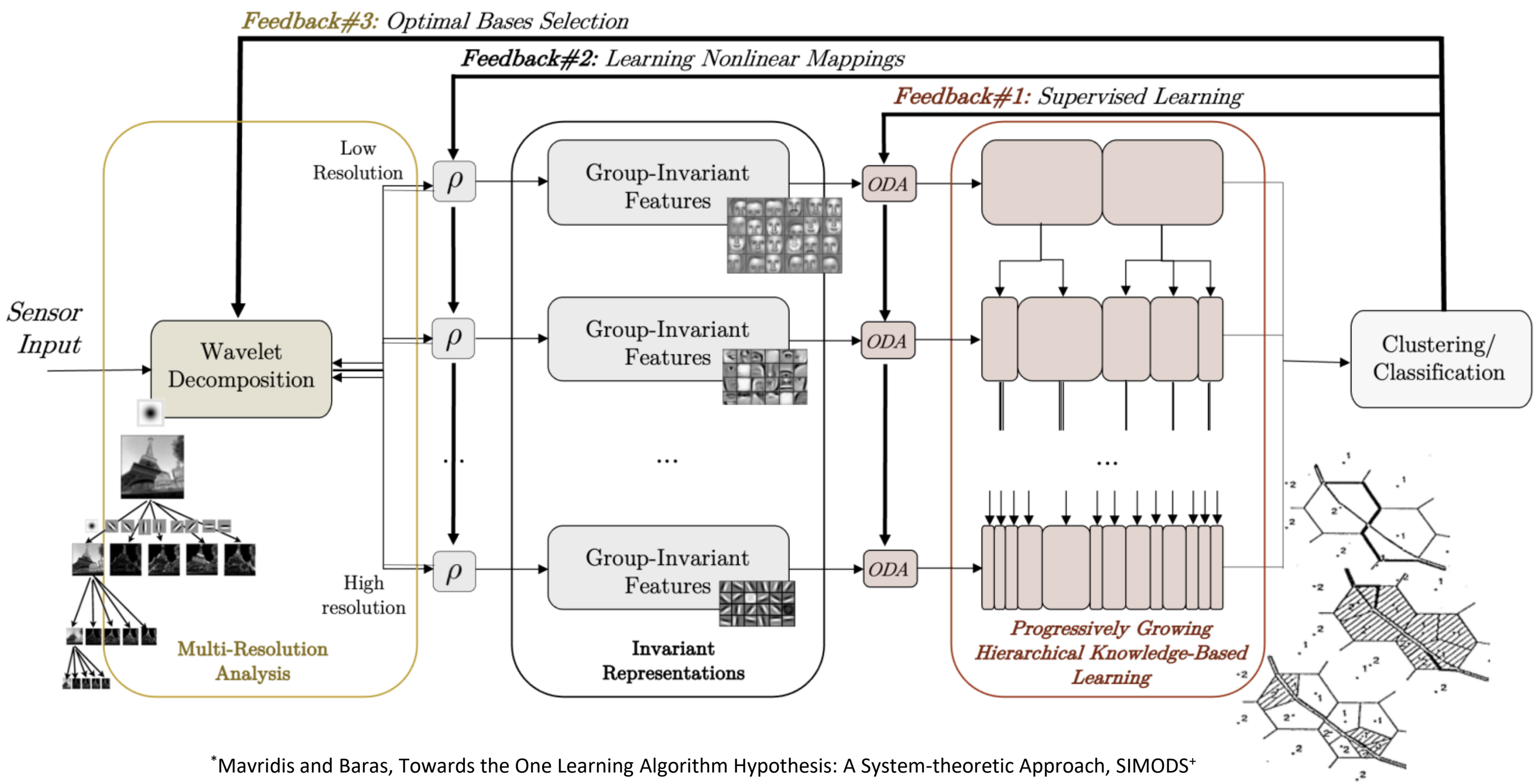
Bregman Divergences
- Works in Vector Spaces & Modules

Stochastic Approximation
- Online Learning Rule
- Needs Fewer Samples

Wavelets & Tree
- Hierarchical Invariant Representations

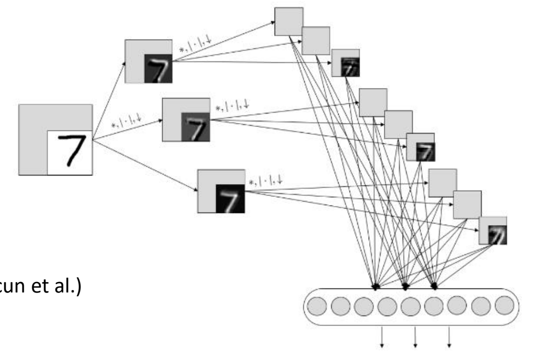


A Universal Learning Architecture (revisited)



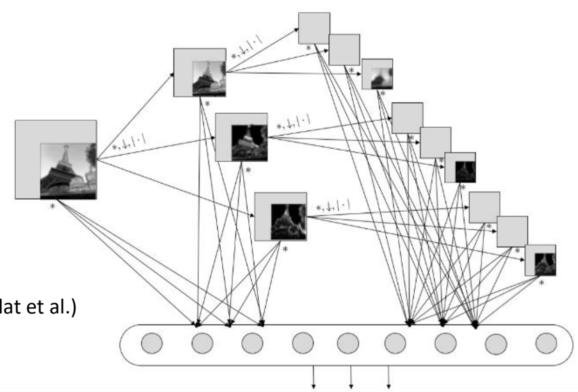
*Mavridis and Baras, Towards the One Learning Algorithm Hypothesis: A System-theoretic Approach, SIMODS+

A Deep Learning Architecture?



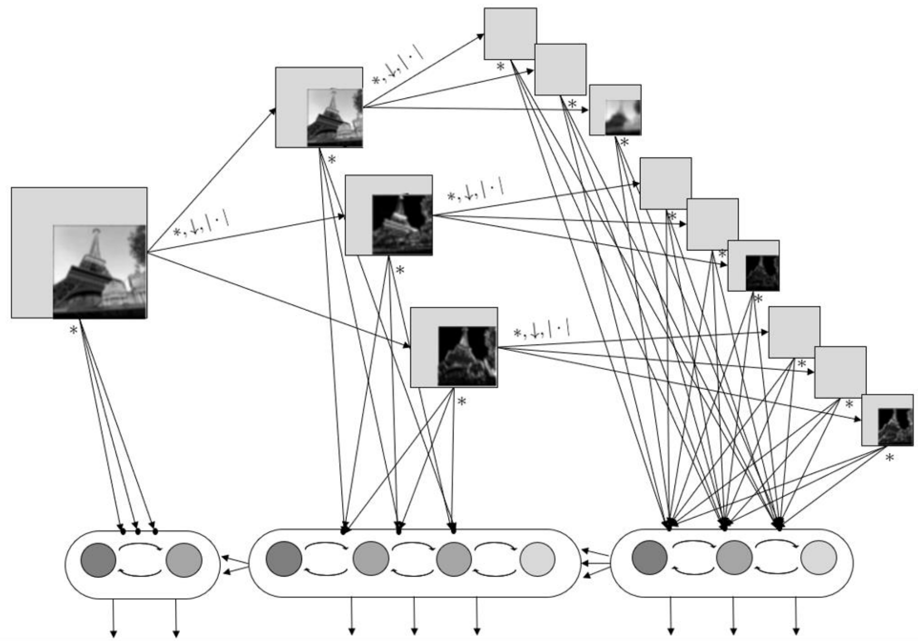
(Lecun et al.)

Deep Convolutional Network



(Mallat et al.)

Scattering Convolutional Network



Our Approach

Applications

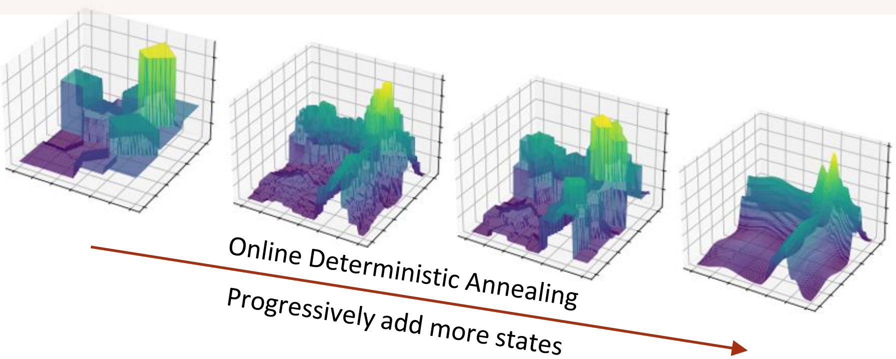
- **Robust ML –** against missing data, noise, attacks
- **Face recognition**
- **Simultaneous** sound direction of arrival, instrument paying, note playing (or person speaking, vowel identification)
- **Robust Reinforcement Learning**
- **CPS Security**
- **Robotics & Multi-Agent Systems**
- **Learning with Progressively Growing Knowledge Representations** for Decision-Making Systems
- Towards a Neuroscience-inspired Universal Learning Algorithm: **Hierarchical, Memory-based, Progressive, Interpretable, Robust**
- Adaptive Space Aggregation for **Memory-Efficient Reinforcement Learning** in Robot Control
- Progressive **Graph Partitioning** and Image Segmentation
- **Community detection** on graphs
- **Hardware implementation** via hybrid (digital and neuromorphic) chips

Explainable Reinforcement Learning

<u>**Optimal Control Problem:**</u> Given an MDP $(\mathcal{X}, \mathcal{U}, \mathcal{P}, C)$ $\mathcal{X} \times \mathcal{U} \in \mathbb{R}^{d \times m}$

solve: $$\min_u \; J(u) := \mathbb{E}\left[\sum_{l=\tau}^{\infty} \gamma^l C(x_l, u_l)\right]\Bigg|_{\tau=0} := Q(x_\tau, u_\tau)\Bigg|_{\tau=0}$$

- **Q-Learning** $Q_{j+1}(x, u') = Q_j(x, u') + \alpha_j[C(x, u') + \gamma \min_u Q_j(x', u) - Q_j(x, u')]$
    - Assumes Discrete Space
    - Is a stochastic approximation algorithm
- ~~Ad hoc discretization~~ → **Adaptive State/Action Aggregation with ODA**



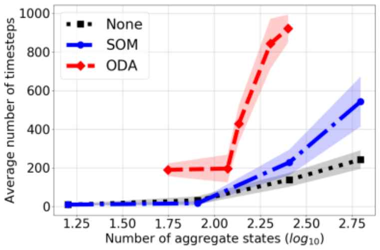
Explainable Reinforcement Learning

Optimal Control Problem: Given an MDP $(\mathcal{X}, \mathcal{U}, \mathcal{P}, C)$ $\mathcal{X} \times \mathcal{U} \in \mathbb{R}^{d\times m}$

solve: $\min_u J(u) := \mathbb{E}\left[\sum_{l=\tau}^{\infty} \gamma^l C(x_l, u_l)\right]\Big|_{\tau=0} := Q(x_\tau, u_\tau)\Big|_{\tau=0}$

- **Q-Learning** $Q_{j+1}(x,u') = Q_j(x,u') + \alpha_j[C(x,u') + \gamma \min_u Q_j(x',u) - Q_j(x,u')]$
  - Assumes Discrete Space
  - Is a stochastic approximation algorithm
- ~~Ad hoc discretization~~ → **Adaptive State/Action Aggregation with ODA**
- **Stochastic Approximation in Two Timescales**
  - **Fast Component: *Q-Learning***
  - **Slow Component: *ODA***

$$\begin{cases} x_{n+1} = x_n + \alpha(n)\left[f(x_n,y_n) + M_{n+1}^{(x)}\right] \\ y_{n+1} = y_n + \beta(n)\left[g(x_n,y_n) + M_{n+1}^{(y)}\right] \end{cases}, \quad \frac{\beta(n)}{\alpha(n)} \to 0$$



Mavridis, Baras, Maximum-Entropy Progressive State Aggregation for Reinforcement Learning, IEEE CDC 2021.
Mavridis, Baras, Annealing Optimization for Progressive Learning with Stochastic Approximation, IEEE TAC 2022.

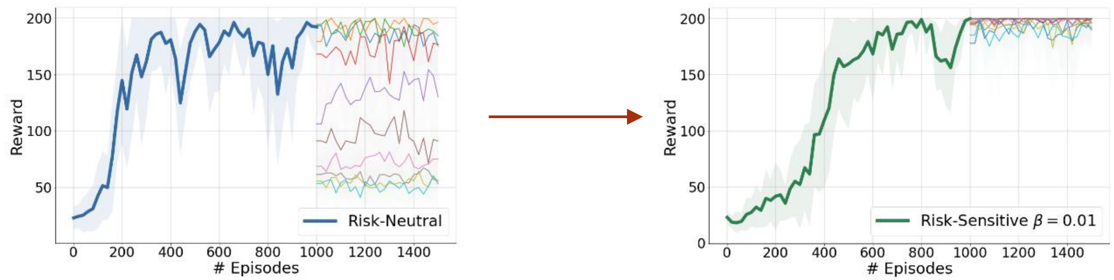
Risk-Sensitive Reinforcement Learning

<u>Optimal Control Problem:</u> Given an MDP $(\mathcal{X}, \mathcal{U}, \mathcal{P}, C)$ $\mathcal{X} \times \mathcal{U} \in \mathbb{R}^{d \times m}$

solve: $\min_u \; J(u) := \mathbb{E}\left[\sum_{l=\tau}^{\infty} \gamma^l C(x_l, u_l)\right]\Bigg|_{\tau=0} := Q(x_\tau, u_\tau)\Big|_{\tau=0}$

$$\min_u \frac{1}{\beta} \log \mathbb{E}\left[e^{\beta J}\right] = \min_u \begin{cases} \sup_{P_\nu \in \mathcal{P}(\Omega)} \left\{\mathbb{E}_{P_\nu}[J] - \frac{1}{\beta} D_{KL}(P_\nu, P_\mu)\right\}, & \beta > 0 \\ \inf_{P_\nu \in \mathcal{P}(\Omega)} \left\{\mathbb{E}_{P_\nu}[J] - \frac{1}{\beta} D_{KL}(P_\nu, P_\mu)\right\}, & \beta < 0 \end{cases}$$

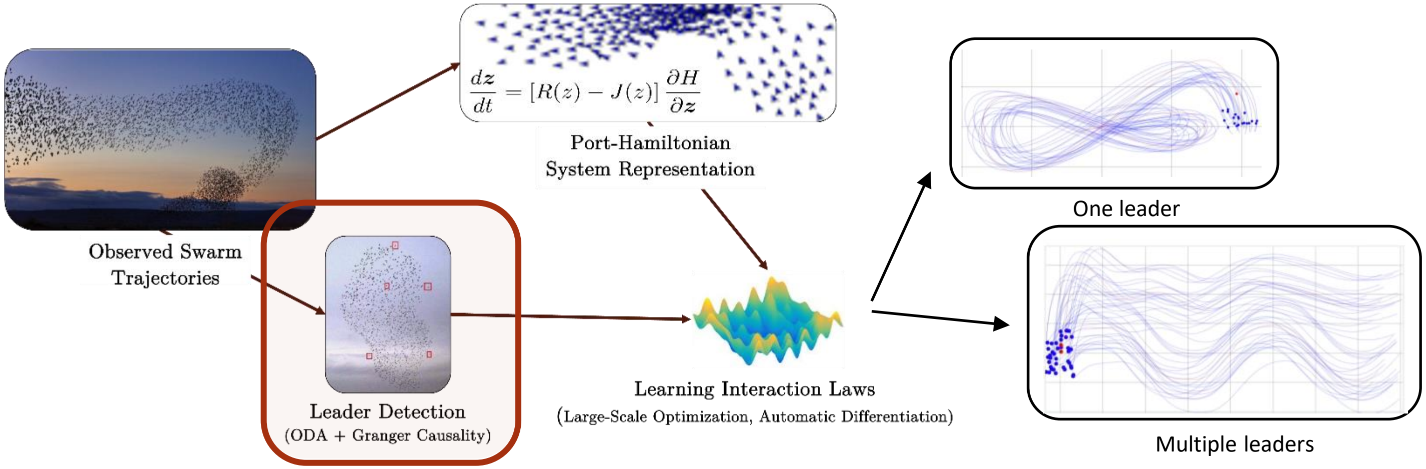
- **Multiplicative Bellman equation:** $V_\beta^*(x_k) = \min_u \beta \left\{ e^{\beta C(x_k, u_k)} \mathbb{E}\left[e^{\beta \gamma V_\beta^*(x_{k+1})} \mid x_k\right] \right\}$



Noorani , Baras, et al., Risk-Sensitive Policy-Gradient Reinforcement Learning with Exponential Criteria [under review].
Noorani, Baras, et al., Risk-Sensitive Reinforcement Learning for Coordination Games [under review].

Application in Robotics & Multi-Agent Systems

Swarm Coordination Laws and Leader Detection

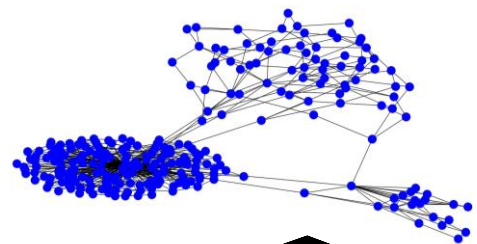


- Application: **Defense against adversarial UAV swarm attacks**

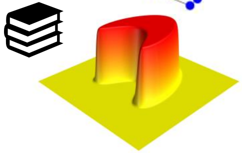
Mavridis , Baras, et al., Detection of Dynamically Changing Leaders in Complex Swarms from Observed Dynamic Data, Springer 2020.
Mavridis , Baras, et al., Learning Swarm Interaction Dynamics from Density Evolution, IEEE TCNS.

Application in Robotics & Multi-Agent Systems

Community Detection on Graphs



- **Adaptive Spectral Clustering**
  - ODA on spectral clustering features
  - Distributed approximation of spectral clustering features



Cyber-Physical Security: Attack Identification in Dynamic Games

$$\dot{x}(t) = Ax(t) + Bu(t) + K_i d_i(t),\ x(0) = x_0,\ t \geq 0$$

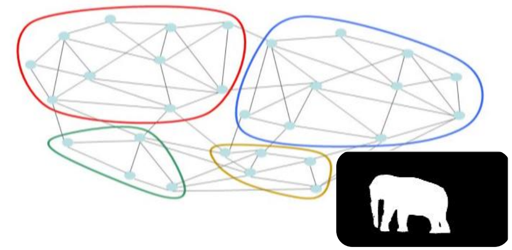
$$V(x) = \min_u \max_d \int_t^\infty \left(x^\mathrm{T} Q x + u^\mathrm{T} R u - \gamma^2 \|d\|^2\right) \mathrm{d}\tau.$$

Tractable Solution: Bounded Rationality + Attack Identification

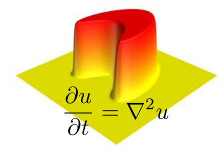
Mavridis, Baras, Progressive Graph Partitioning Based on Information Diffusion, IEEE CDC 2021.

Mavridis, Baras, et al., Attack Identification for Cyber-Physical Security in Dynamic Games under Cognitive Hierarchy [under review].

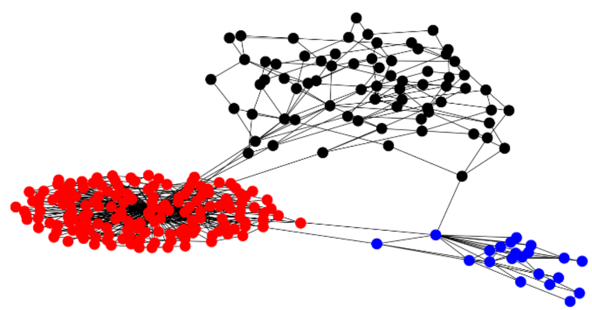
Application in Graph Partitioning



- Spectral Clustering → **Graph Cuts**, **Image Segmentation**
- **Distributed approximation of spectral clustering features**
  - Simulated heat diffusion on graphs $u_i(t+1) = u_i(t) - \sum_{j \in \mathcal{N}(i)} L_{ij} u_j(t)$



- **Adaptive Spectral Clustering using ODA**
  - Progressively growing model (adjusts number of clusters)
  - Online Learning (no need for graph knowledge a priori)
  - Avoids poor local minima



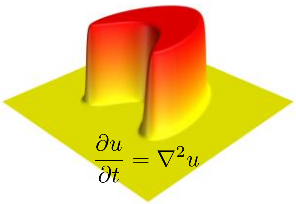
*Mavridis and Baras, Progressive Graph Partitioning Based on Information Diffusion, CDC 2021

Information Diffusion on Graphs

- Discretized Heat equation on graphs: $u_i(t+1) = u_i(t) - \sum_{j \in \mathcal{N}(i)} L_{ij} u_j(t)$

Solution: $u_i(t) = c_1 + (1-\lambda_2)^t \hat{v}_i^{(2)} + \ldots + (1-\lambda_N)^t \hat{v}_i^{(N)}$

where $|(1-\lambda_i)| < 1, \text{ for } i = 2, \ldots, N$



$$u_i^{\infty} = c_1$$

$$\rightarrow \quad \sum_{t=0}^{\infty} u_i(t) - u_i^{\infty} = \sum_{j=2}^{N} \frac{1}{\lambda_j} \hat{v}_i^{(j)} \quad \longrightarrow \quad \boxed{\hat{x}_i := \sum_{t=0}^{N_c} u_i(t) - u_i(N_c)}$$

Weighted sum of eigenvectors

Learning features for Spectral Clustering

Future Directions: Advancing AI and ML – our Approach

- Rigorous Mathematics for Deep Networks – Universal Architecture emerging
- Non von-Neumann computing – do not separate CPU form Memory – Synaptic NN, in-memory processing
- Universal ML -- Integrate Deep NN and Synaptic NN
- Knowledge Representation and Reasoning: Integrate Knowledge Graphs and Semantic Vector Spaces
- Progressive Learning, Knowledge Compacting
- Link Machine Learning with Knowledge Representation and Reasoning

Future Directions

1 ➢ **Hierarchical and Safe Decision-Making**
- Progressively transition from fast sub-optimal to optimal solutions
- Constructing hierarchical and invariant data representations

2 ➢ **Risk-Sensitive & Explainable Reinforcement Learning**
- Connection to Robust Control
- Explainable Policies → Error Correction
- Partially-Observable Systems using the “information state”

3 ➢ **Network Dynamics and Structure**
- Importance of Leaders and Self-Organization
- Heterogeneous Graph Consensus / Decentralized Auctions (Traffic Control)
- Distributed Learning

4 ➢ **Coordination Games**
- Risk-Sensitivity and Trust in Coordination Game Equilibria
- Signaling (Implicit Communication) and Optimal Control

5 ➢ **Intelligent Transportation**
- Mixed-Traffic Control
- Real-time Communication-based CAV Consensus for optimal decisions

6 ➢ **Human-Robot Interaction (& Collaboration)**
- Safety, Real-time Adaptation
- Learning from Human Demonstration

7 ➢ **Augment Human Decision Makers with Machine Intelligence**
- Interpretable Learning models
- Knowledge Representation and Reasoning
- Situational awareness, e.g., assistive robotics, battlefield applications

References https://johnbaras.com/

C.N. Mavridis and J.S. Baras, “Online Deterministic Annealing for Classification and Clustering”, *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1-10, Online: Jan. 7, 2022. DOI: 10.1109/TNNLS.2021.3138676

C. Mavridis, A. Tirumalai, J.S. Baras, “Learning Swarm Interaction Dynamics from Density Evolution”, *IEEE Transactions on Control of Network Systems*, pp. 1-12, DOI: 10.1109/TCNS.2022.3198784, August 16, 2022

C.N. Mavridis and J.S. Baras, “Annealing Optimization for Progressive Learning with Stochastic Approximation”, submitted to *IEEE Transactions on Automatic Control (IEEE TAC).*

I. Matei, C.N. Mavridis, J.S. Baras & M. Zhenirovskyy, “Inferring Particle Interaction Physical Models and Their Dynamical Properties”, *Proc. 2019 IEEE Conference on Decision and Control (CDC)*, pp. 4615-4621, Nice, France, December 11-13, 2019.

I. Matei, M. M. Zhenirovskyy, J. de Kleer, C. Somarakis, and J.S. Baras, “Learning Physical Laws: The Case of Micron Size Particles in Dielectric Fluid”, *Proceedings of the 2020 American Control Conference*, pp. 2949-2954, Denver, CO, July 1-3, 2020.

C.N. Mavridis, and J.S. Baras, “Convergence of Stochastic Vector Quantization and Learning Vector Quantization under Bregman Divergences,” *Proc. 21st Intern. Fed. of Autom. Control World Congress (IFAC 2020)*, pp. 2244-2249, Berlin, July 12-17, 2020.

References https://johnbaras.com/

C. Mavridis, N. Suriyarachchi, and J.S. Baras, "Detection of Dynamically Changing Leaders in Complex Swarms from Observed Dynamic Data", *Proceedings of the GameSec 2020 Conference on Decision andGame Theory for Security*", LCNS 12513, pp. 223-240, Virtual (UMD), October 28-30, 2020.

C. Mavridis, A. Tirumalai, and J. S. Baras, "Learning Interaction Dynamics from Particle Trajectories and Density Evolution", *Proceedings of the 59th IEEE Conference on Decision and Control*, pp. 1014-1019, Jeju Island, Republic of Korea, December 8-11, 2020.

C.N. Mavridis and J.S. Baras, "Vector Quantization for Adaptive State Aggregation in Reinforcement Learning", *Proceedings of the 2021 American Control Conference*, pp. 2187-2192, New Orleans, LA, May 26-28, 2021

C.N. Mavridis and J.S. Baras, "Maximum-Entropy Progressive State Aggregation for Reinforcement Learning," *Proceedings 60th IEEE Conference on Decision and Control (CDC 2021)*, Invited Session on "Learning with Guarantees in Control and Decision-making", pp. 5144-5149, Austin, TX, December 13-15, 2021. DOI: 10.1109/CDC45484.2021.9682927.

C.N. Mavridis and J.S. Baras, "Progressive Graph Partitioning Based on Information Diffusion", *Proceedings of the 60th IEEE Conference on Decision and Control (CDC 2021)*, pp. 37-42, Austin, TX, December 13-15, 2021. DOI: 10.1109/CDC45484.2021.9682799.

References https://johnbaras.com/

C.N. Mavridis, E. Noorani, and J.S. Baras, Risk Sensitivity and Entropy Regularization I Prototype-based Learning," *Proceedings 30th Mediterranean Conference on Control and Automation (MED 2022)*, pp. 194-199, Athens, Greece, June 28 - July 1, 2022.

C.N. Mavridis, G.P. Kontoudis, and J.S. Baras, "Sparse Gaussian Process Regression using Progressively Growing Learning Representations", *Proc. 61st IEEE Conference on Decision and Control (CDC 2022)*, pp. 1454-1459, Cancun, Mexico, Dec. 6-9, 2022.

E. Noorani, C.N. Mavridis and J.S. Baras, "Exponential TD Learning: A Risk-Sensitive Actor-Critic Reinforcement Learning Algotithm", *to appear Proc. 2023 American Control Conference*, San Diego, CA, May 31 - June 2, 2023.

C. N. Mavridis and J.S. Baras, "Multi-Resolution Online Deterministic Annealing: A Hierarchical and Progressive Learning Architecture [under review].

C.N. Mavridis and J.S. Baras, "Towards the One Learning Algorithm Hypothesis: A System-theoretic Approach [under review].

The Institute for Systems Research

UNIVERSITY OF MARYLAND 1856

Thank you!

baras@umd.edu

301-405-6606

https://johnbaras.com/

Questions?