Explainable Learning with Hierarchical Online Deterministic Annealing

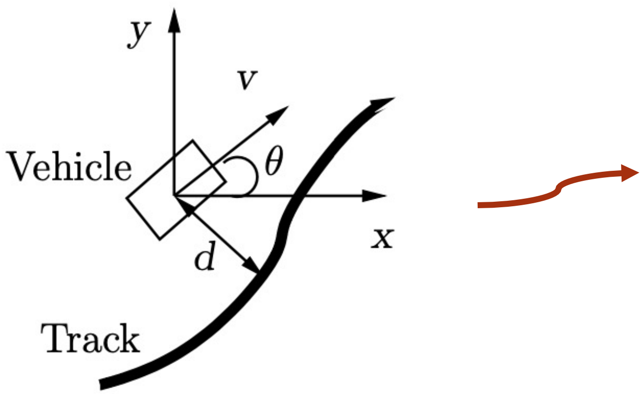
Christos N. Mavridis and John S. Baras

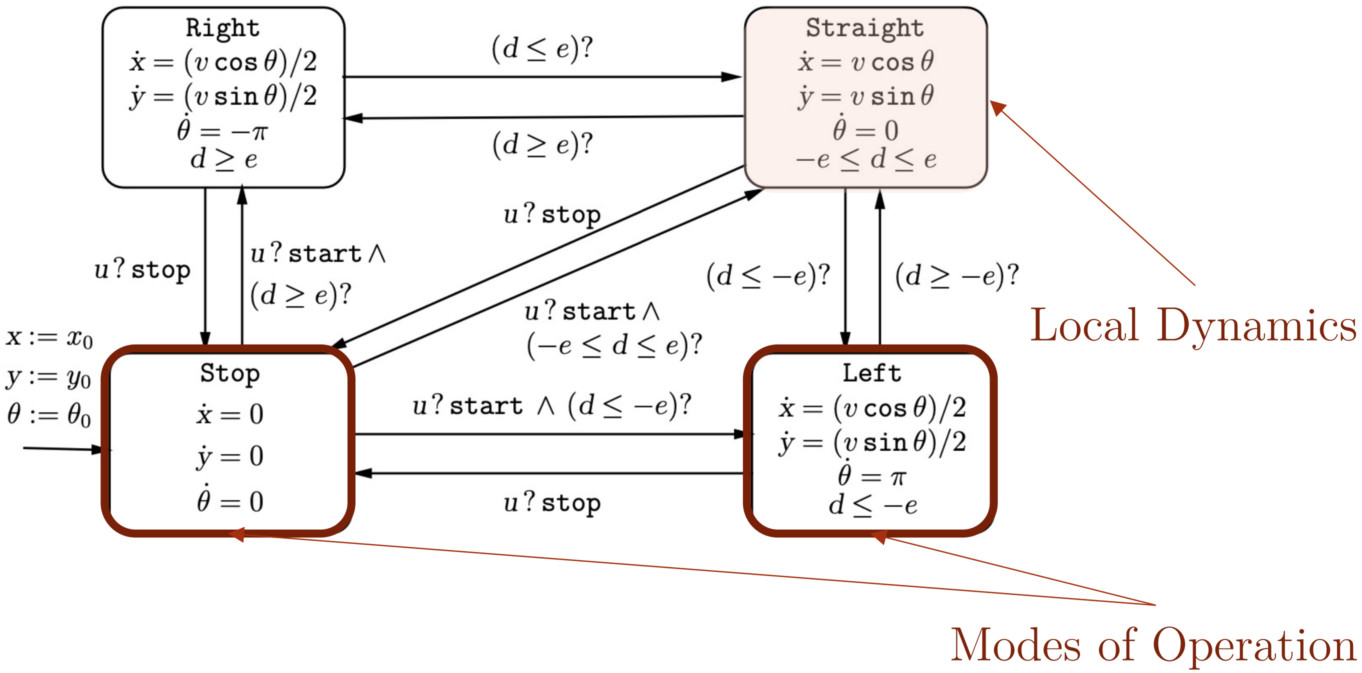
Department of Electrical and Computer Engineering
Institute for Systems Research
University of Maryland

ECML PKDD 2023
Uncertainty meets Explainability

Explainable Learning – The Control-Theoretic Perspective

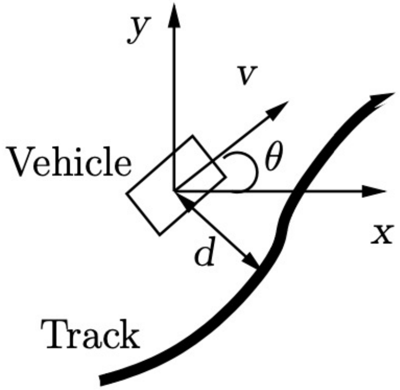
➢ Autonomous Vehicle Control

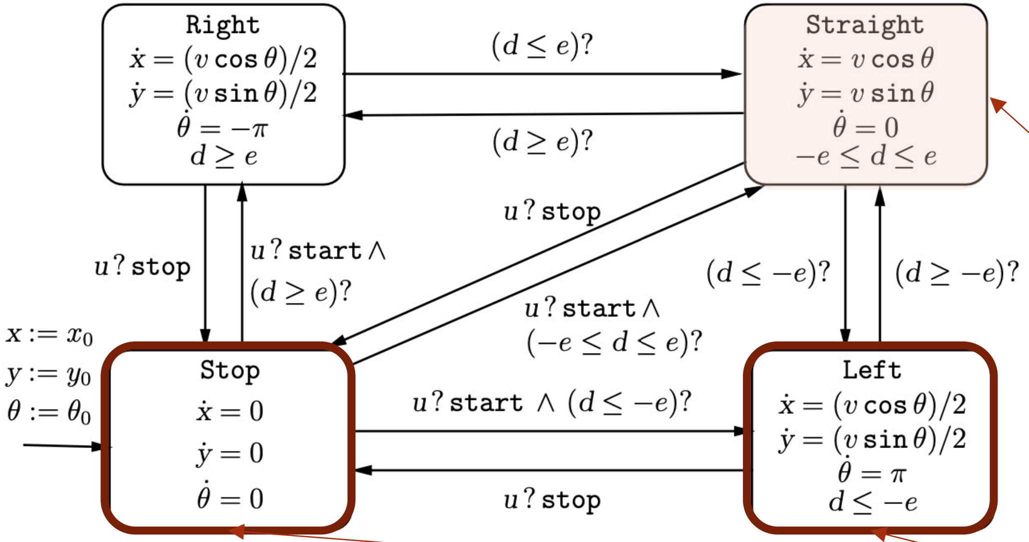




Explainable Learning – The Control-Theoretic Perspective

- **Autonomous Vehicle Control**





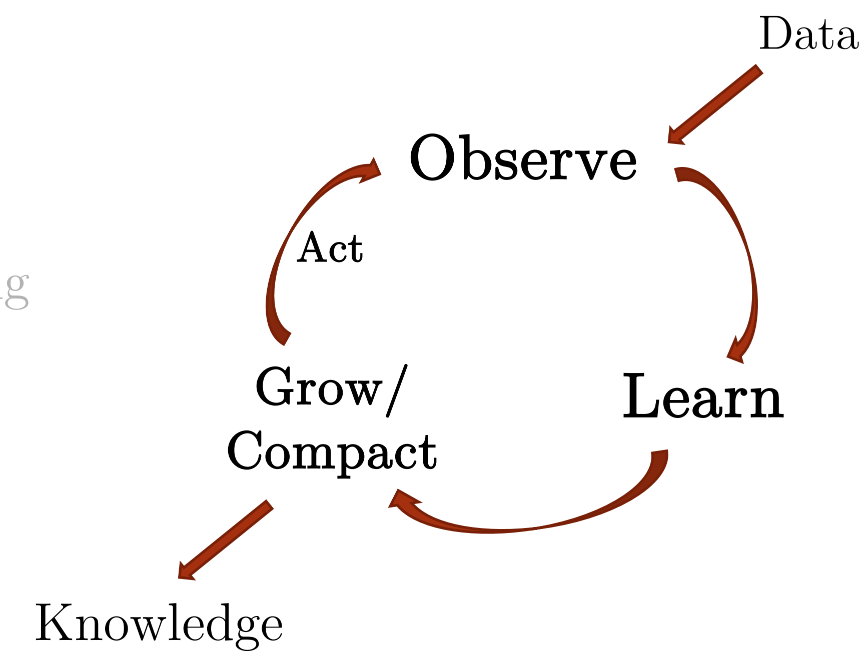
Local Dynamics

Modes of Operation

- **Intelligent Autonomous Systems:**
  - How many modes?
  - Local System Identification?
  - Simultaneous, real-time learning?

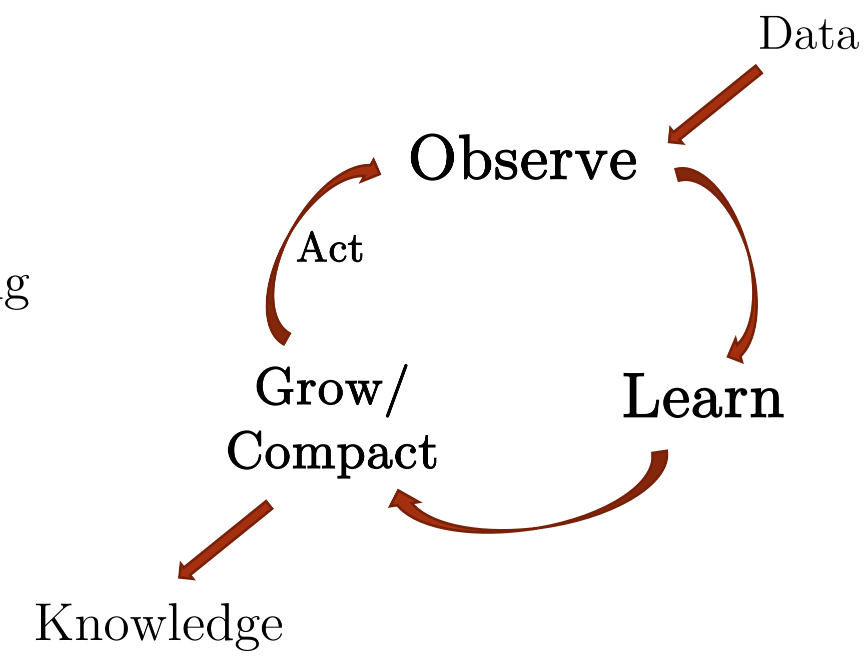
Learning Properties in Cyber-Physical Systems

- 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
  - Hierarchical Learning?



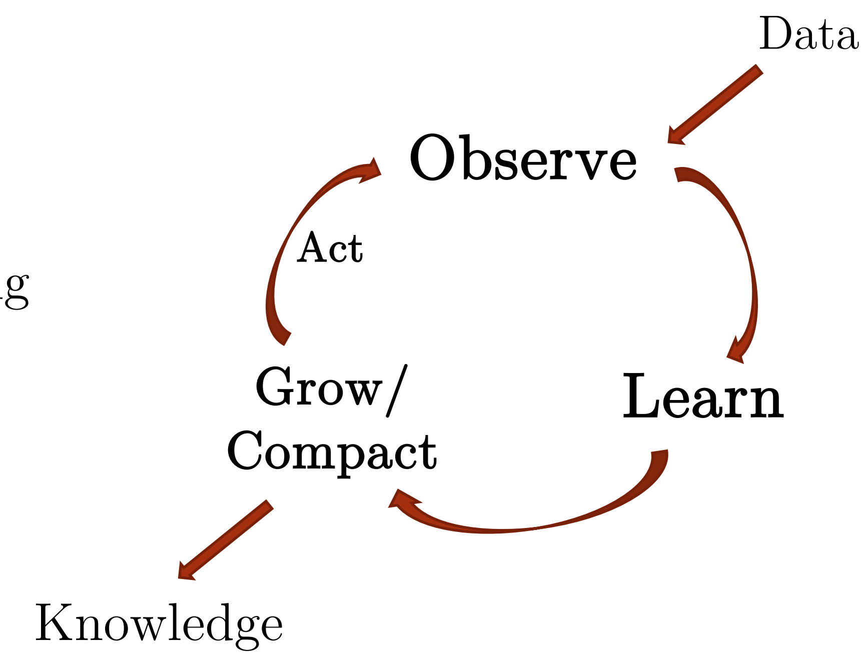
Learning Properties in Cyber-Physical Systems

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



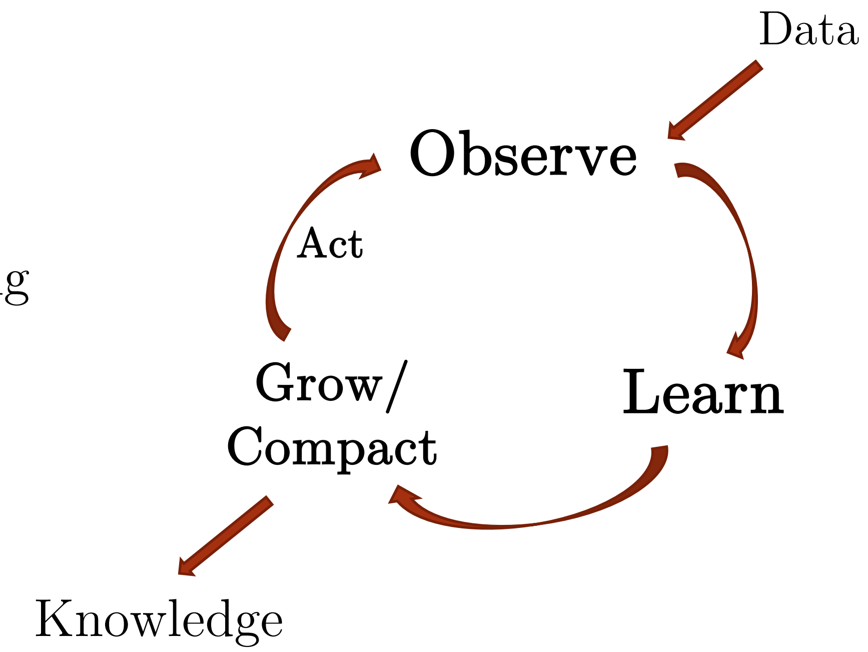
Learning Properties in Cyber-Physical Systems

- **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
  - Hierarchical Learning?



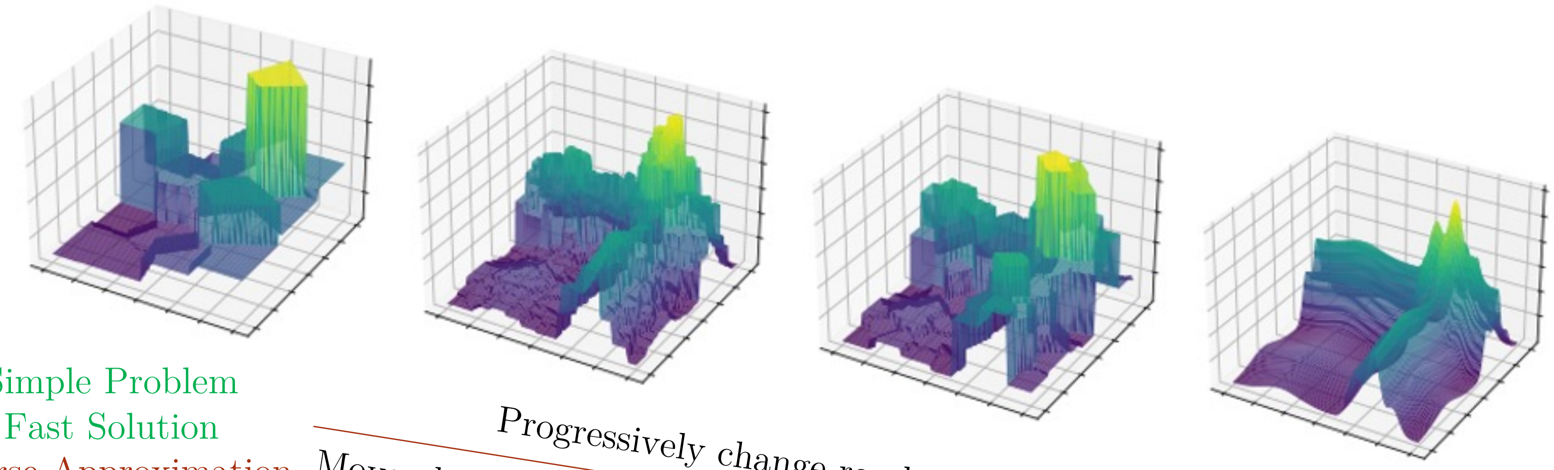
Learning Properties in Cyber-Physical Systems

- **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
  - Hierarchical Learning?



Towards Explainable Hierarchical Learning

- **Goal: Hierarchically Approximate Optimal Solutions***



Simple Problem
Fast Solution
Coarse Approximation

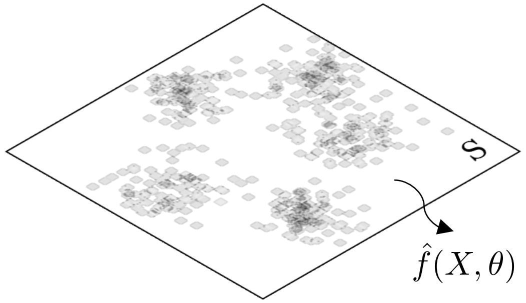
Progressively change resolution
Move along the Performance-Complexity Pareto Curve

Complex Problem
Slow to Solve
Detailed Approximation

* function approximation, reinforcement learning, game policies, system identification, clustering/classification

Towards Explainable Hierarchical Learning (II)

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



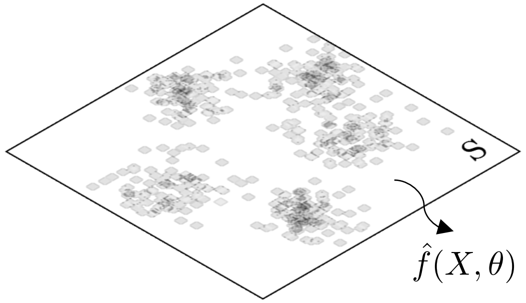
$$\min_{\theta} \mathbb{E}\left[d\left(f(X), \hat{f}(X,\theta)\right)\right]$$

$$y = \hat{f}(x), \;\; x \in S$$

Highly Complex & Non-linear

Towards Explainable Hierarchical Learning (II)

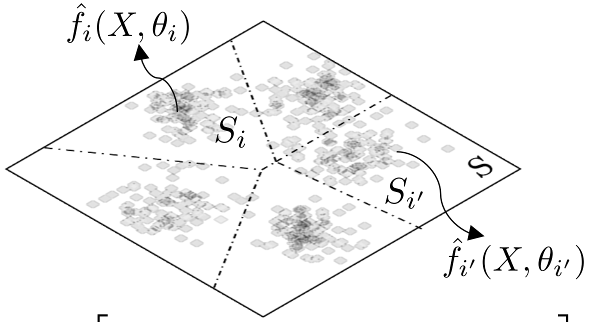
- **Divide and Conquer:** Partition the space and use local models



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

$$y = \hat{f}(x), \ x \in S$$

Highly Complex & Non-linear



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

Simpler local models

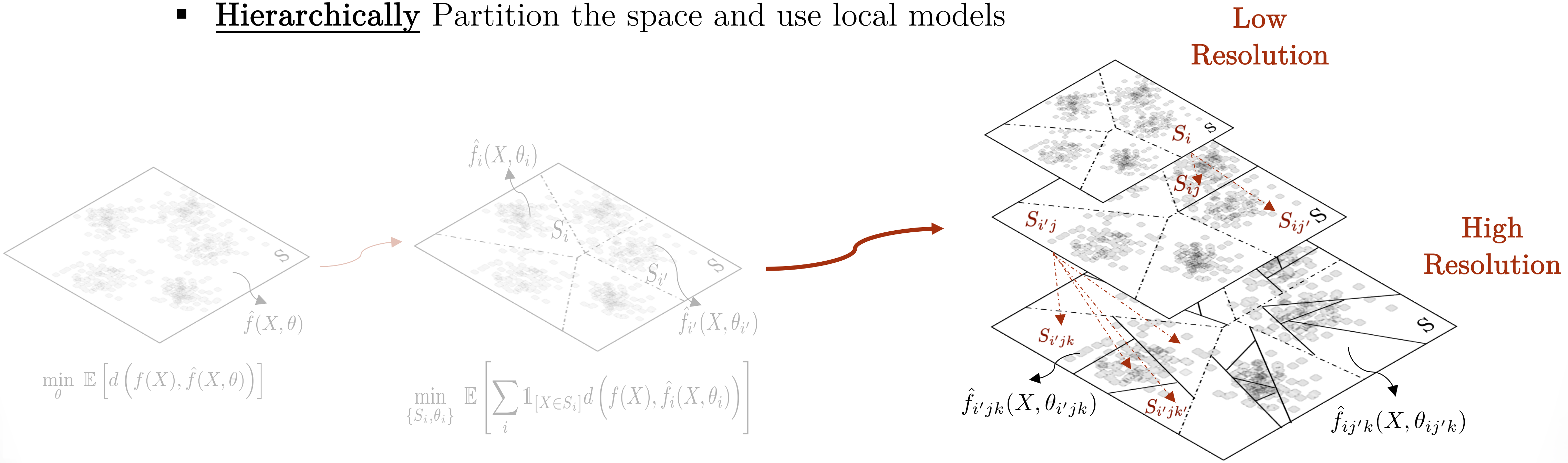
$$y = \begin{cases} \hat{f}_1(x), \ x \in R_1 \\ \hat{f}_2(x), \ x \in R_2 \\ \vdots \\ \hat{f}_n(x), \ x \in R_n \end{cases}$$

Structure = Explainability

Towards Explainable Hierarchical Learning (II)

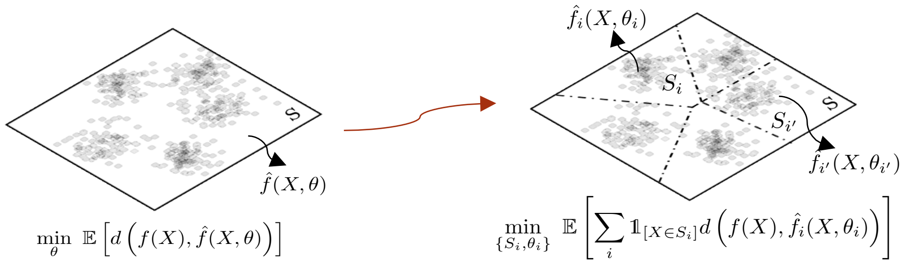
➢ Divide and Conquer

- **Hierarchically** Partition the space and use local models



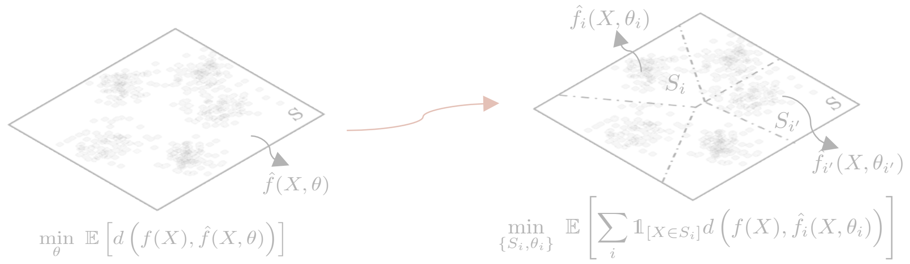
Towards Explainable Hierarchical Learning (III)

➤ **Problems with Simultaneous Partitioning and Local Learning?**



Towards Explainable Hierarchical Learning (III)

➢ Problems with Simultaneous Partitioning and Local Learning?



➢ Problems:

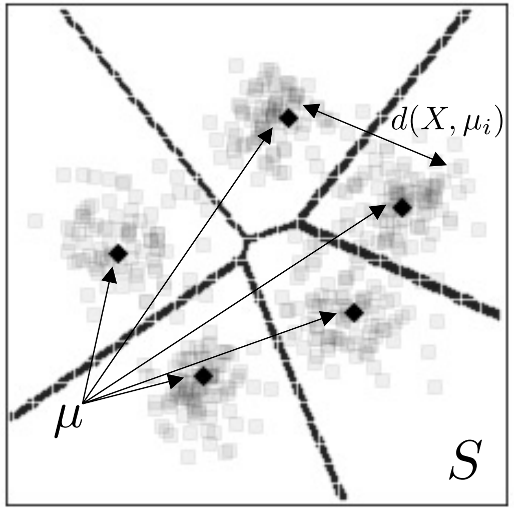
- **How many regions?**
  - Start with few and add as needed?
- **Optimal parameters?**
  - Local minima? Gradients?
  - Robustness?
- **Simultaneously learn local models?**

→ Online Deterministic Annealing

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.* $Q \in S$

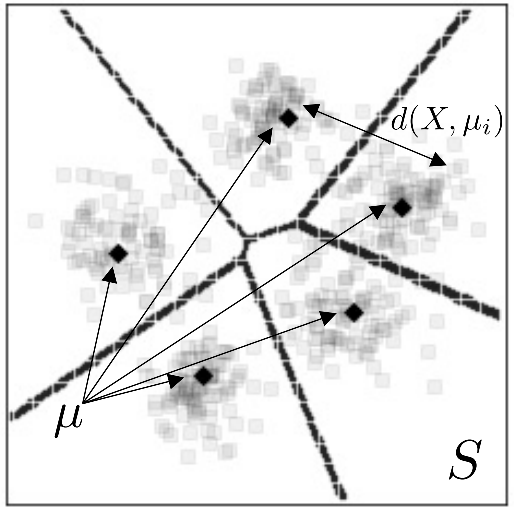
  defined by: $p(\mu_i|x) = \mathbb{P}\left[Q = \mu_i | X = x\right]$
- *Dissimilarity:* $d : S \times S \to [0, \infty)$



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.* $Q \in S$

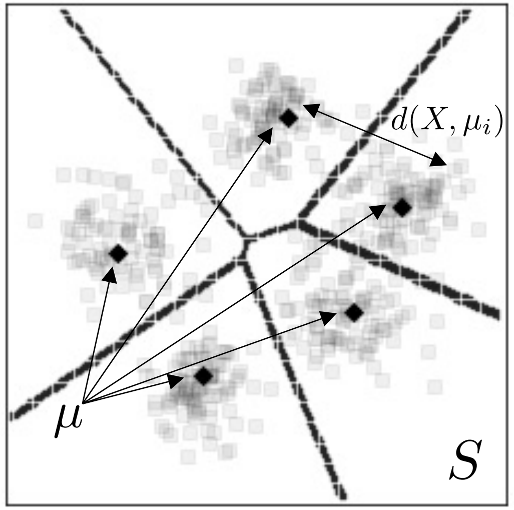
  defined by: $p(\mu_i|x) = \mathbb{P}\left[Q = \mu_i | X = x\right]$
- *Dissimilarity:* $d : S \times S \to [0, \infty)$



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.* $Q \in S$

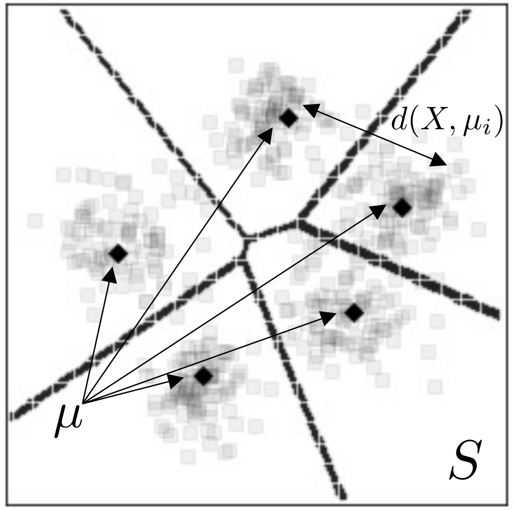
  defined by: $p(\mu_i|x) = \mathbb{P}\left[Q = \mu_i | X = x\right]$
- *Dissimilarity:* $d : S \times S \to [0, \infty)$



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:* $d : S \times S \to [0, \infty)$



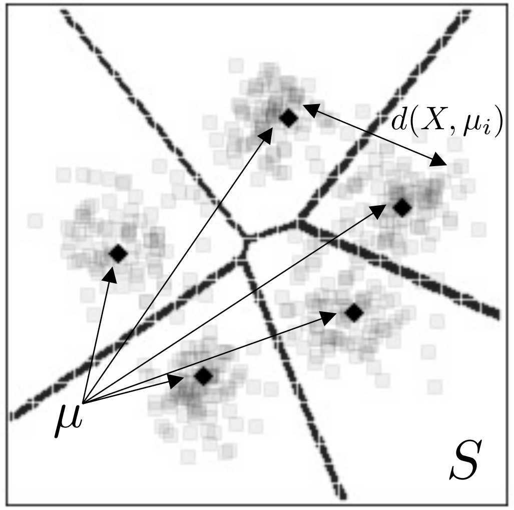
Clustering?

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

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:* $d : S \times S \to [0, \infty)$



Clustering?

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

Online Deterministic Annealing

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

Adaptive
Robust
Progressive

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

Why Maximum Entropy?

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

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

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

Risk-Sensitivity

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

- **Robustness** w.r.t. initial conditions, input perturbations.
- **Bifurcation:** Progressively grow set of models

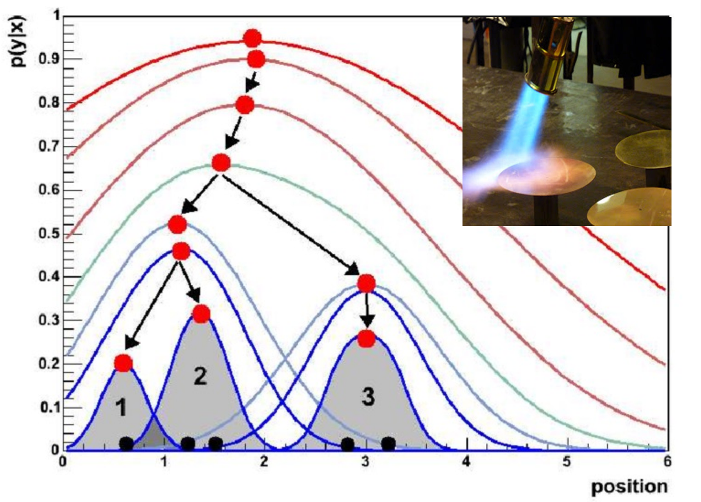
*Mavridis et al., Risk Sensitivity and Entropy Regularization in Prototype-based Learning, IEEE MED 2022.

Online Deterministic Annealing

<u>Online Deterministic Annealing</u>

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

$$\begin{cases} D(X,Q) & : \text{Distortion} \\ H(X,Q) & : \text{Entropy} \end{cases}$$

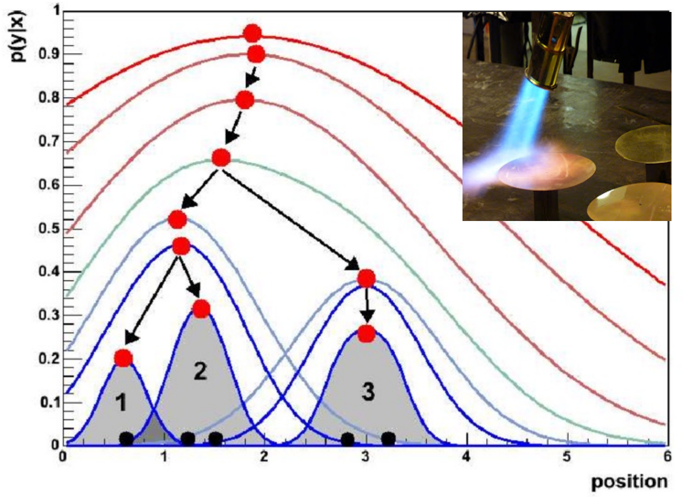


Online Deterministic Annealing

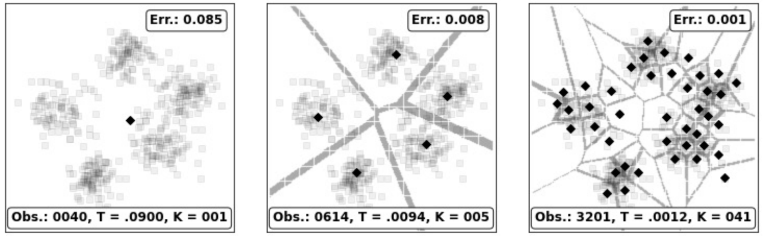
Online Deterministic Annealing

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

$\begin{cases} D(X,Q) & : \text{Distortion} \\ H(X,Q) & : \text{Entropy} \end{cases}$



- **Lagrange (Temperature) Coefficient** T
  - ➢ Controls Performance/Complexity Tradeoff
  - ➢ Simulates Annealing Optimization (Temperature)
  - ➢ Stochastic Approximation
    - Simultaneous local system identification
  - ➢ Triggers Bifurcation
    - Progressively adjust number of regions/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*
  $p^*(\mu_i|x) = \frac{e^{-\frac{d(x,\mu_i)}{T}}}{\sum_j e^{-\frac{d(x,\mu_j)}{T}}}, \quad \forall x \in S.$

- **Theorem.** *The solution to* $\min_\mu F^*(\mu)$ *is given by*

$$\mu_i^* = \mathbb{E}\left[X|\mu_i\right] = \frac{\int x p(x) p^*(\mu_i|x)\ dx}{p^*(\mu_i)}$$

centroid form

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 dynamic stochastic process created by the recursive updates*

$$\mu_i(n+1) = \frac{\beta(n)}{\rho_i(n)}\left[\frac{\sigma_i(n+1)}{\rho_i(n+1)}(\rho_i(n) - \hat{p}(\mu_i|x_n)) + (x_n\hat{p}(\mu_i|x_n) - \sigma_i(n)\right]$$

where the quantities $\rho_i(n)$, $\sigma_i(n)$, and $\hat{p}(\mu_i|x_n)$ are recursively updated by:

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

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

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

$$\mu_i(n) = \frac{\sigma_i(n) \to \mathbb{E}\left[\mathbb{1}_{[\mu]}X\right]}{\rho_i(n) \to \mathbb{P}[\mu]}$$

Stochastic Approximation: Gradient-Free !

Online Deterministic Annealing (IV)

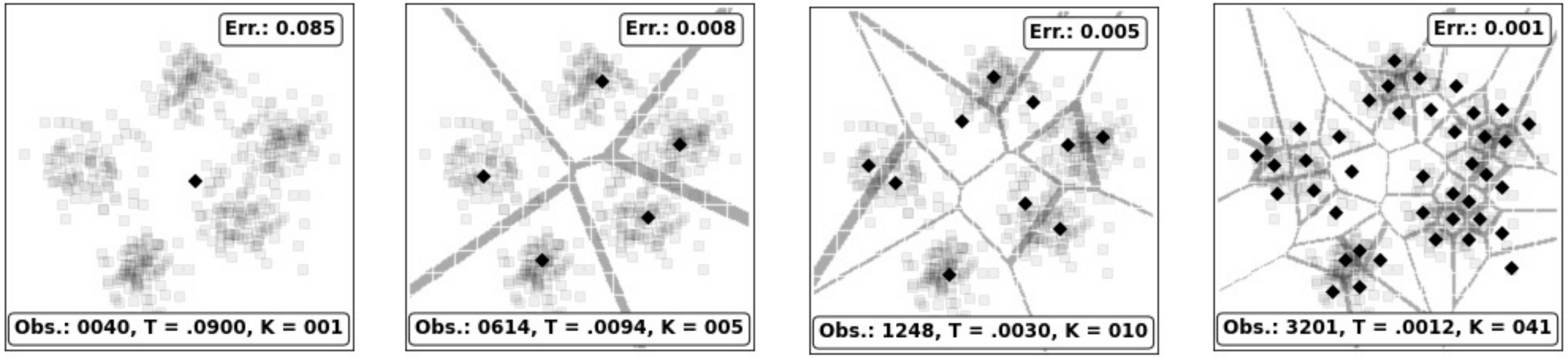
Bifurcation and the number of codevectors

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

  $\cdots$

  $\min F_{T_0} := D - T_0 H$, $T_i < T_{i+1}$: Decreasing Temperature

- **Remark.** *As $T \to \infty$, we get $\mu_i = \mathbb{E}[f(X)]$, $\forall i$, 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.*



Performance-Complexity Trade-off

Online Deterministic Annealing (V)

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:

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

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

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

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

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

$$\mu_{t+1} = \mu_t + \beta(t)\left[g(\theta_t, \mu_t) + M_{t+1}^{(\mu)}\right]$$

Slow SA

Partition

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:

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

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

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

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

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

$$\mu_{t+1} = \mu_t + \beta(t) \left[g(\theta_t, \mu_t) + M_{t+1}^{(\mu)}\right]$$

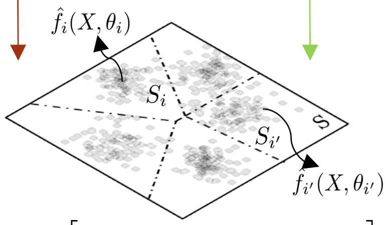
$$\theta_{t+1} = \theta_t + \alpha(t) \left[f(\theta_t, \mu_t) + M_{t+1}^{(\theta)}\right]$$

Slow SA → Partition

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

Fast SA → Local models

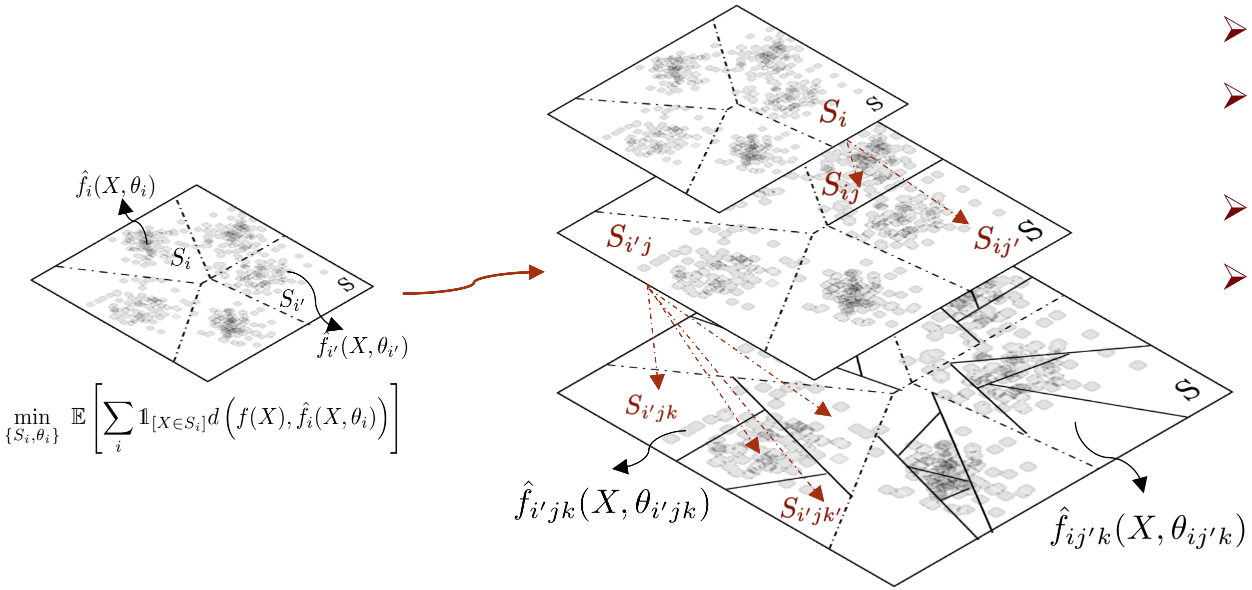
Fast SA: Gradient Descent, Q-Learning, ...



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

Hierarchical Online Deterministic Annealing

Tree-Structured Hierarchical Learning

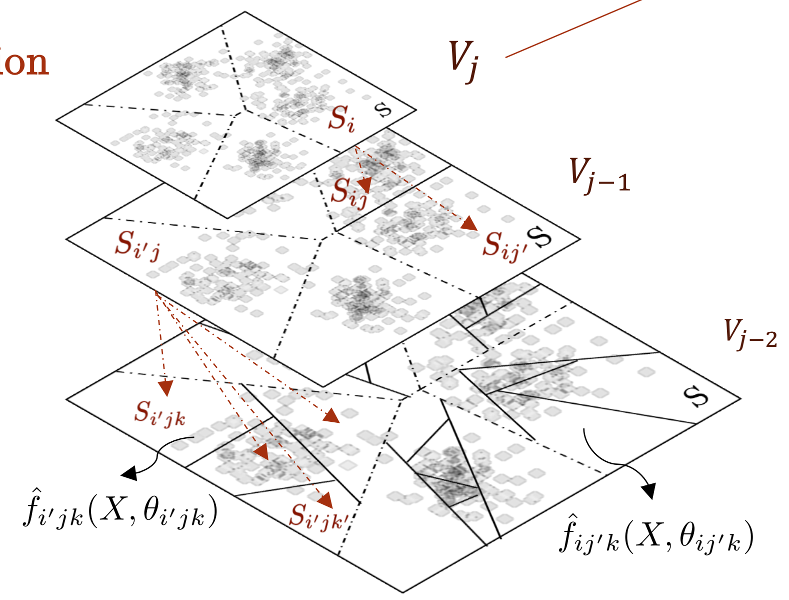


- Constructive (Structured Representation)
- Provably Consistent
- Localization
  - Emphasis on regions with high error
- Asynchronous/Parallel Computation
- Reduced Complexity $O\left(\frac{k^{\tilde{l}}-1}{k(k-1)} N_c (2\bar{k})^2 d\right)$

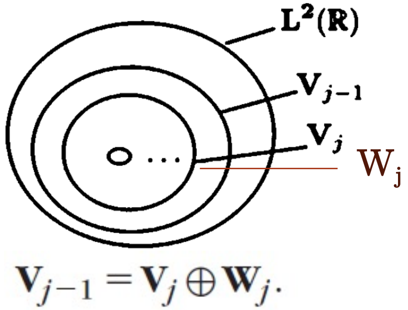
$$\bar{k} = \sum_{n=0}^{1/\tilde{l}\log_2 K_{max}} 2^n$$

Hierarchical Online Deterministic Annealing

Multi-Resolution Hierarchical Learning



Example: Group-convolutional Wavelets

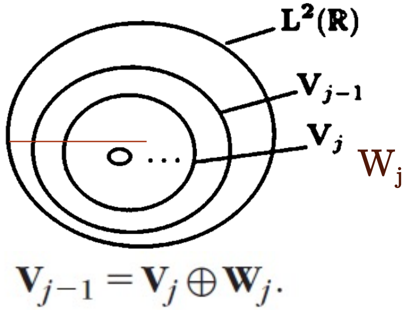


- Constructive (Structured Representation)
- Provably Consistent
- Localization
  - Emphasis on regions with high error
- Asynchronous/Parallel Computation
- Reduced Complexity $O\left(\frac{k^{\bar{l}}-1}{k(k-1)} N_c (2\bar{k})^2 d)\right)$

$$\bar{k} = \sum_{n=0}^{1/\bar{l} \log_2 K_{max}} 2^n$$

Group-Convolutional Wavelets

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

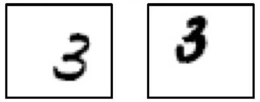


- **Convolution on Groups**

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

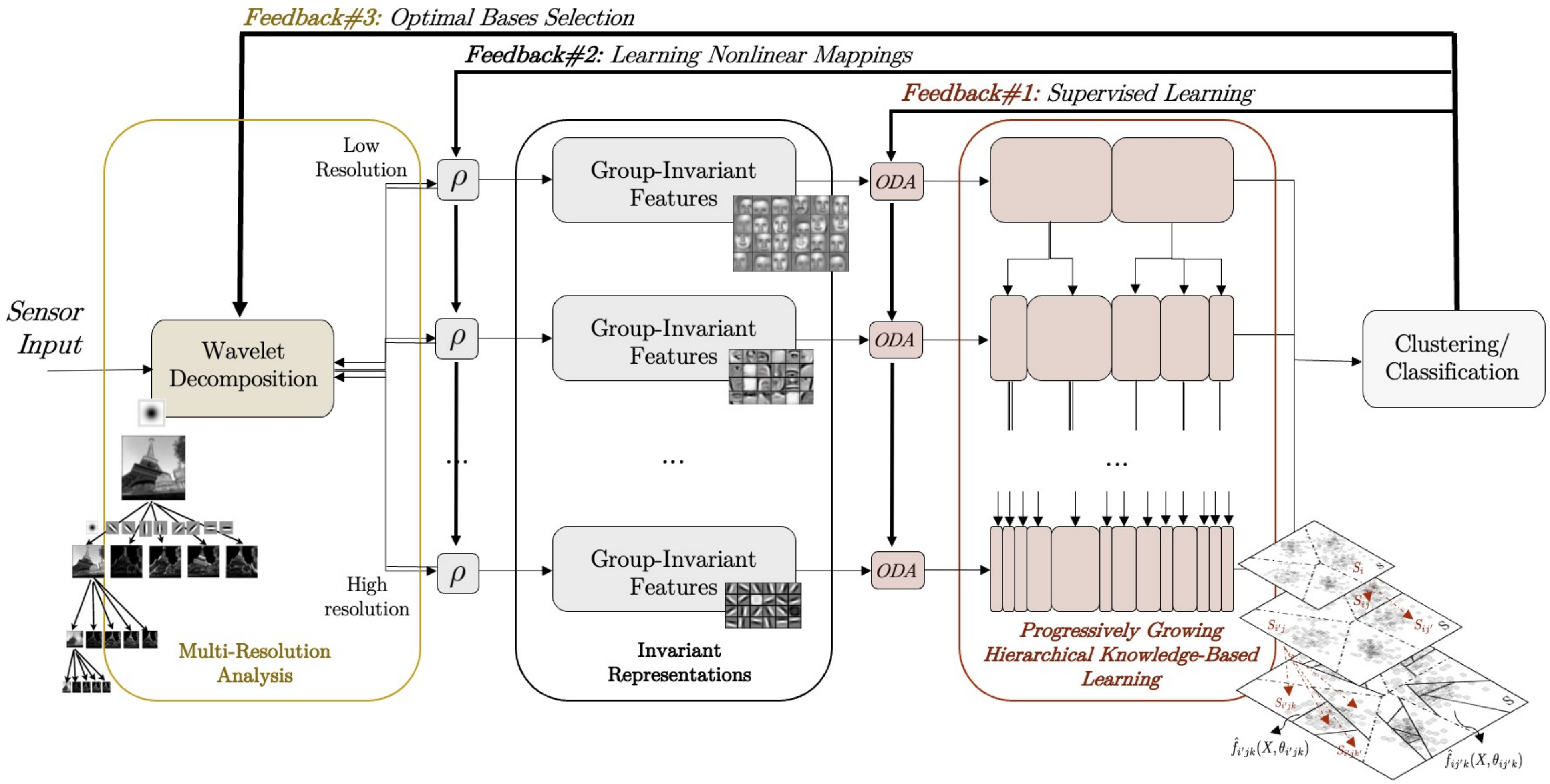
where for a Lie Group G: $g \in G \to g.f(x) := f(g^{-1}x)$

- **Locally Group-Invariant Representations**

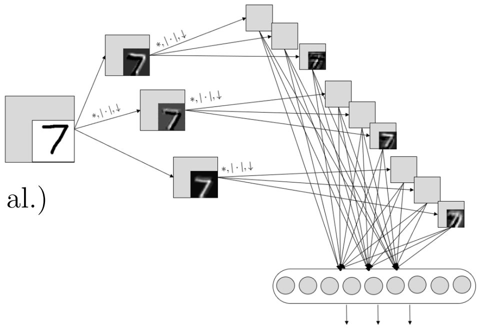


Repeat
- Build group-covariant representations (**wavelets**)
- Make them locally invariant (**non-linearity** + **averaging**)

Closed-Loop Hierarchical Learning Architecture

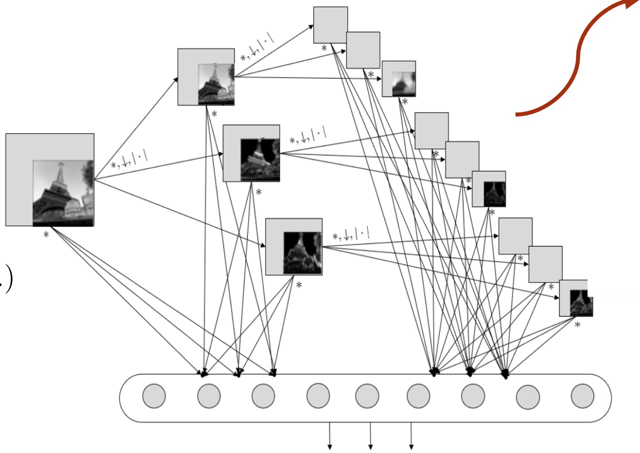


A Deep Learning Architecture



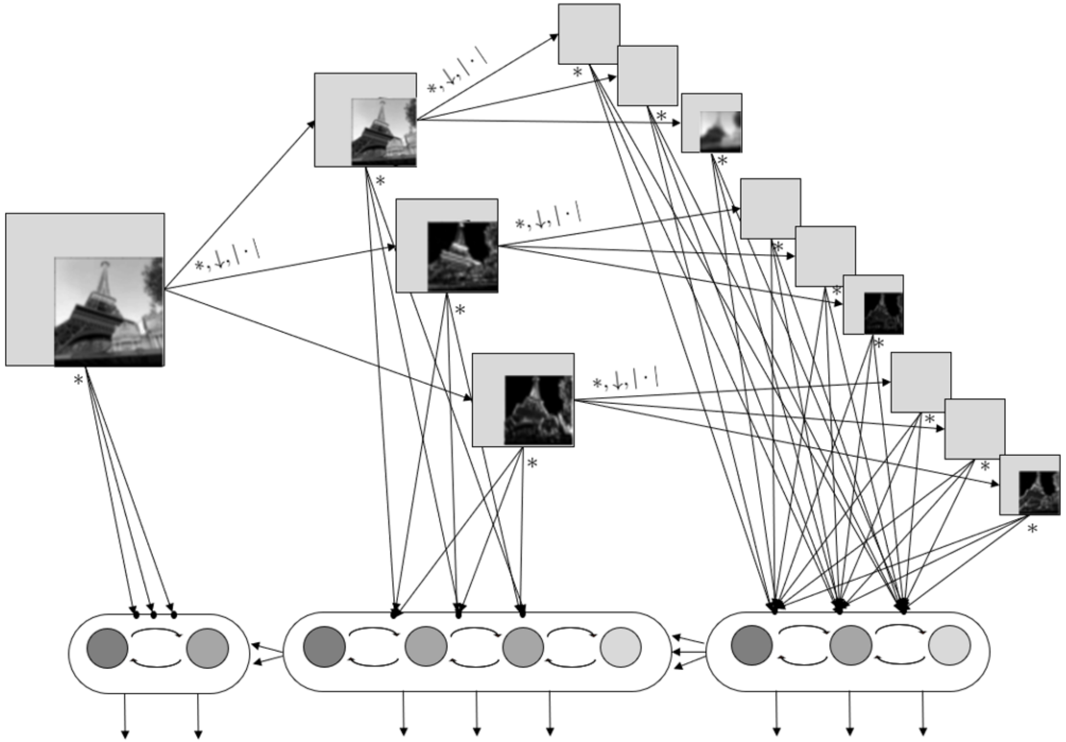
(Lecun et al.)

Deep Convolutional Network



(Mallat et al.)

Scattering Convolutional Network



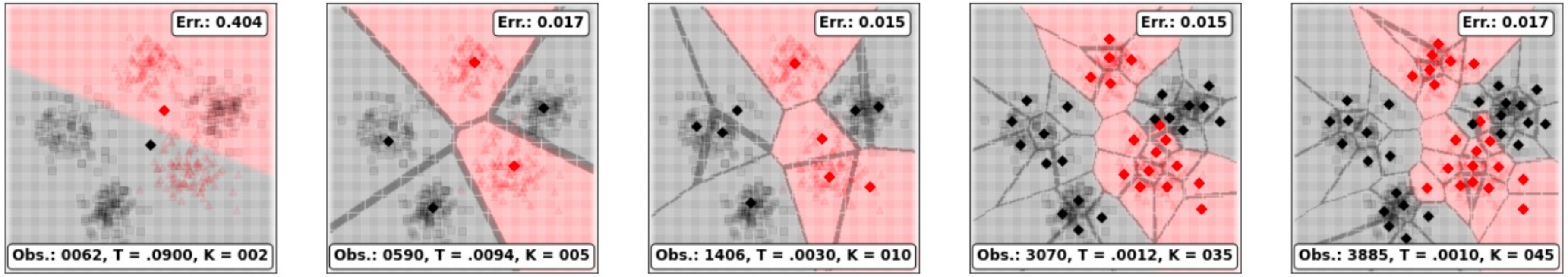
Our Approach

$*, \downarrow, |\cdot|$ →

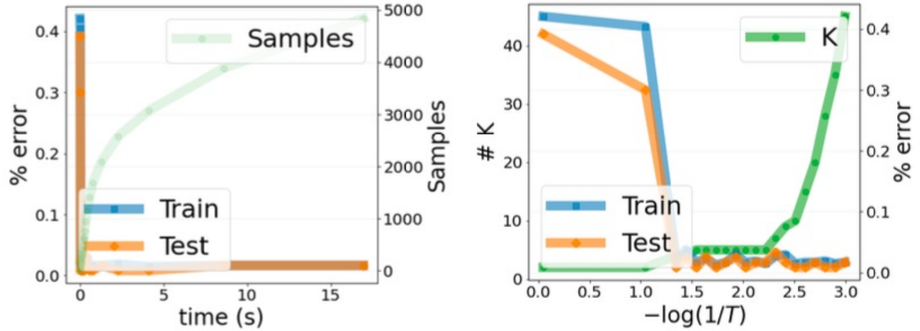
Simulation Results

➢ **Single Resolution.** Binary Classification on Mixture of Gaussians.

Performance-Complexity Trade-off



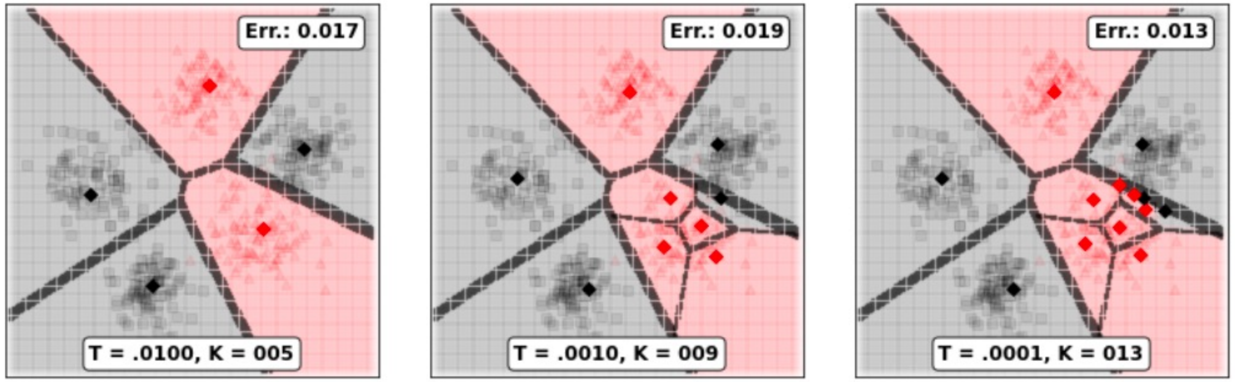
(a) Evolution of the algorithm in the data space.



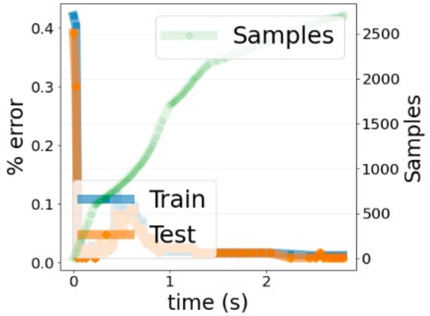
(b) Performance curves.

Simulation Results (II)

- **Single Resolution – Tree-Structured.** Binary Classification on Mixture of Gaussians.



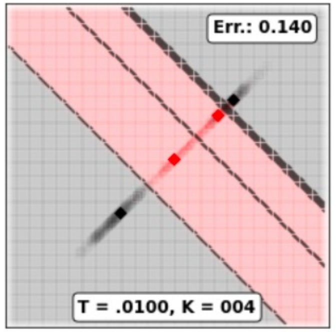
(a) Evolution of the algorithm in the data space.



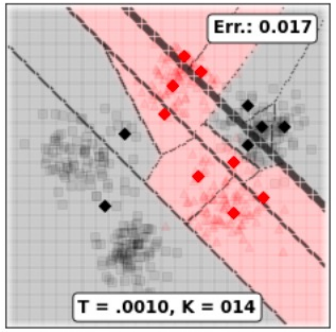
(b) Performance curves.

Simulation Results (III)

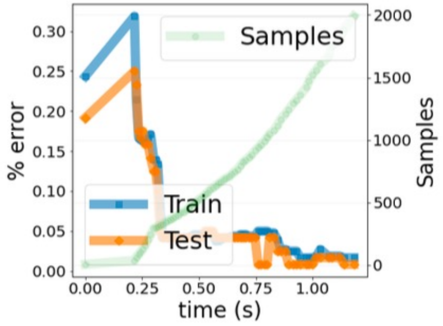
➢ **Multiple Resolutions w/ PCA.** Binary Classification on Mixture of Gaussians.



(a) Convergence of first layer with low-resolution features.



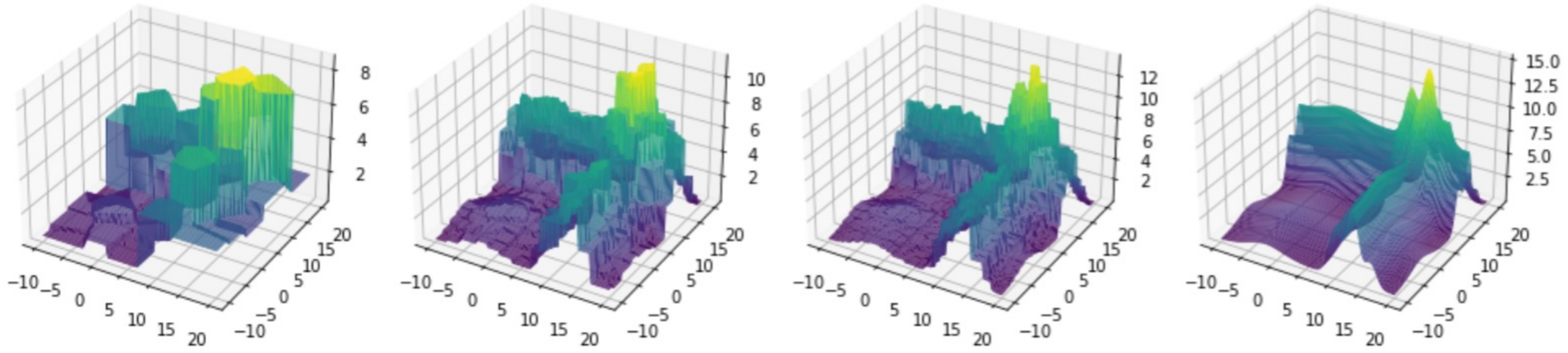
(b) Convergence of second layer with high-resolution features.



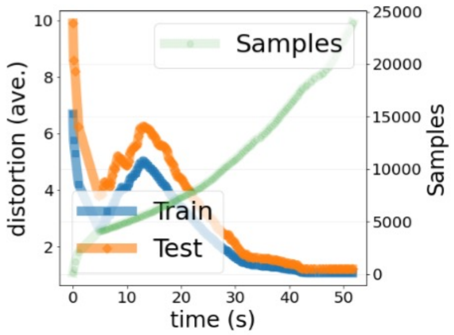
(c) Performance curve.

Simulation Results (IV)

➢ **Single Resolution.** 2D Regression with Constant Local Models.



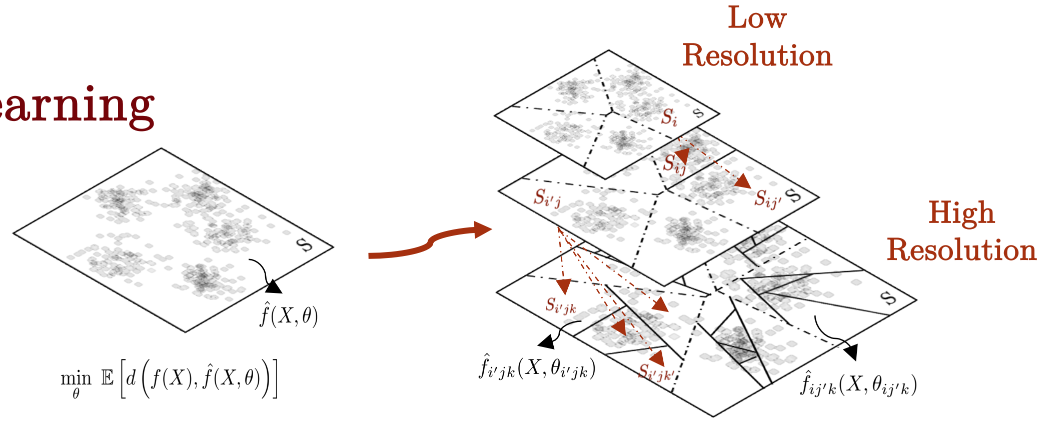
(a) Evolution of the algorithm in the data space (original function on the right).



(b) Performance curves.

Thank you!

- **Simultaneous Partitioning and Local Learning**
  - Explainability
  - Robustness w.r.t. Init. & Noise



- **Hierarchical Online Deterministic Annealing**
  - Multi-Resolution Partitioning
  - Online, Adaptive, Gradient-Free
  - Simultaneous local model training

Questions?

Christos Mavridis

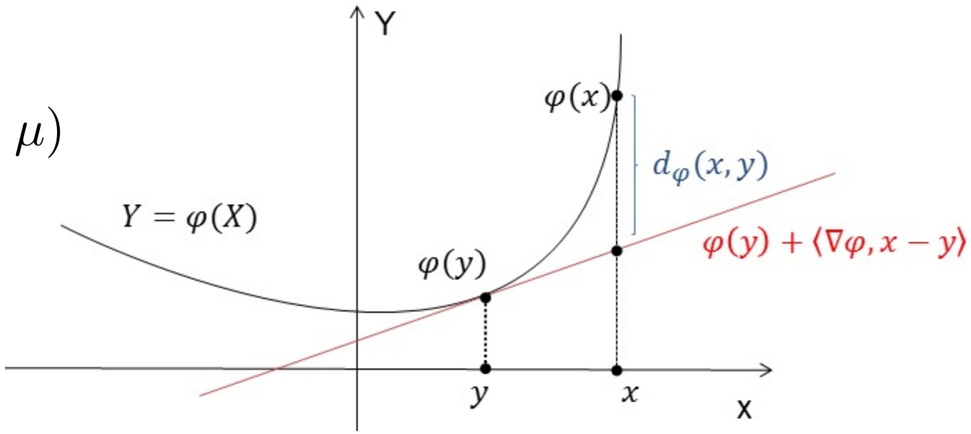
mavridis@umd.edu

mavridis@kth.se

https://mavridischristos.github.io/

Bregman Divergences

- $$d_\phi(x,\mu) = \phi(x) - \phi(\mu) - \frac{\partial \phi}{\partial \mu}(\mu)(x-\mu)$$
  - **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.*

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

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:

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

□

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