Representation Policy Iteration:
A Unified Framework for Learning Behavior and Representation

Sridhar Mahadevan

Unified Theories of Cognition

(Newell, William James Lectures, Harvard)

TIME SCALE OF HUMAN ACTION

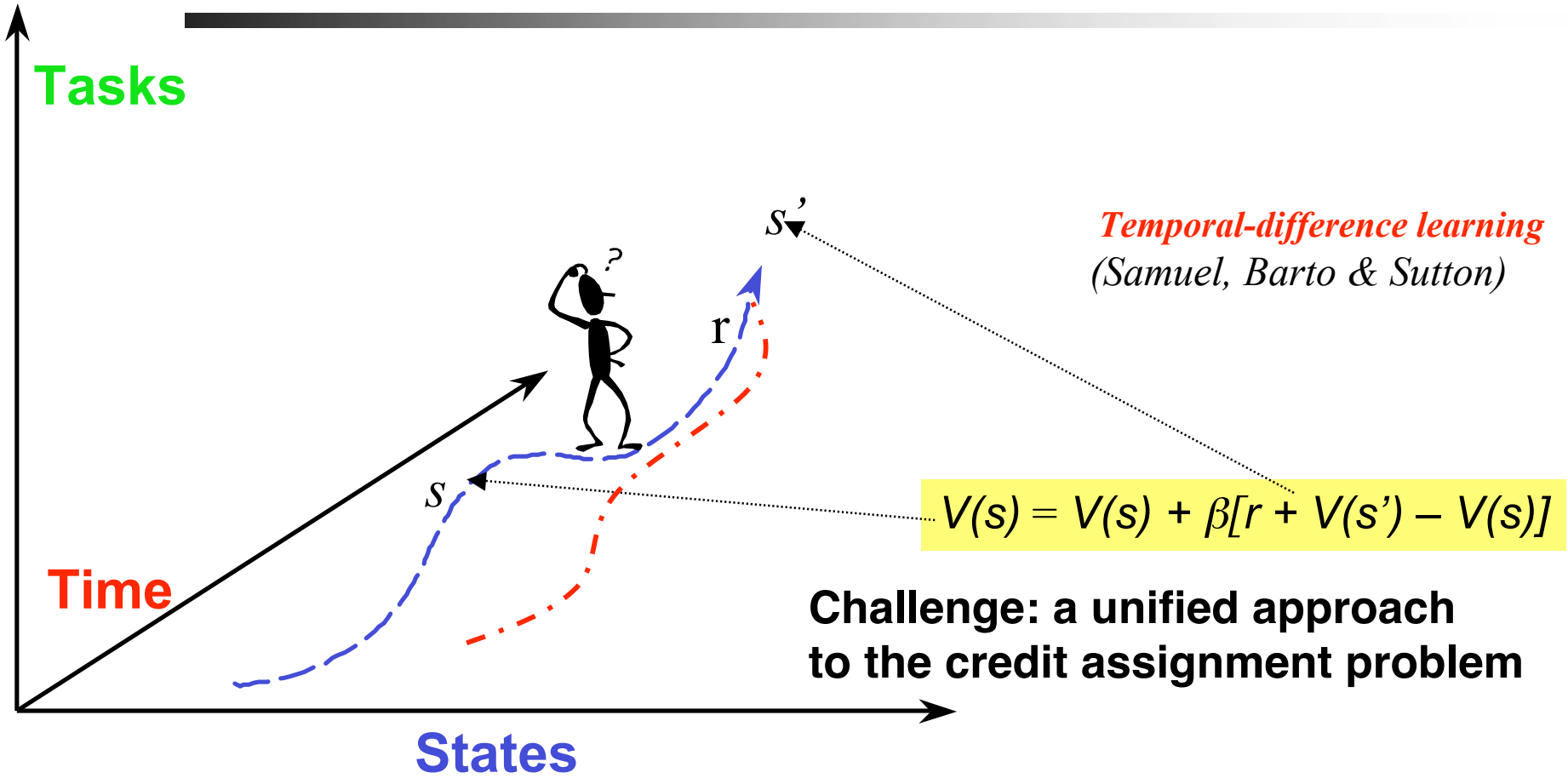
Scale (sec)	Time Units	System	World (theory)
10^{7}	months		
10^{6}	weeks		SOCIAL BAND
10^{5}	days		
10^{4}	hours	Task	
10^{3}	10 min	Task	RATIONAL BAND
10^{2}	minutes	Task	
10^{1}	10 sec	Unit task	
10^{0}	1 sec	Operations	COGNITIVE BAND
10^{-1}	100 ms	Deliberate act	
10^{-2}	10 ms	Neural circuit	
10^{-3}	1 ms	Neuron	BIOLOGICAL BAND
10^{-4}	100 µs	Organelle	

How are we able to reason about the world across such a large dynamic range of time-scales?

How can AI systems automatically construct temporal abstractions given the uncertain nature of actions?

Credit Assignment Problem

(Minsky, Steps Toward AI, 1960)



Temporal-difference learning
(Samuel, Barto & Sutton)

$$V(s) = V(s) + \beta[r + V(s') - V(s)]$$

Challenge: a unified approach to the credit assignment problem

Learning to Play Checkers

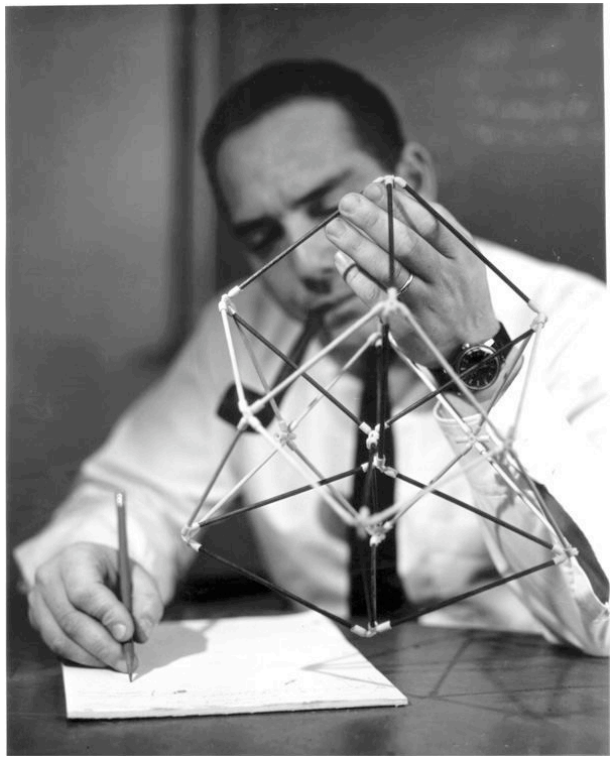
(Arthur Samuel, 1950s)

- ``*Samuel completed the first checkers program on an IBM 701, and when it was about to be demonstrated, Thomas J. Watson Sr., the founder and President of IBM, remarked that the demonstration would raise the price of IBM stock 15 points. It did.''* -- John McCarthy

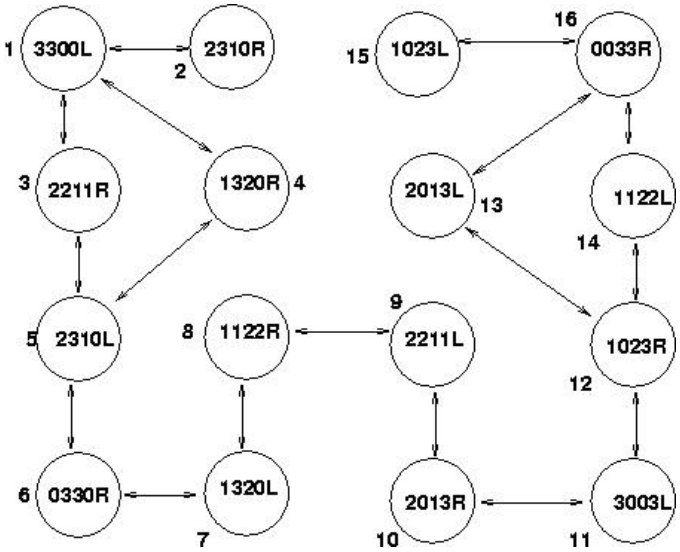
- Samuel pioneered value function approximation using a fixed polynomial architecture

Representation Learning by Global State Space Analysis

(Saul Amarel, 1960s)



Missionaries and Cannibal



Find symmetries and bottlenecks

Transfer of Learning: From Task-Specific to Task-Independent Value Functions

Task-Independent Learning

Task-Specific Learning

Bottlenecks, symmetries

Pseudo-rewards, shaping

Global Rewards

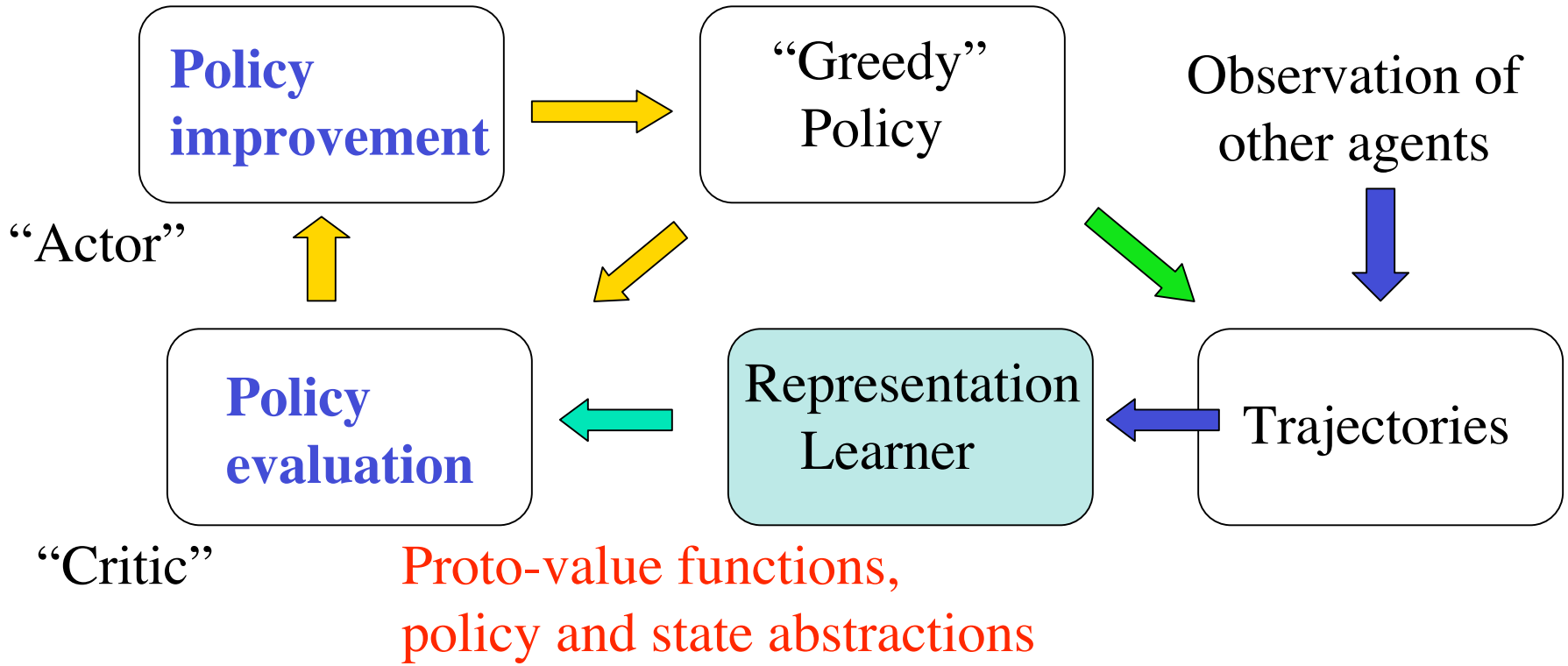
Proto-value functions
(Mahadevan, 2005)

Local value functions
(Dietterich, 2000)

Global value functions
(TD-Gammon, Tesauro, 1992)

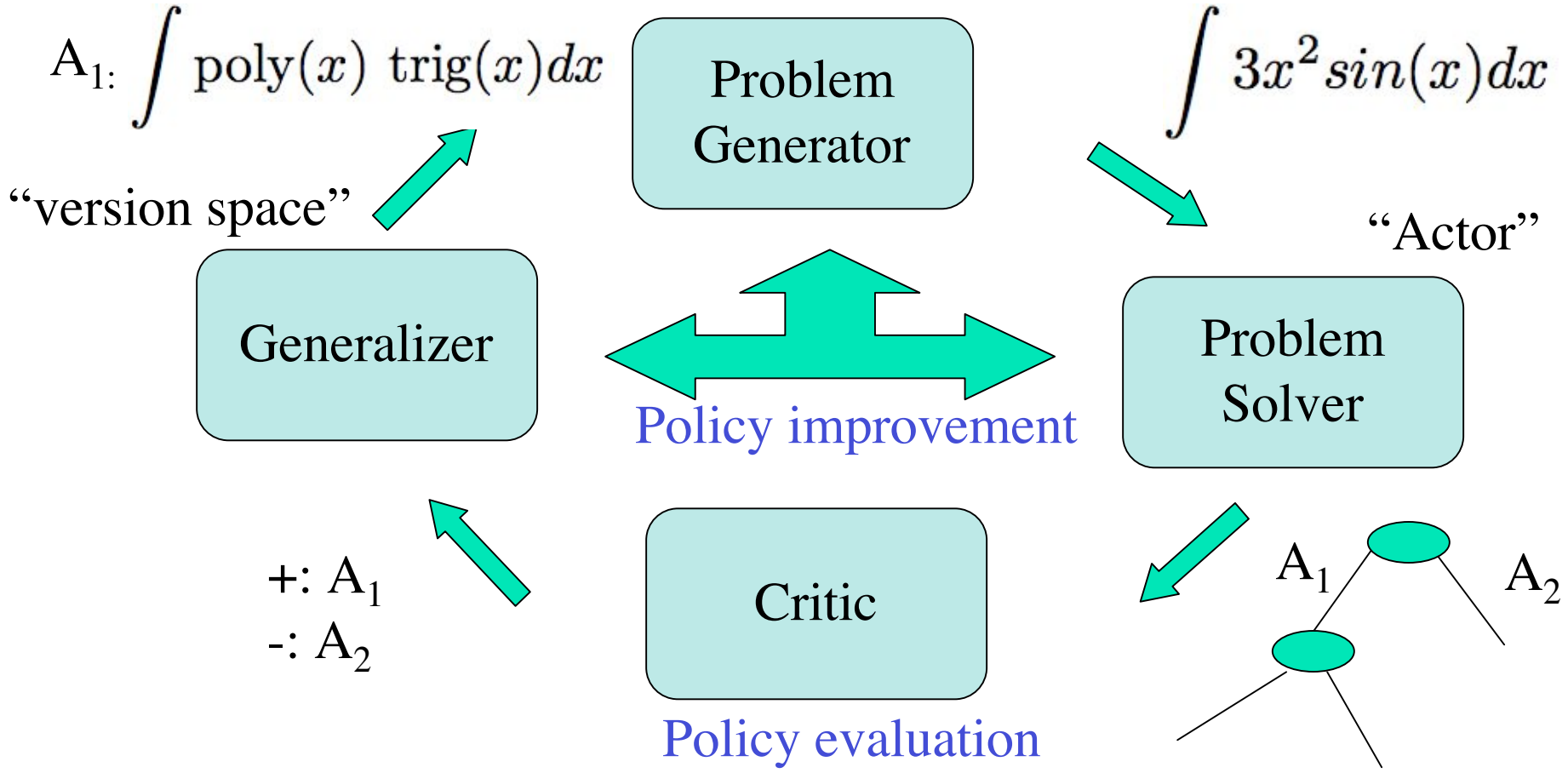
Diffusion wavelets
(Coifman and Maggioni, 2005)

Representation Policy Iteration

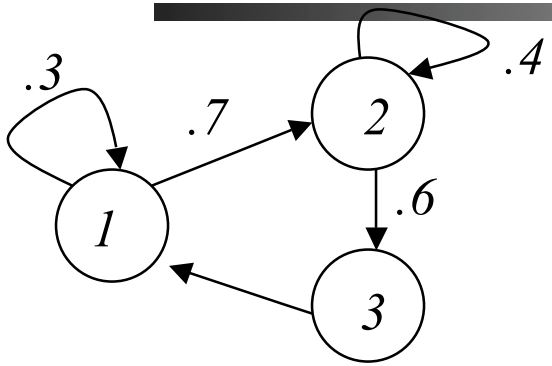


LEX: Learning Symbolic Integration

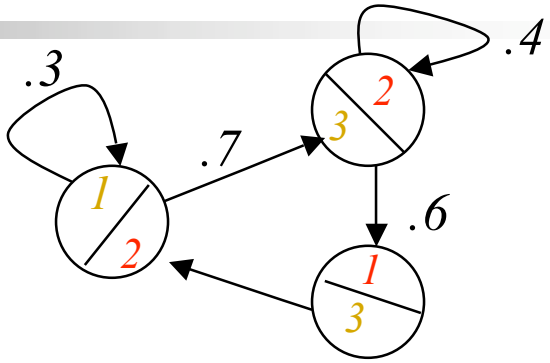
(Mitchell, Utgoff, and Banerji, 1983)



Probabilistic Finite State Models

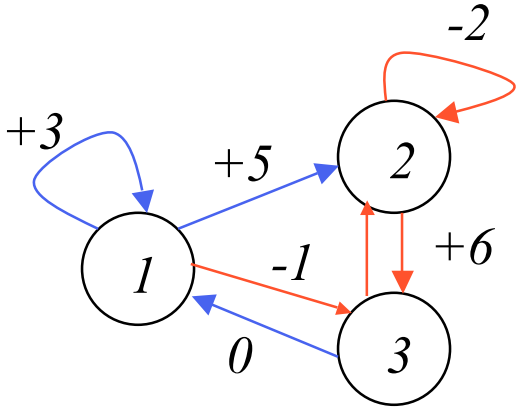


Markov Chain

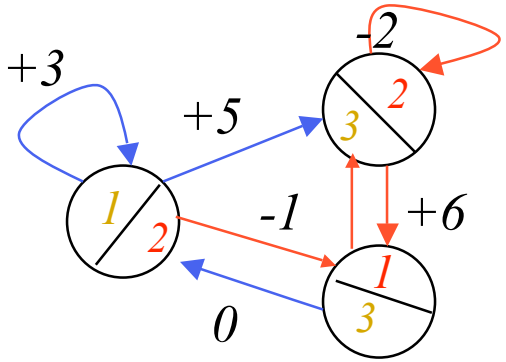


Hidden Markov Model

How to extend these models to capture temporal/spatial abstractions?

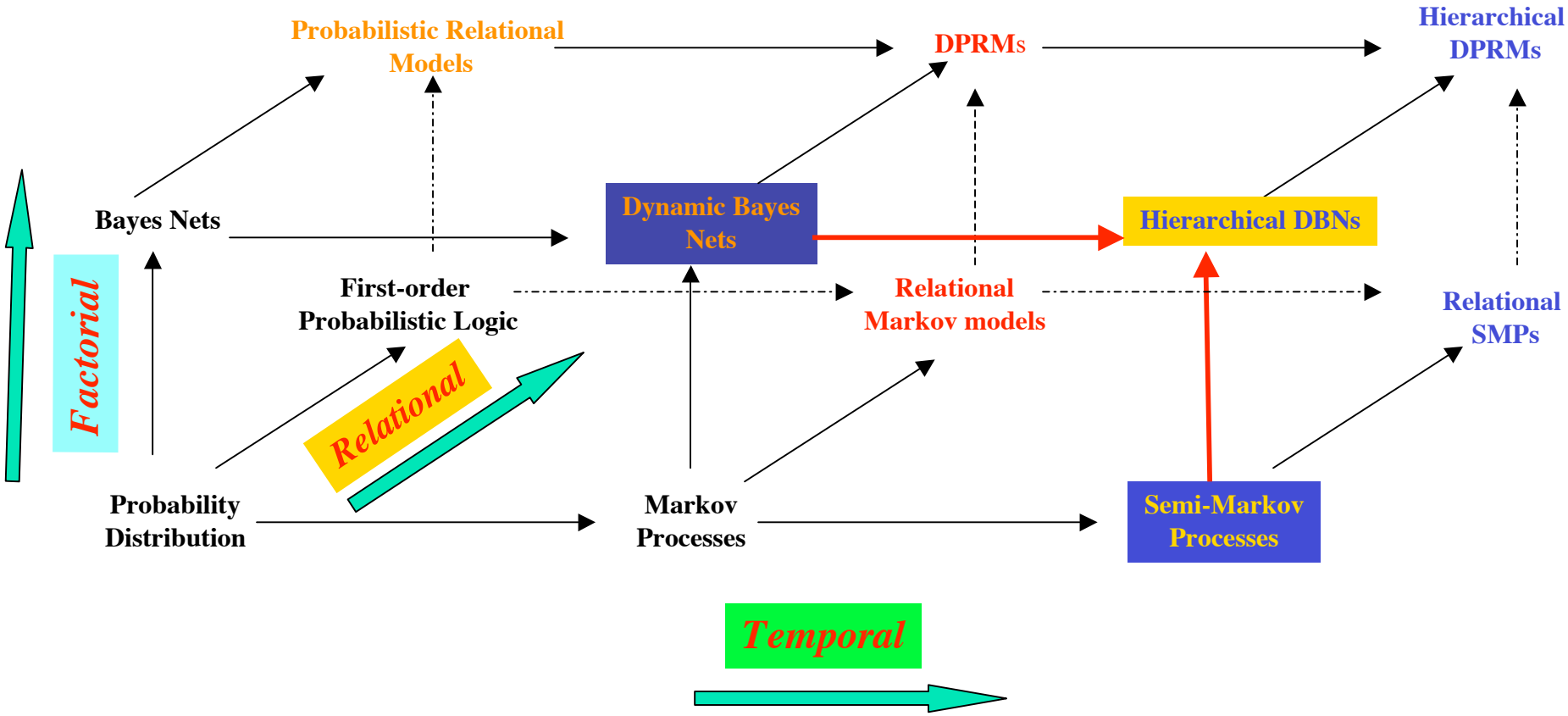


Markov Decision Process



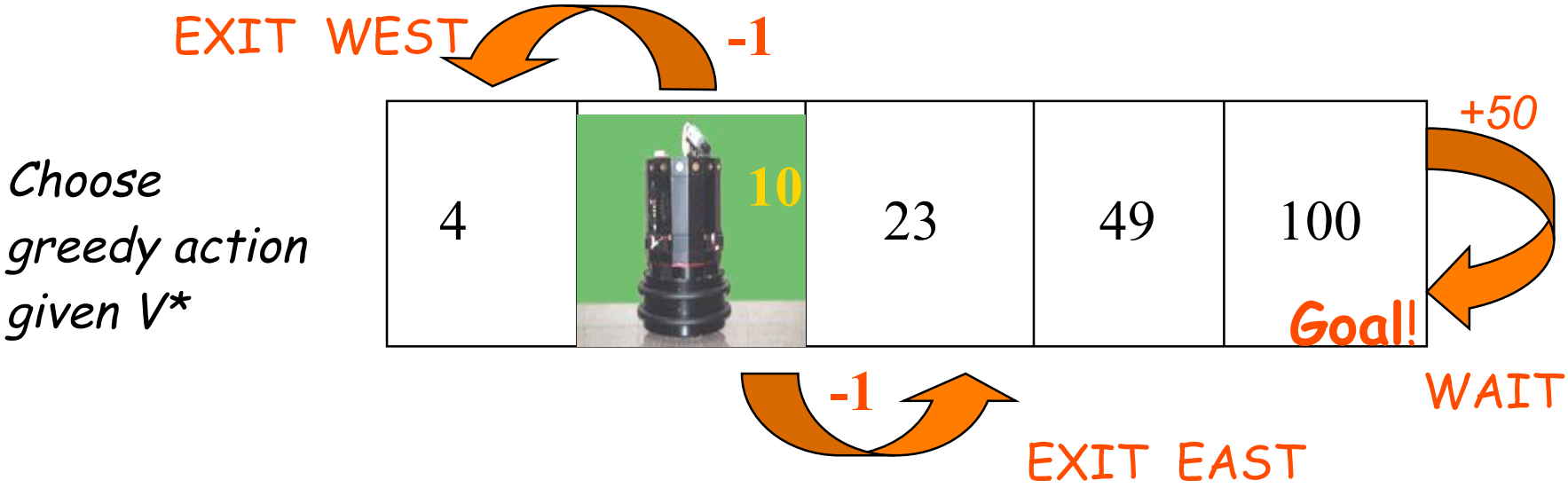
Partially Observable Markov Decision Process

Dynamic Abstractions



Robot Navigation using Global Value Functions

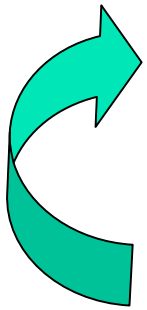
$$V^*(x) = \max_{a \in A(x)} \left(r(x,a) + \gamma \sum_y P^a_{xy} V^*(y) \right)$$

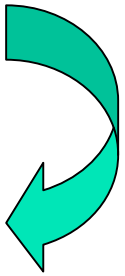


Policy Iteration
(Howard, PhD, MIT, 1959)

Policy Evaluation: ***("Critic")***

$$V^{\pi}(x) = r(x, \pi(x)) + \gamma \sum_{y} P_{xy}^{\pi(x)} V^{\pi}(y)$$

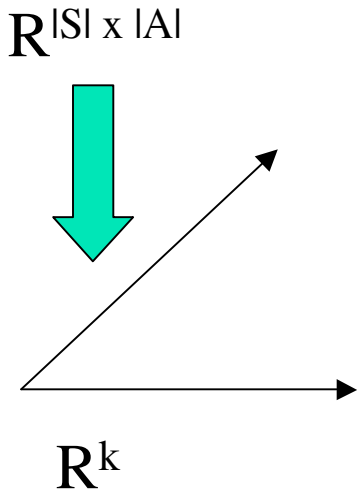




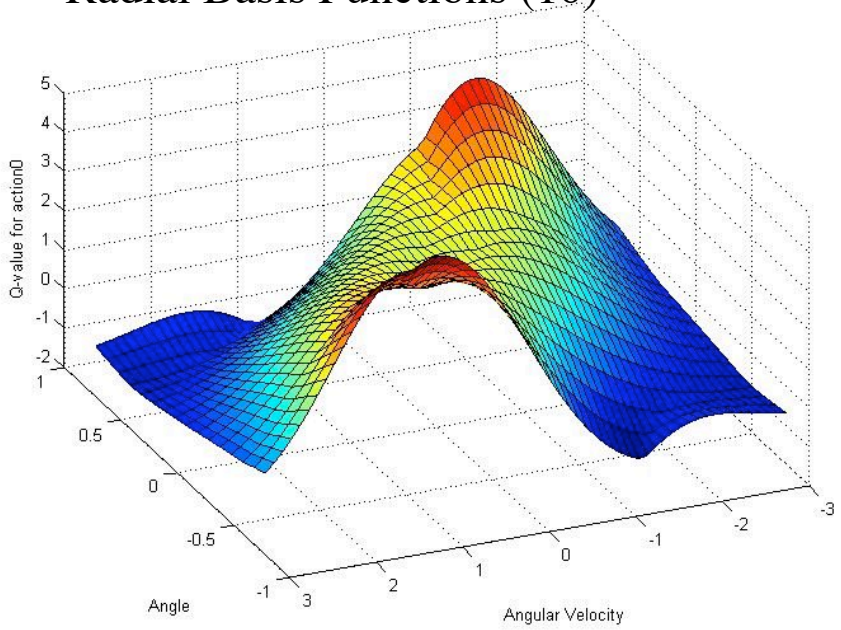
Policy Improvement: ***("Actor")***

$$\pi'(x) = \text{argmax}_a \left(r(x, a) + \gamma \sum_{y} P_{xy}^{a} V^{\pi}(y) \right)$$

Value Function Approximation

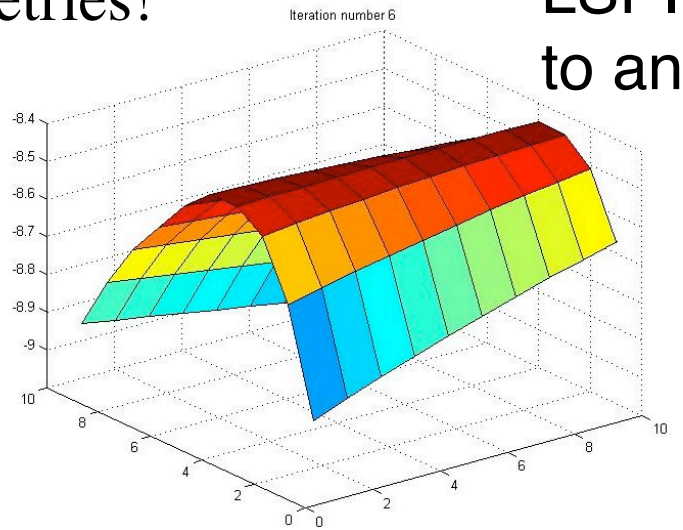


LSPI [Lagoudakis and Parr, JMLR 2003]
Inverted Pendulum with
Radial Basis Functions (10)



Parametric Value Function Approximation Can Fail

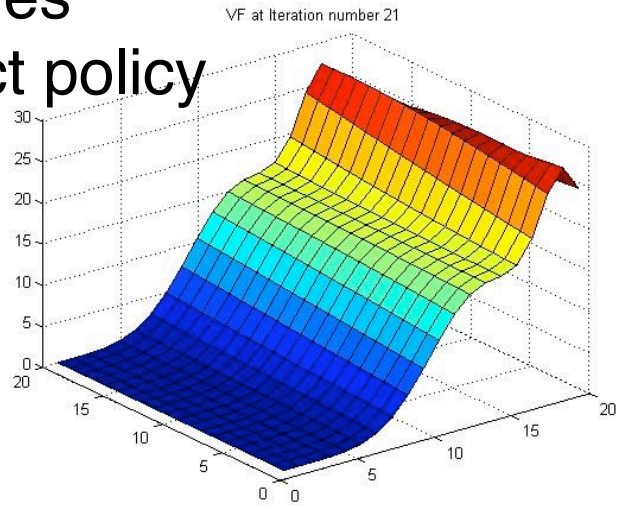
Approximator blind to symmetries!

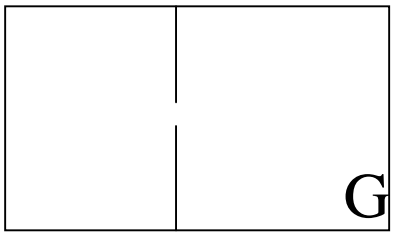


Goal is to get to center of square grid

LSPI converges to an incorrect policy

Approximator blind to bottlenecks!

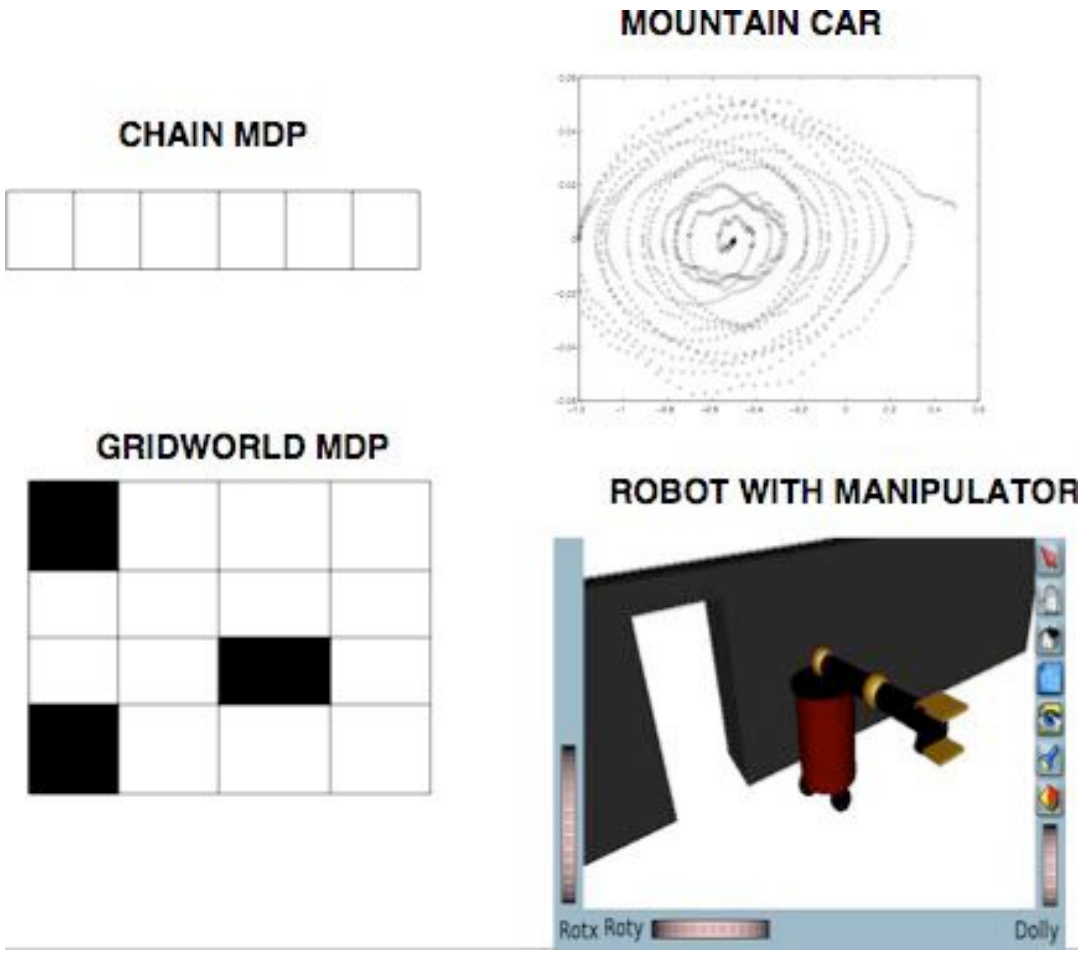




[Drummond, JAIR 2002]

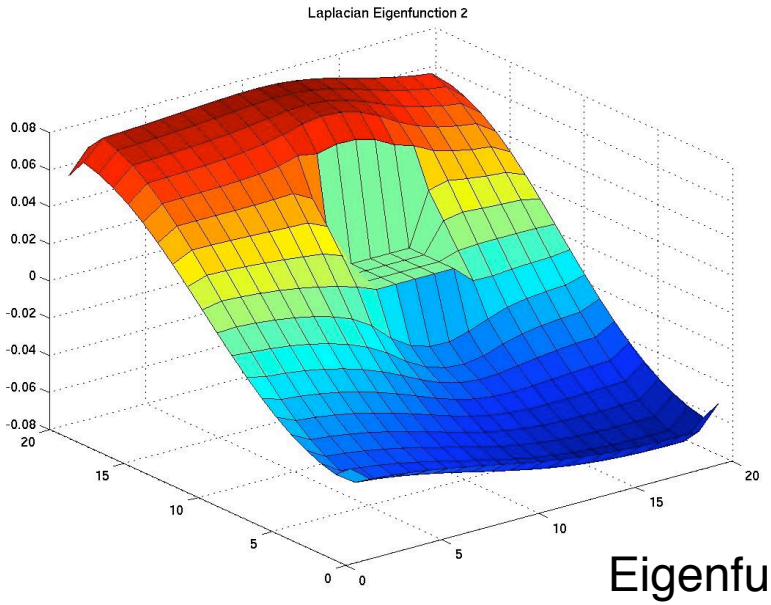
Structural Credit Assignment: Automating Value Function Approximation

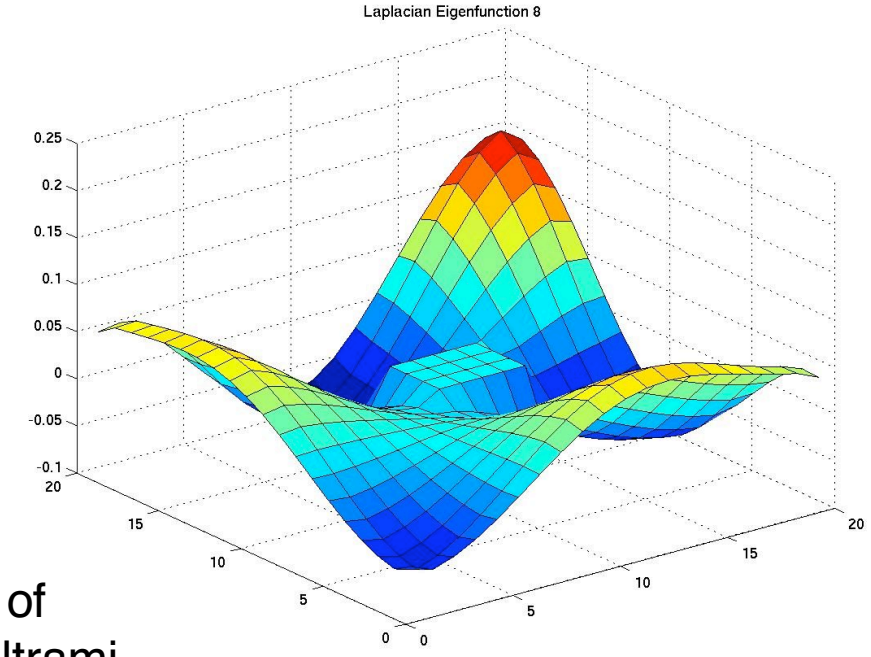
- Many approaches to value function approximation
  - Neural nets, radial basis functions, support vector machines, kernel density estimation, nearest neighbor
- How to automate the design of a function approximator?
- We want to go beyond model selection!



Proto-Value Functions

(Mahadevan: AAAI 2005, ICML 2005, UAI 2005)



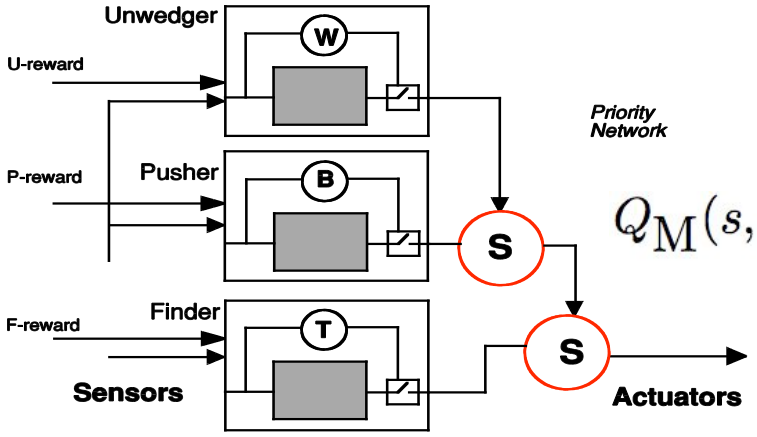


Eigenfunctions of the Laplace-Beltrami operator [Belkin and Niyogi, MLJ 2004]

Task-Level Credit Assignment: Local Value Functions

(Mahadevan and Connell, AAAI 1991; AIJ 1992)

- One approach to across-task transfer is to decompose the *global* value function into *local* value functions
- Q-learning [Watkins, 1989] combined with a handcoded behavioral task decomposition learns much more quickly [Arkin, Brooks]



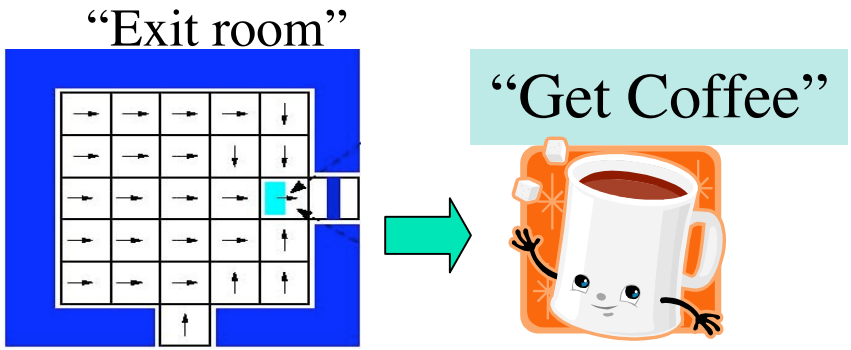
$$Q_{\mathrm{M}}(s,a) = (1-\beta)Q_{\mathrm{M}}(s,a) + \beta(r_M(s,a) + \max_{a'} Q_{\mathrm{M}}(s',a'))$$

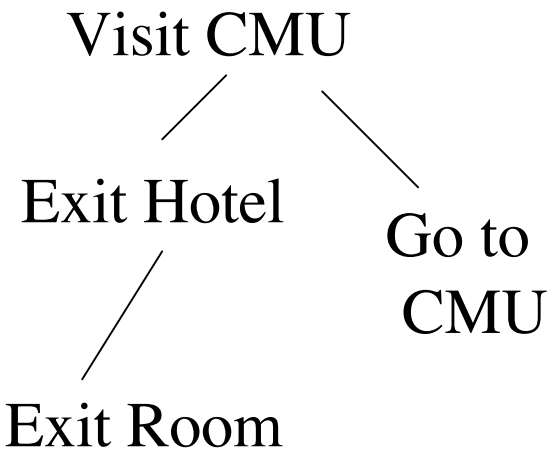
Modeling Temporally Extended Actions

(Barto and Mahadevan, Discrete-Event Systems, 2003)

- Semi-Markov decision process
  - S: set of states
  - A: set of *activities* (or behaviors)
  - P: S x A x S x N → [0,1]: multi-step transition probability
  - R: S x A → $\Re$: expected reward over duration of activity

[Kaelbling, ICML 1993]
[Parr and Russell, NIPS 1998]
[Sutton, Precup, and Singh, AIJ 1999]
[Dietterich, JAIR 2000]





How to Discover Temporal Abstractions?

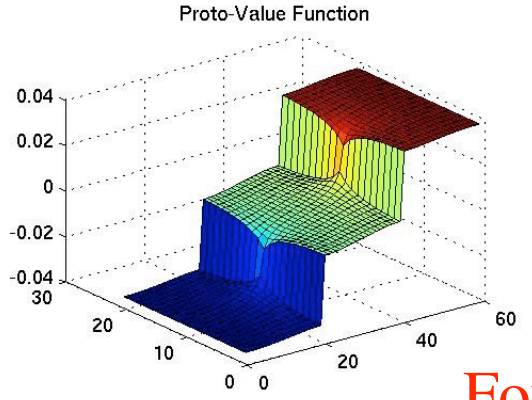
- Find bottlenecks and symmetries in state spaces
  - [McGovern, U.Mass PhD, 2002; Balaraman, U.Mass, PhD 2004]
- Rank state variables by rate of change
  - [Hengst, ICML 2002]
- Use graph clustering and graph partitioning
  - [Menache et al, ECML 2002; Simsek, Wolfe, and Barto, ICML 2005]
- Can such approaches be formalized and generalized to arbitrary (continuous or discrete) state spaces?
  - Build a model of the underlying state space manifold
  - Fourier eigenfunctions of the Laplacian [Belkin and Niyogi, MLJ 2004]
  - Diffusion wavelets based on dilations of the random walk operator on a manifold [Coifman and Maggioni, ACHA, 2005]

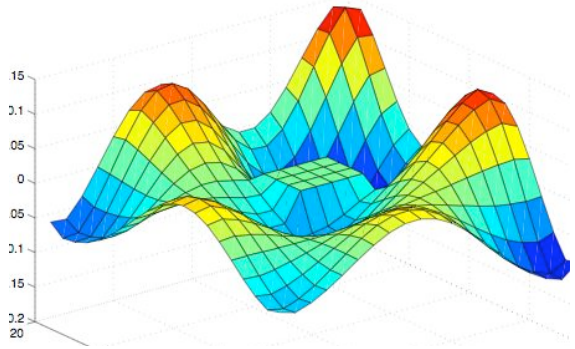
Proto-Value Functions

(Mahadevan, AAAI 2005, ICML 2005, UAI 2005)

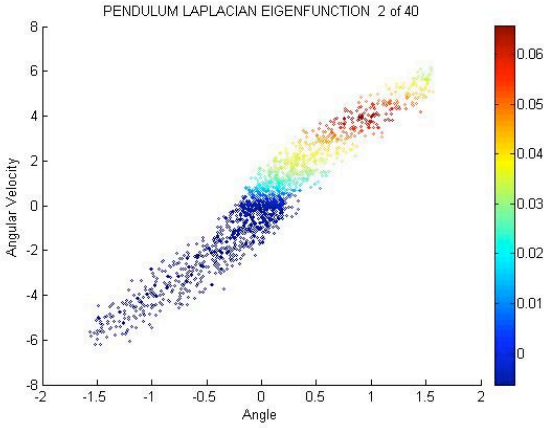
Proto-value functions are the representational building blocks of all value functions on a given environment.

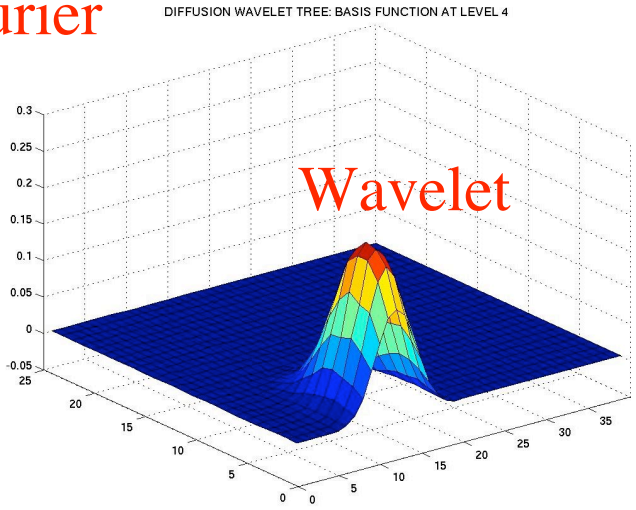
Proto-value functions are based on modeling the state space *manifold*





Fourier





Wavelet

Spectral Graph Theory

- Spectral graph theory is the study of graphs using ideas from continuous manifolds [Chung, AMS 1996]
- Given an undirected graph G = (V, E,W), the normalized graph Laplacian is defined as

$$L = D^{-1/2} (D - W) D^{-1/2}$$

where W is a weight matrix of G

D is the diagonal matrix of row sums of W

  - Proto-value functions are computed by solving the equation

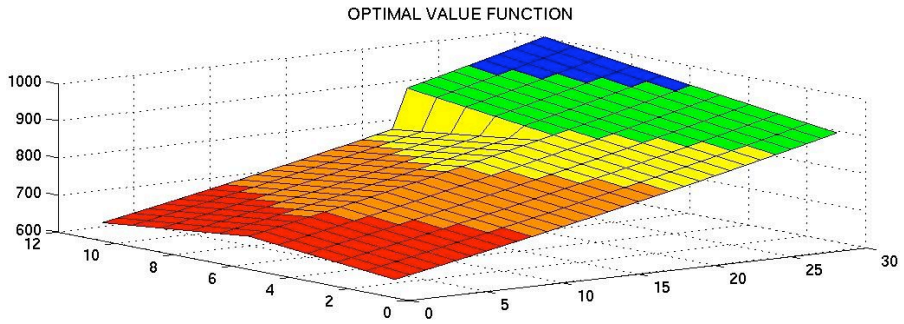
$$L f = \lambda f$$

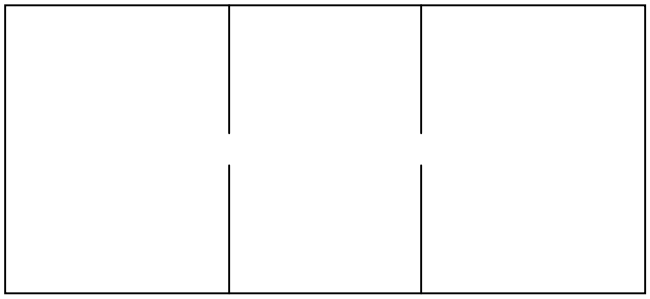
Markov Diffusion Process

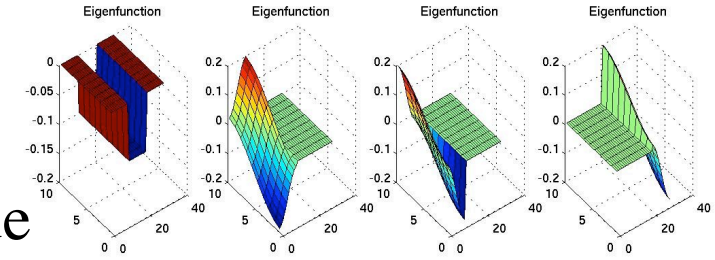
- Given an undirected graph G = (V, E, W), a Markov diffusion process is a random walk on the graph
  - Random walk is defined as $T = D^{-1}W$
  - D is a diagonal matrix of row sums of W
- Graph Laplacian
  - Combinatorial Laplacian: $L = D - W$
  - Normalized Laplacian: $L = D^{-1/2} (D - W) D^{-1/2} = I - D^{-1/2} W D^{-1/2}$
- Note that

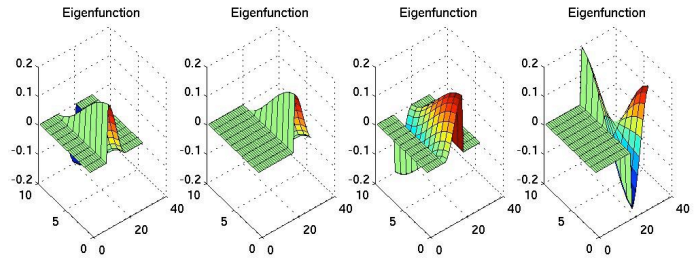
  $T = D^{-1}W = D^{-1/2} (D^{-1/2} W D^{-1/2}) D^{1/2} = D^{-1/2} (I - L) D^{1/2}$

How are Proto-Value Functions Learned?

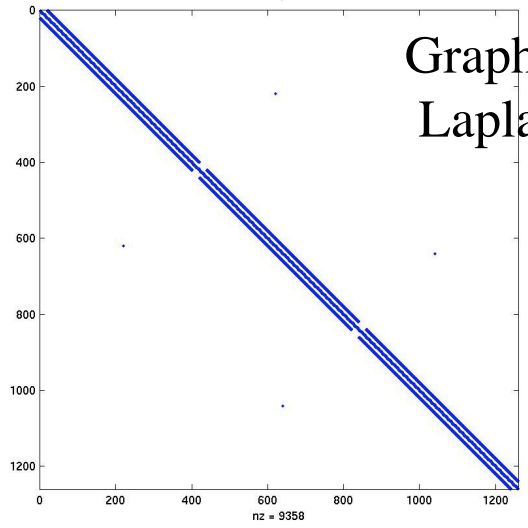








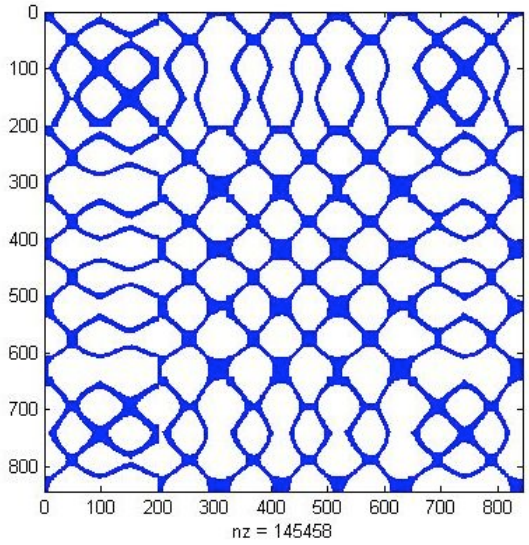
Proto-value functions



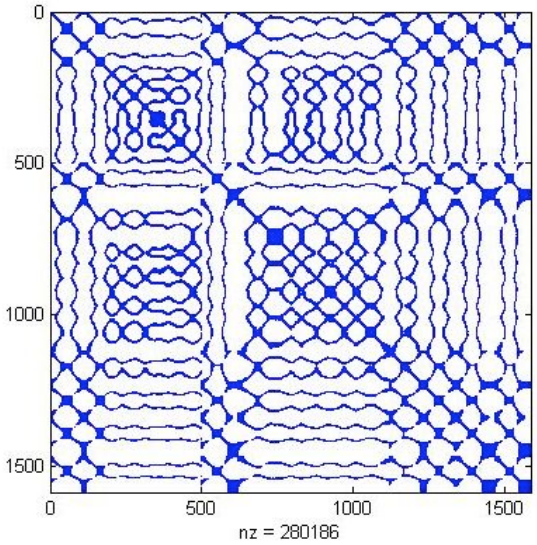
Graph Laplacian

Normalized Graph Laplacian

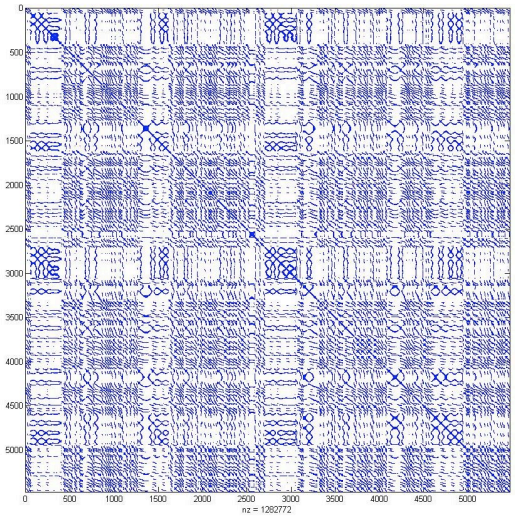
Mountain Car Problem (2D continuous state space)



800 x 800



1500 x 1500



3000 x 3000

Representation Policy Iteration

(Mahadevan, UAI 2005)

- Learn a set of proto-value functions from a sample of transitions generated from a random walk (or from watching an expert)

- These basis functions can then be used in an approximate policy iteration algorithm, such as Least Squares Policy Iteration [Lagoudakis and Parr, JMLR 2003]

Least-Squares Policy Iteration

(Boyan, ICML 1999; Lagoudakis and Parr, JMLR 2003)

Do a random walk generating a set of transitions D = (s_t, a_t, r, s_t')

$$\tilde{A}^{t+1} = \tilde{A}^t + \phi(s_t, a_t)\left(\phi(s_t, a_t) - \gamma\phi(s_t', \pi(s_t'))\right)^T$$

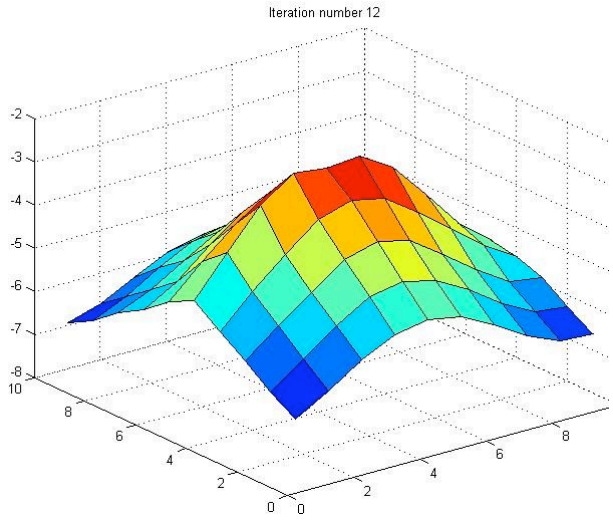
$$\tilde{b}^{t+1} = \tilde{b}^t + \phi(s_t, a_t) r_t$$

Solve the equation: $\tilde{A}w^\pi = \tilde{b}$

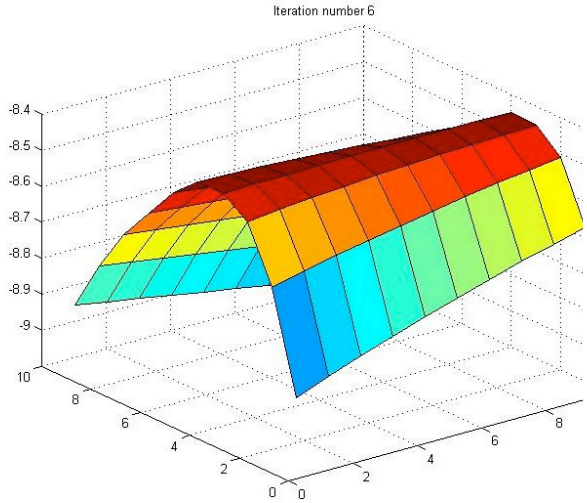
$$Q^\pi(x, a) \approx \sum_{i=1}^{k} \phi(s, a) w_i^\pi$$

Results of RPI on a Grid World

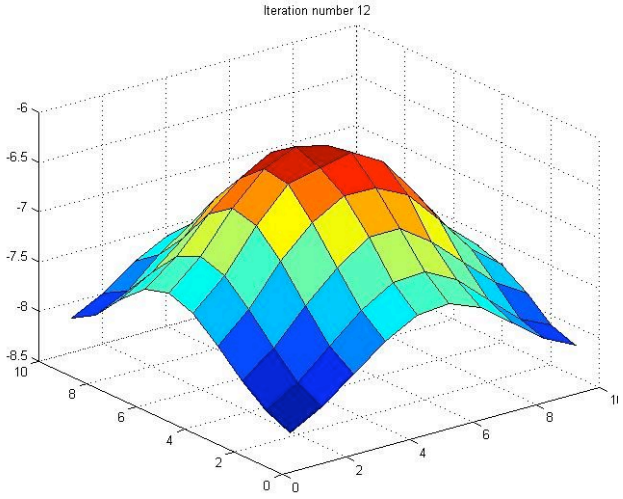
LSPI RBF



LSPI Polynomial

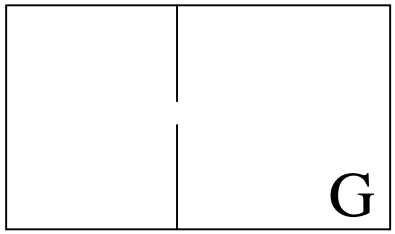


RPI

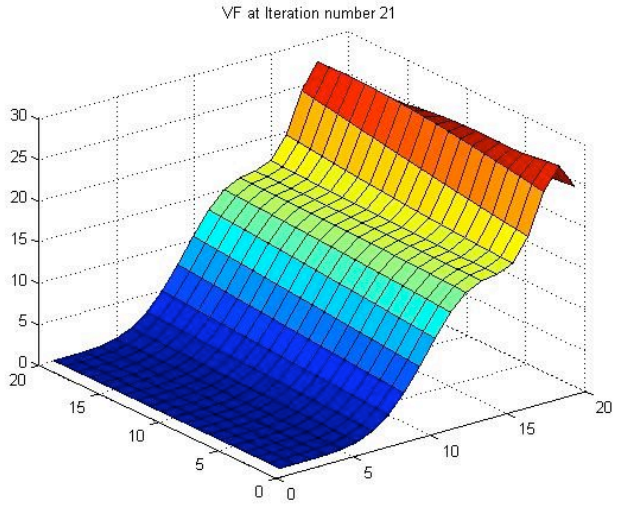


Goal is to get to center of square grid

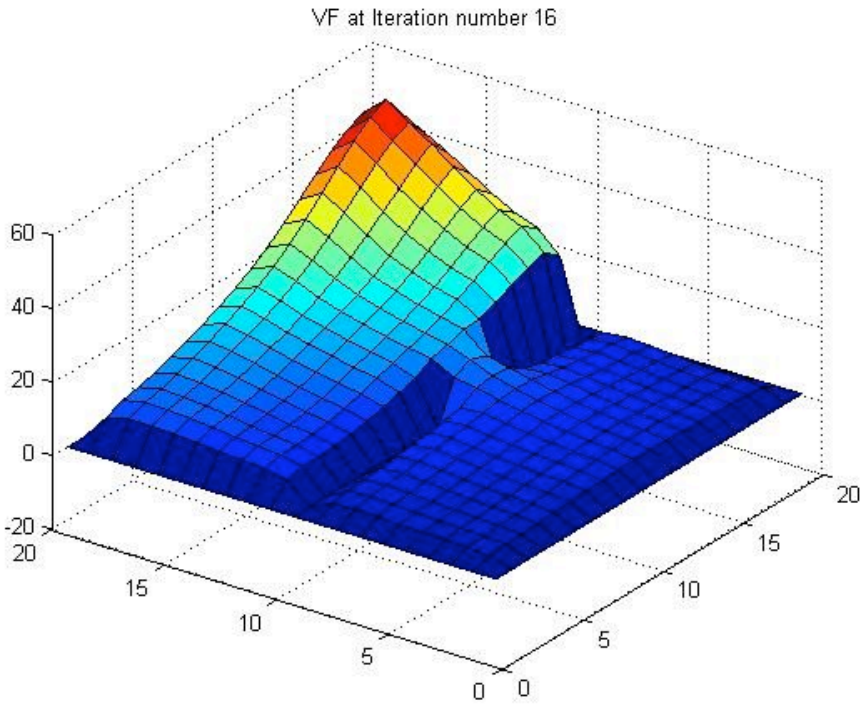
Results of RPI on Two Room World



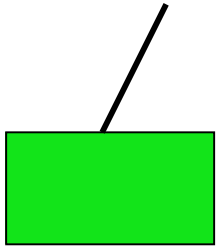
LSPI with polynomial approximation



Nonlinearity due to bottleneck is nicely captured by RPI!

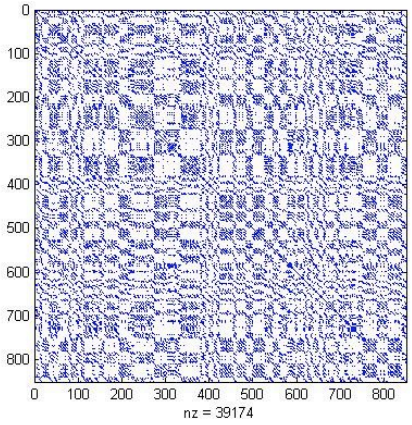


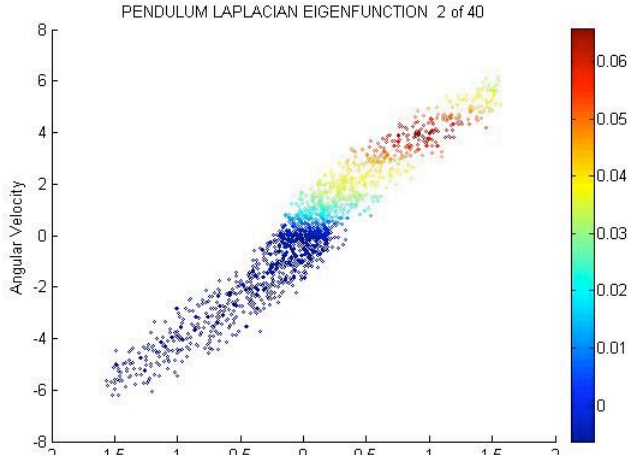
RPI : Inverted Pendulum Task

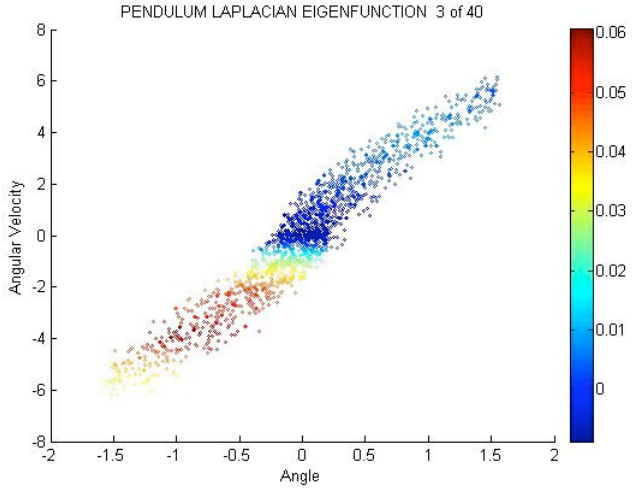


Sample transitions from random walk (~ 800-1600)

Normalized Graph Laplacian matrix

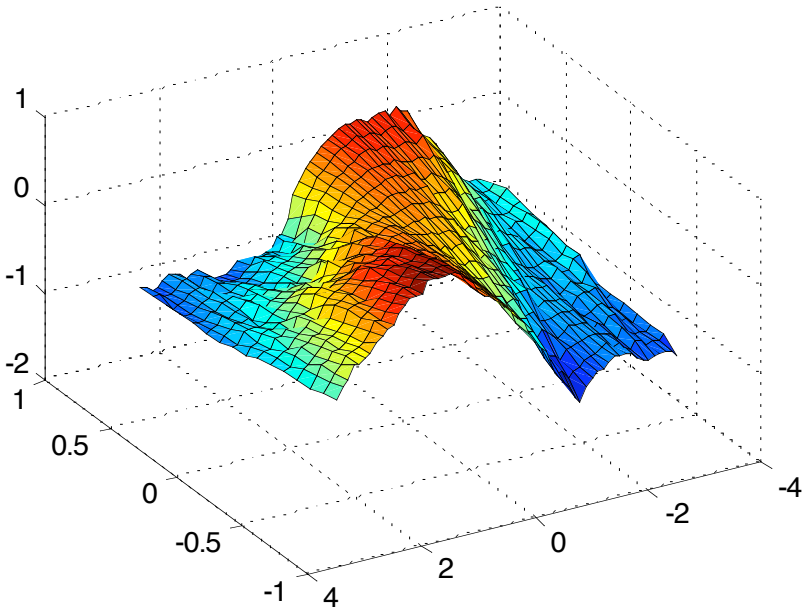




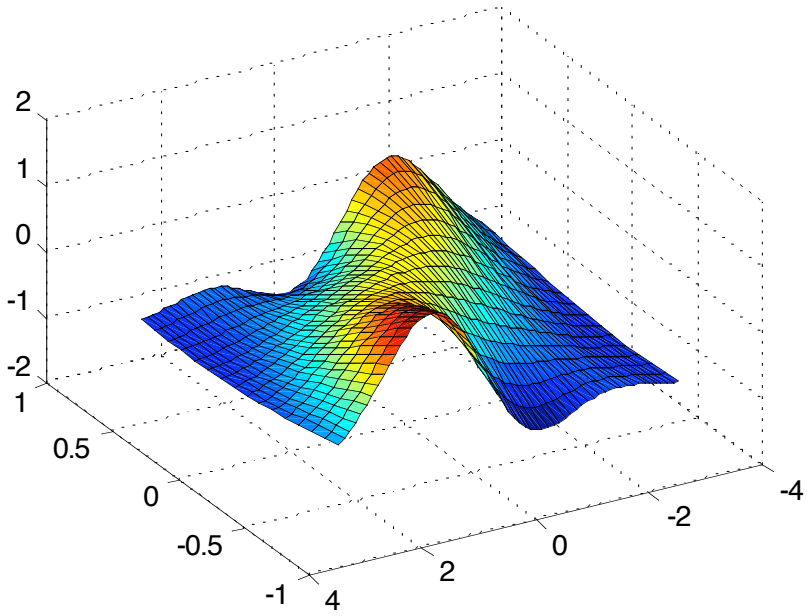


Proto value functions

RPI: Inverted Pendulum



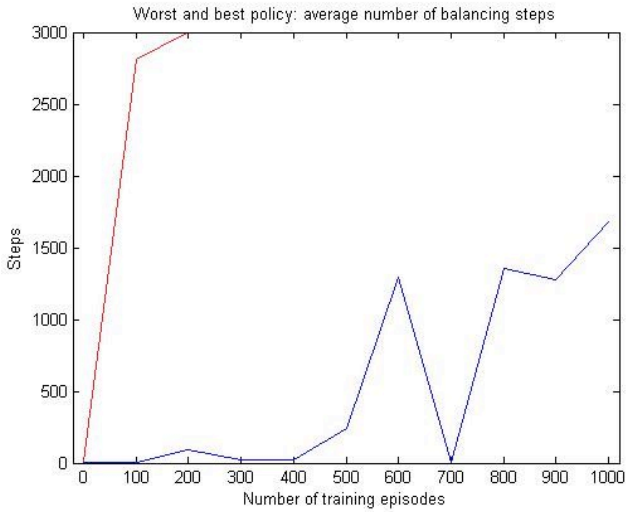
Learned Proto-Value Functions



Handcoded Radial Basis Functions

Handcoded vs. Learned Representations: Inverted Pendulum

LSPI

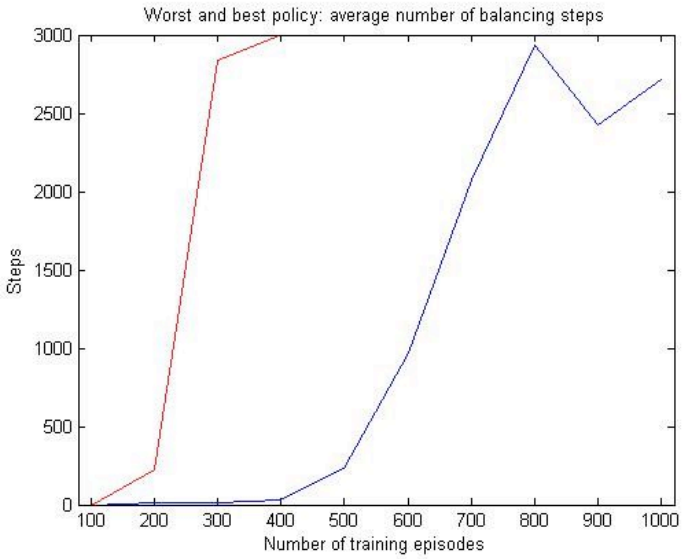


Handcoded (RBF)

Each episode consists of a random walk till the pole is dropped.

Results show the best and worst policy learned over the 10 runs, measured in terms of the number of steps the pole remained upright.

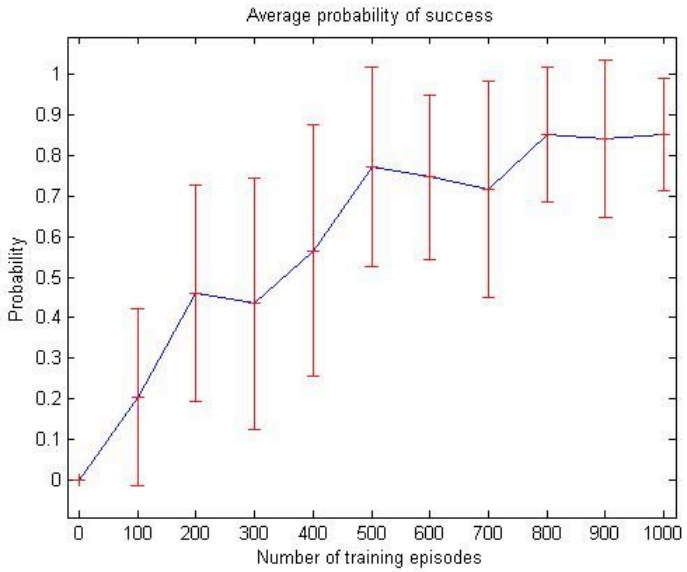
RPI



Learned (Laplacian)

Handcoded vs. Learned Representations: Inverted Pendulum

LSPI



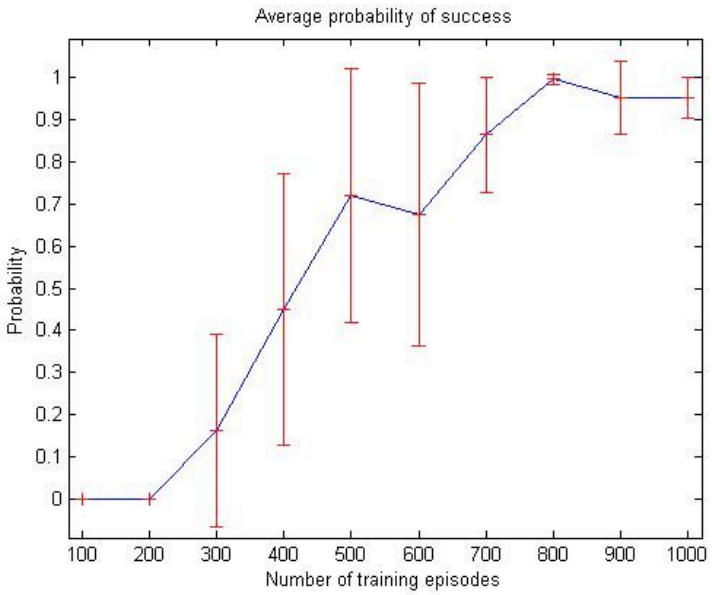
Handcoded (RBF)

Each episode consists of a random walk till the pole is dropped.

Success is measured by measuring the percentage of times the pole is balanced upright for 3000 steps, averaged over 20 trials.

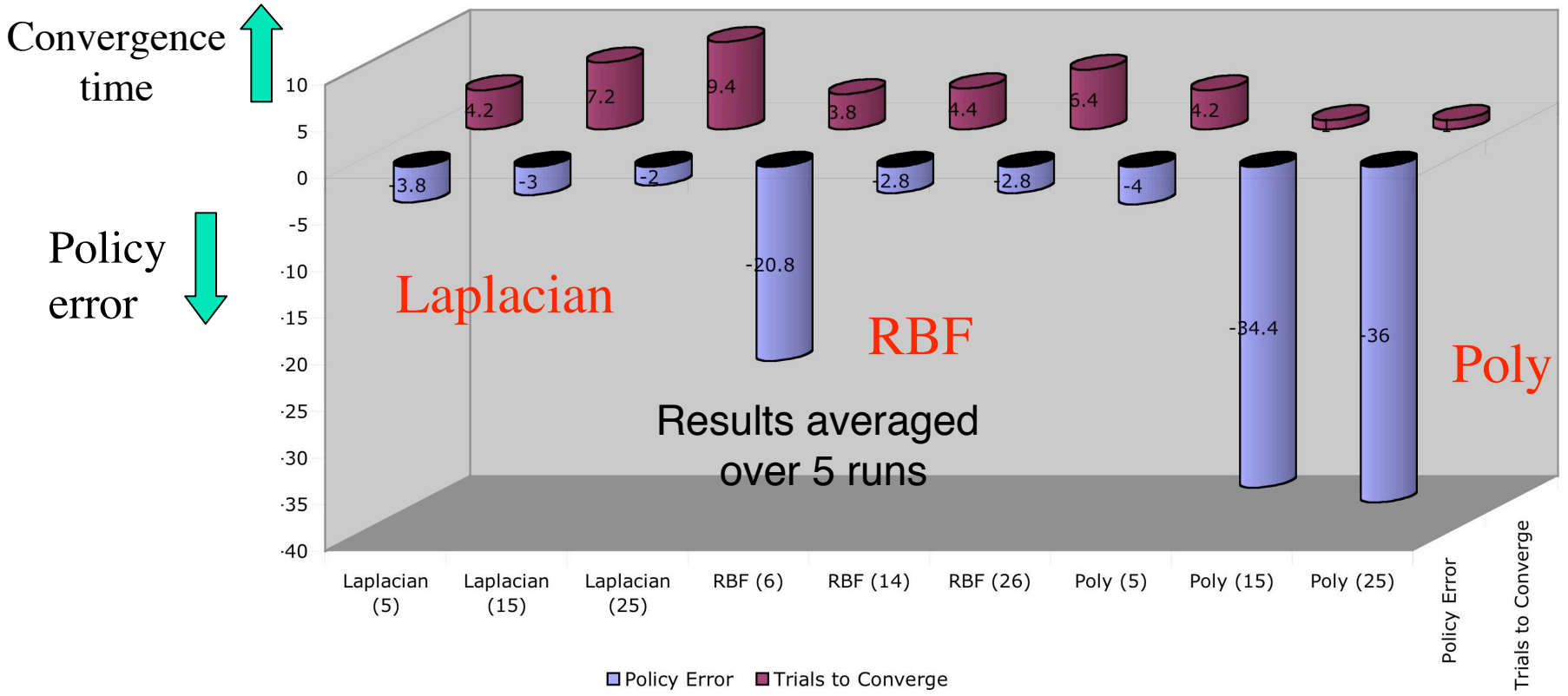
Results shown are averaged over 10 runs.

RPI



Learned (Laplacian)

Learned vs. Handcoded Representations: Chain MDP



Multilevel Proto-Value Functions using Diffusion Wavelets

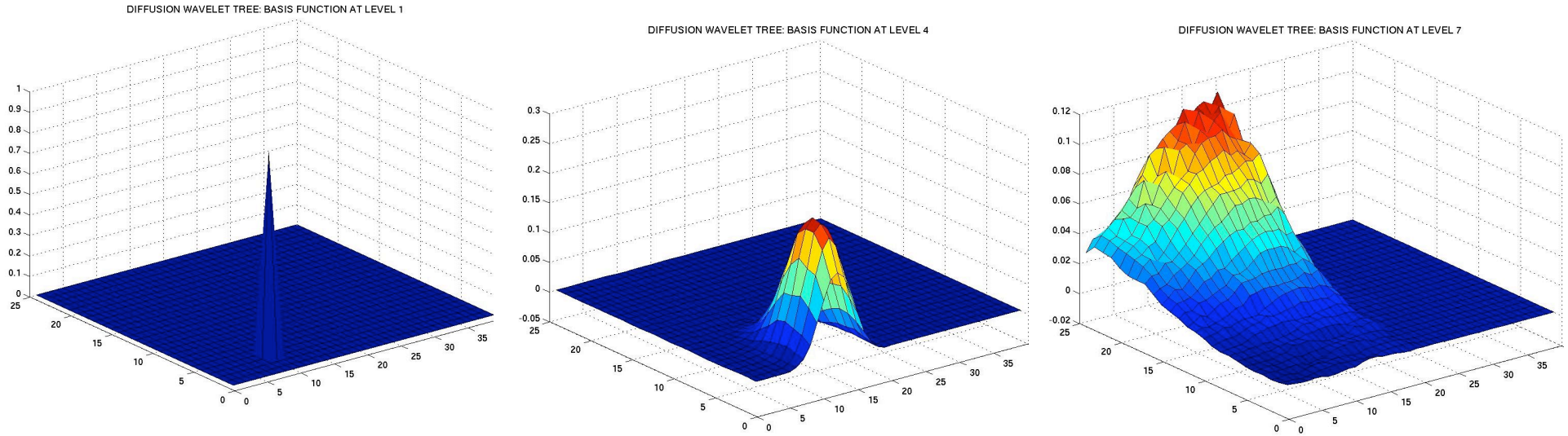
(Coifman and Maggioni, ACHA 2005; Maggioni and Mahadevan, 2005)

- Laplacian eigenfunctions are global Fourier basis functions over the state space
- Wavelets are another popular tool for harmonic analysis that use compact basis functions
- Diffusion wavelets generalize standard wavelets to graphs and manifolds
- They provide a formal basis for multilevel proto-value functions, as well as hierarchical abstraction of Markov processes on graphs

Learning Proto-Value Functions using Diffusion Wavelets

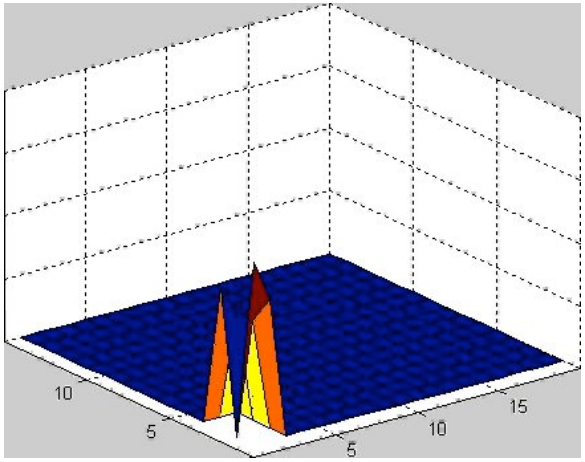
(Mahadevan and Maggioni, 2005)

Level 1 Level 4 Level 7

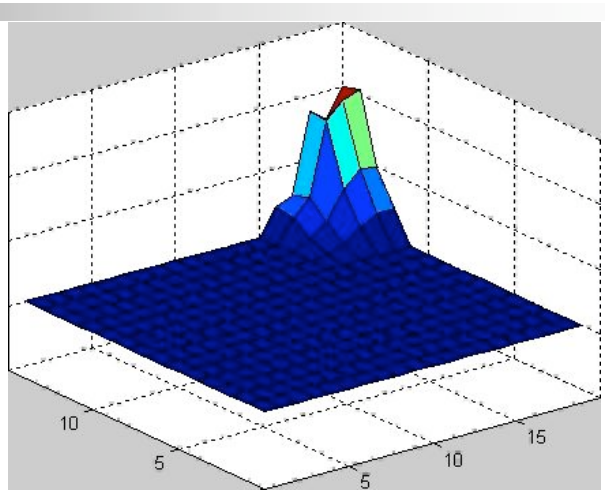


Two room environment

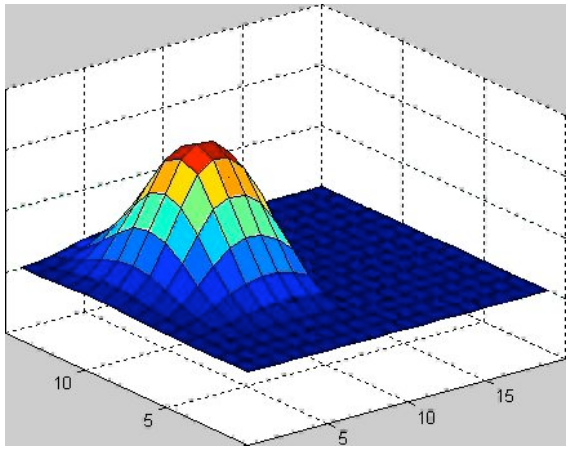
Learning Proto-Value Functions using Diffusion Wavelets



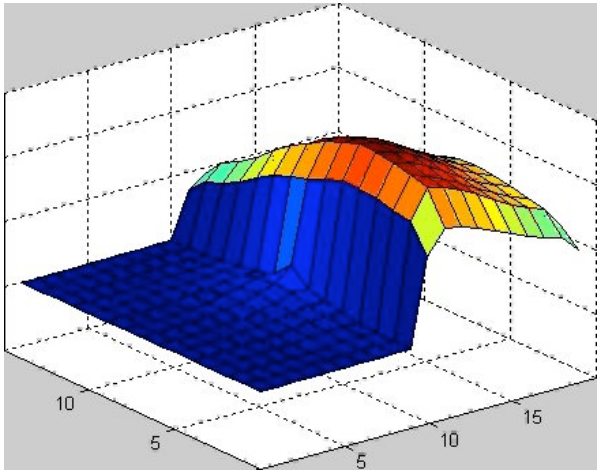
Level 2



Level 3



Level 4



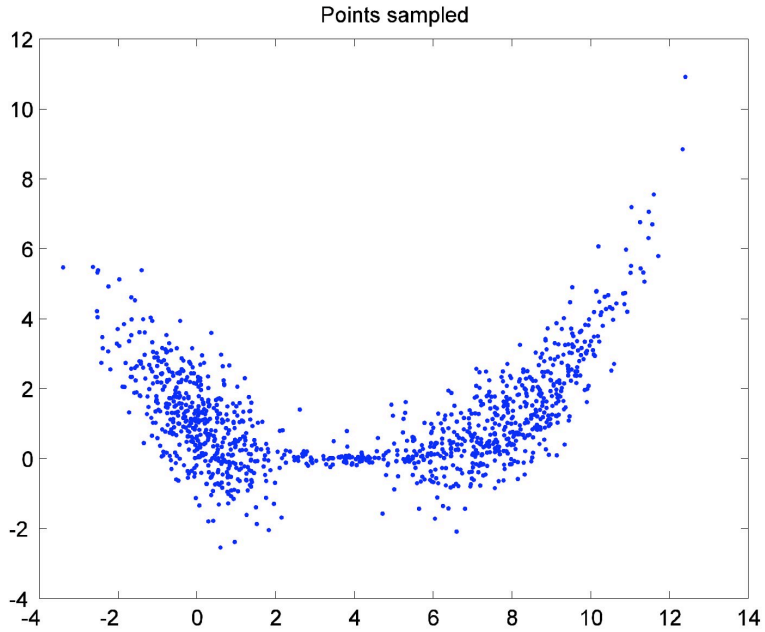
Level 5

Learning Diffusion Wavelets

- Diffusion wavelets are constructed using a layered spectral analysis of reversible random operator T
- The key idea is that powers of T have faster spectral decay than T, and require a smaller set of basis functions
- The algorithm takes as input a precision parameter ε
- It returns a diffusion wavelet tree that constructs a hierarchical layered set of representations of T at varying levels of abstraction

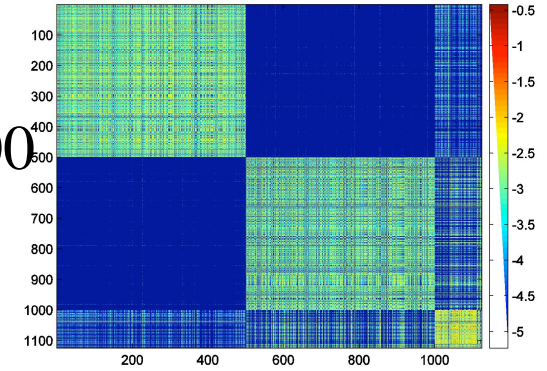
Multiscale Analysis of Random Walks using Diffusion Wavelets

(Maggioni and Mahadevan, U.Mass TR 2005)



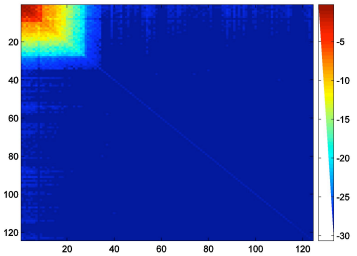
Sampling of a continuous two-room environment

1000x1000



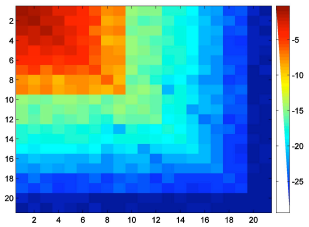
T

100 x 100



T^4

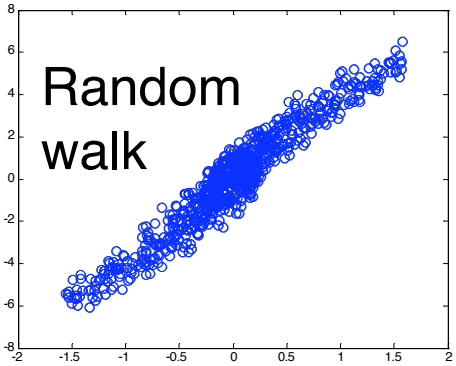
20 x 20



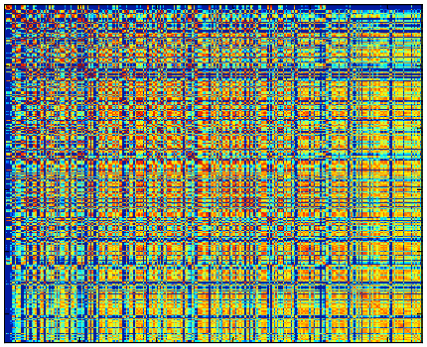
T^6

Abstracting Random Walks using Diffusion Wavelets

Inverted pendulum

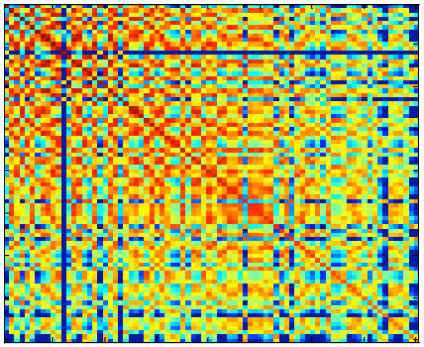


270 x 270



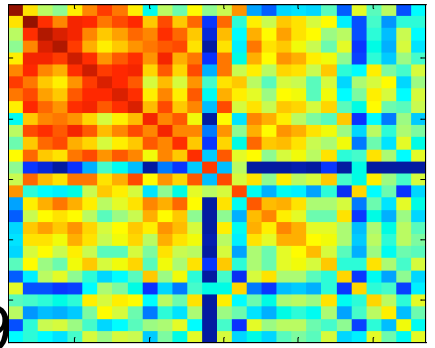
Level 3

80 x80

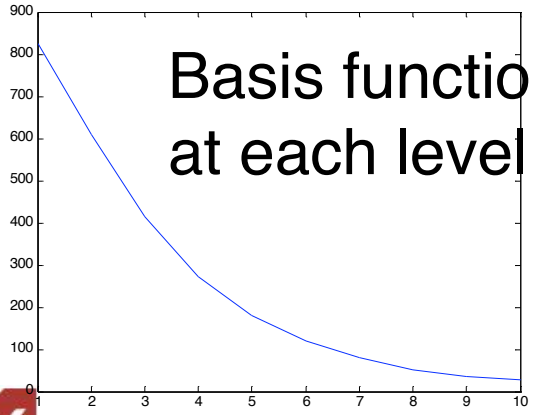


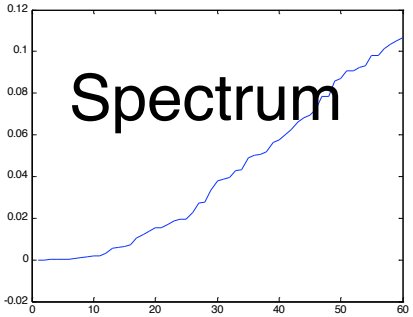
Level 6

28 x 28



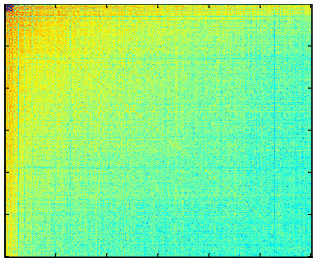
Level 9





Abstracting Random Walks using Diffusion Wavelets

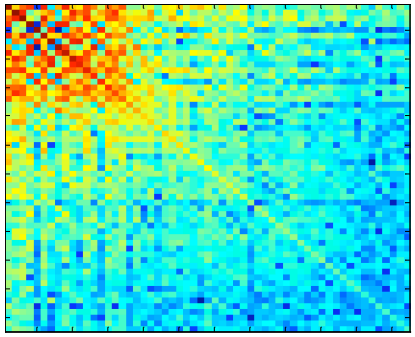
Mountain Car



600 x 600

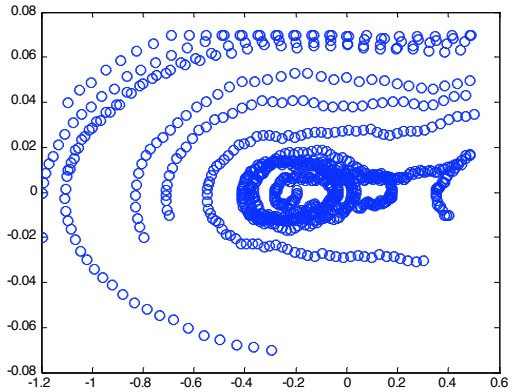
Level 2

55 x 55



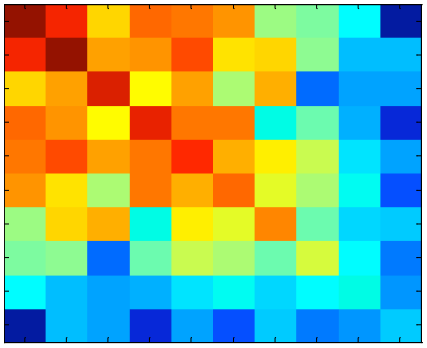
Level 4

Velocity



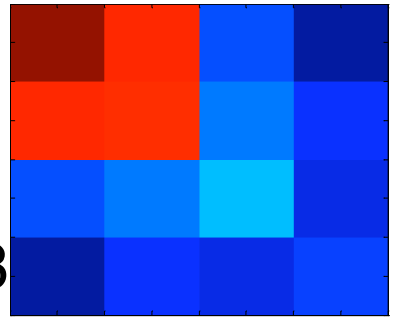
Position

10 x 10



Level 6

5x5



Level 8

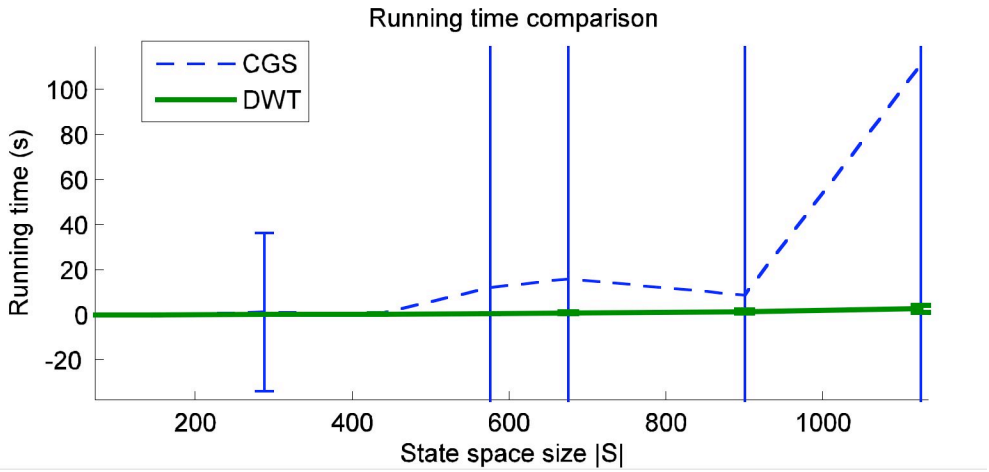
A Faster Critic: Policy Evaluation with Diffusion Wavelets

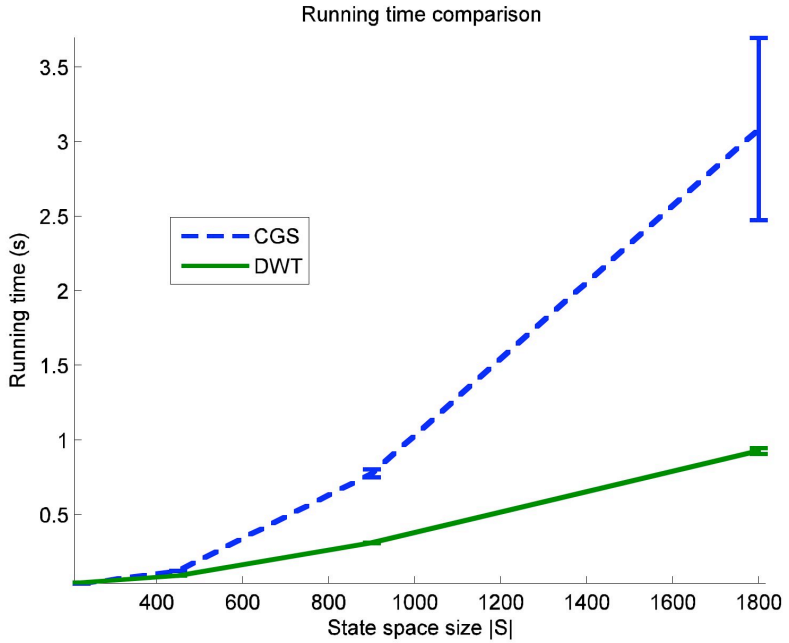
(Maggioni and Mahadevan, U.Mass TR 2005)

$$V^\pi = R^\pi + \gamma P^\pi V^\pi = (I - \gamma P^\pi)^{-1} R^\pi$$

- Traditionally, policy evaluation requires solving a system of |S| linear equations, which takes $O(|S|^3)$
- By constructing a diffusion wavelet tree from the transition matrix, it is possible to develop a significantly faster policy evaluation algorithm that runs in $O(|S| \log^2 |S|)$
  - Precomputation phase: **task-independent**
  - Inversion step: **reward-specific**

Direct Policy Evaluation with Diffusion Wavelets





Summary: Proto-Value Functions

- A major challenge facing AI research is to design a unified framework for learning representation and behavior
  - Representation Policy Iteration
- Proto-value functions
  - Fourier and wavelet representations that capture large-scale geometry of the state space manifold
  - Provides a unified solution to the problem of temporal, structural, and task-level credit assignment
- Extensions of proto-value functions
  - Factored Markov decision processes [Boutilier, Guestrin]
  - Relational domains [Getoor, Jensen, Koller]

Acknowledgements

- Diffusion wavelets software: Mauro Maggioni (Yale Math Department)
- PhD Students:
  - Mohammad Ghavamzadeh, Jeff Johns, Victoria Manfredi, Sarah Osentoski, Khashayar Rohanimanesh, Suchi Saria (Stanford), Georgios Theocharous (Intel))
- Colleagues at U.Mass, Amherst:
  - Andrew Barto, Oliver Brock, Roderic Grupen, Victor Lesser, Andrew McCallum, Paul Utgoff, Shlomo Zilberstein
- PhD Students in the Autonomous Learning Laboratory
  - Ravi Balaraman, Amy McGovern, Mike Rosenstein, Ozgur Simsek, Alicia Wolfe
- Funding:
  - NSF (ECS and ROLE programs), DARPA (MARS, Robot-2020)