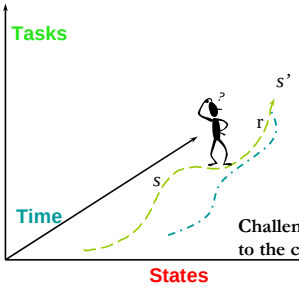
Learning Representation & Behavior:
Manifold and Spectral Methods for
Markov Decision Processes and Reinforcement Learning

Sridhar Mahadevan, U. Mass, Amherst

Outline

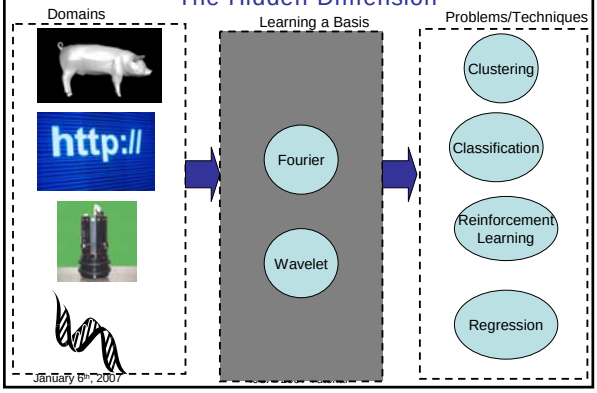
- Learning Representations
  - Harmonic analysis on graphs
  - Fourier and wavelet bases on graph
- Least squares methods for solving MDPs
  - Bellman residual and fixpoint approaches
  - Least-squares policy iteration
- Algorithms
  - Representation Policy Iteration (Mahadevan, UAI 2005)
  - Fast Policy Evaluation (Maggioni and Mahadevan, ICML 2006)
- Experimental Results
- Future Directions

Credit Assignment Problem
(Minsky, Steps Toward AI, 1960)

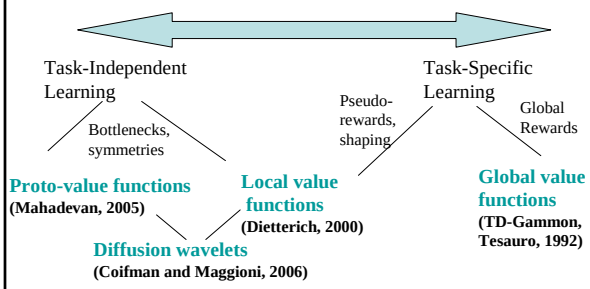


Challenge: Need a unified approach to the credit assignment problem

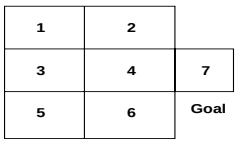
Learning Representations: The Hidden Dimension



Combining Reward-Specific and Reward-Independent Learning



How to find a good basis?



$e_1 = [1, \ldots, 0]$

$e_i = [0, \ldots, i, \ldots, 0]$

Any function on this graph is a vector in R^7

The question we want to ask is how to construct a basis set for approximating functions on this graph

Solution 1: use the *unit basis*

Solution 2: use *polynomials or RBFs*

Neither of these exploit geometry

Polynomial Bases

(Samuel, 1959; Koller and Parr, UAI 2000)

One basis function applied to all states

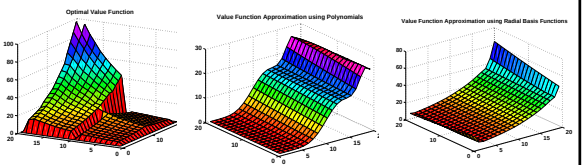
Φ

All basis functions applied to one state

1		1	1	1	1	1
1		4	8	16	32	64
1		9	27	81	243	729
1		16	64	256	1024	4096
1		25	125	625	3125	15625
1		36	216	1296	7776	46656
1		49	343	2401	16807	117649
i^0	i	i^2	i^3	i^4	i^5	i^6

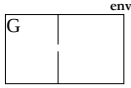
Parametric VFA Can Easily Fail!

(Dayan, Neural Comp, 1993; Drummond, JAIR 2003)



OPTIMAL VF

Multi-room environment



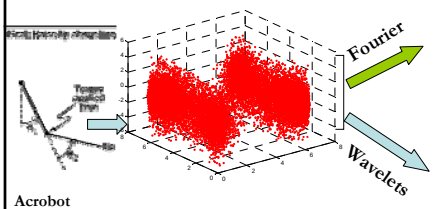
POLYNOMIAL

RADIAL BASIS FUNCTION

These approaches measure distances in ambient space, not on the manifold!

Learning Representation and Behavior

(Mahadevan, AAAI,ICML,UAI 2005; Mahadevan & Maggioni, NIPS 2005; Maggioni and Mahadevan, ICML 2006)



Acrobot

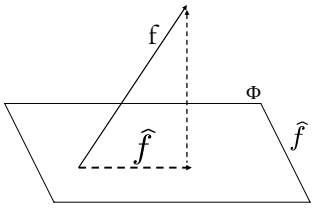
Samples from random walk on manifold in state space

$\Delta\phi = \lambda\phi$

Eigenfunctions of Laplacian on state space: global basis functions

Dilations of diffusion operator: local multiscale basis functions

Least-Squares Projection



Theory of least-squares tells us how to find the closest vector in a subspace to a given vector

$$\hat{f} = \Phi(\Phi^T\Phi)^{-1}\Phi^T f$$

What is an optimal subspace Φ for approximating f on a graph?

Standard bases (polynomials, RBFs) don't exploit geometry of graph

Fourier Analysis on Graphs

(Fiedler, 1973; Cvetkovic et al, 1980; Chung, 1997)

1	2	
3	4	7
5	6	Goal

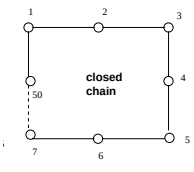
Laplacian Matrix = D - A

2	-1	-1	0	0	0	0
-1	2	0	-1	0	0	0
-1	0	3	-1	-1	0	0
0	-1	-1	4	0	-1	-1
0	0	-1	0	2	-1	0
0	0	0	-1	-1	2	0
0	0	0	-1	0	0	1

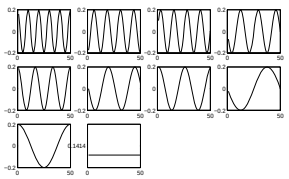
Row sums

Negative of weights

One-Dimensional Chain



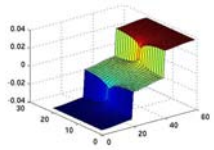
Cayley graph



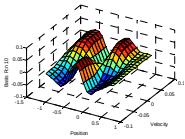
Fourier Basis: Eigenvectors of the Graph Laplacian

Eigenvectors of Graph Laplacian: Discrete and Continuous Domains

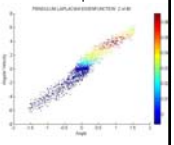
Three rooms with bottlenecks

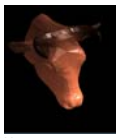


Mountain Car Problem



Inverted pendulum





3D Modeling
In Graphics

Matrix Decompositions

- Spectral Theorem: $\Phi = U \Lambda U^T$
  - U is an orthonormal basis of eigenvectors
- Singular value decomposition: $\Phi = U \Sigma V^T$
  - U is an orthonormal basis of eigenvectors of $\Phi \Phi^T$
  - V is an orthonormal basis of eigenvectors of $\Phi^T \Phi$
- Gram-Schmidt: $\Phi = Q R$
  - Q is an orthonormal basis, R is a lower triangular matrix
- Linear system $\Phi x = u$
  - Krylov space: $[u, \Phi u, \Phi^2 u, ..., \Phi^n u]$
  - Useful for reward-specific bases [Petrik, IJCAI 2007]

Graph Laplacian = Inverse Covariance

- The Laplacian is *positive semidefinite*

  - The Laplacian for this graph is [1 -1; -1 1]
  - Observe that $f^T L f = (f_1 - f_2)^2$
- Generalizing, the Laplacian of a weighted graph G = (V, E, W) has the property

$$\langle f, Lf \rangle = f^T L f = \Sigma w_{u,v} (f_u - f_v)^2$$

- Laplacian can be viewed as inverse covariance, inducing a distribution on functions on a graph

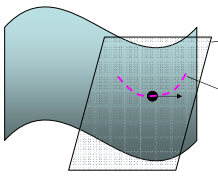
$$P(f) = 1/Z \, e^{-(\beta f^T L f)}$$

Manifold and Spectral Learning

- Spectral methods are based on computing eigenvectors of a normalized "affinity" matrix
  - [Shi and Malik, IEEE PAMI 1997]
  - [Ng, Jordan, and Weiss, NIPS 2001]
  - PageRank [Page, Brin, Motwani, Winograd, 1998]
- Manifold methods model the local geometry of the data by constructing a graph
  - [Roweis and Saul; Tenenbaum, de Silva, Langford, Science 2000]
  - [Belkin and Niyogi, MLJ 2004]
  - [Weinberger, Sha, Saul, ICML 2004]
- These methods are closely related to kernel PCA
  - [Scholkopff, Smola and Muller, 2001]
  - [Bengio et al, Neural Computation, 2004]

Manifolds

$\mathcal{M}^k$: Set that looks "locally" Euclidean



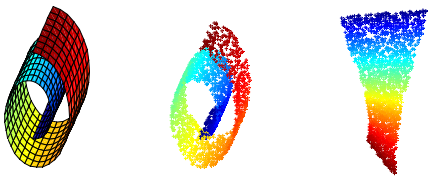
$T_p \mathcal{M}^k$: tangent space (affine subspace of R^n)

$\phi(t): \mathcal{R} \rightarrow \mathcal{M}^k$ (curve)

$f: \mathcal{M}^k \rightarrow \mathcal{R}$

$f(\phi(t)): \mathcal{R} \rightarrow \mathcal{R}$

Nonlinear dimensionality reduction
(LLE, ISOMAP, Laplacian Eigenmaps)



Embedding should preserve "locality"

Graph Embedding

- Consider the following optimization problem mapping, where $y_i \in R$ is a mapping of the i^{th} vertex to the real line

$$\text{Min}_y \sum_{i,j} (y_i - y_j)^2 w_{i,j} \quad \text{s.t. } y^T D y = 1$$

- The best mapping is found by solving the generalized eigenvector problem

$$W \phi = \lambda D \phi$$

- If the graph is connected, this can be written as

$$D^{-1} W \phi = \lambda \phi$$

Range of Laplacian Operators

- Random walk: $P = D^{-1} W$
- Discrete Laplacian: $L_d = I - D^{-1}W$
- Normalized Laplacian: $\mathcal{L} = D^{-1/2} (D - W) D^{-1/2}$
- Directed Combinatorial Laplacian: μ is the Perron vector

$$L = \mu - (\mu P + P^T \mu)/2$$

- Directed Normalized Laplacian: μ is the Perron vector

$$\mathcal{L} = I - (\mu^{1/2} P \mu^{-1/2} + \mu^{-1/2} P^T \mu^{1/2})/2$$

Normalized Graph Laplacian and Random Walks

- Given an undirected weighted graph G = (V, E, W), the random walk on the graph is defined by the transition matrix

$$P = D^{-1}W$$

  – Random walk matrix is not symmetric
- Normalized Graph Laplacian

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

- The random walk matrix has the same eigenvalues as $(I - \mathcal{L})$

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

Operators on Graphs

Operator	Spectrum
Adjacency $= A$	Real, $\lvert\lambda\rvert \leq d_v$
C. Laplacian $= D - A$	PSD, $\lambda \geq 0$
N. Laplacian $= I - D^{-1/2} A D^{-1/2}$	PSD, $\lambda \in (0, 2)$
Random Walk $= D^{-1} A$	$\lambda \in \mathcal{R}$
DWT $= \mathcal{T}$	$\sigma(\mathcal{T}^k)$ decays fast

Optimality of Bases

- Under what conditions is a given basis "optimal"?
  – MSE: A basis Φ is optimal for a distribution D over R^n if

$$\text{MSE} = E_D((X - Y_{\Phi,k})^2)$$

  is the lowest over all other basis choices
  – PCA: Choose highest order eigenvectors of covariance matrix C
  – PCA basis can be computed on a finite sample using SVD
- For points distributed on $X \in R$ or $\in R^2$ resulting in a valid chain graph (1D) or a mesh graph (2D):
  – Fourier (Laplacian) basis is optimal (under some conditions)
  – Covariance C is given by the inverse of the Laplacian [Ben-Chen and Gotsman, ACM Trans. On Graphics, 2005]
- Since the Laplacian is the inverse of covariance, we choose its lowest order eigenvectors

Multiresolution Manifold Learning

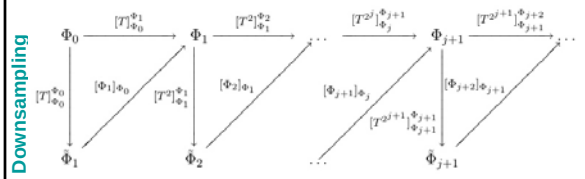
- Fourier methods, like Laplacian manifold or spectral learning, rely on eigenvectors
  – Eigenvectors are useful in analyzing long-term global behavior of a system (e.g, PageRank)
  – They are rather poor at short or medium term transient analysis (or locally discontinuities)
- Wavelet methods [Daubechies, Mallat]
  – Inherently multi-resolution analysis
  – Local basis functions with compact support
- Diffusion wavelets [Coifman and Maggioni, 2004]
  – Extend classical wavelets to graphs and manifolds

Diffusion Wavelets

(Coifman and Maggioni, ACHA, 2006 ;Maggioni and Mahadevan, U.Mass TR, 2006)

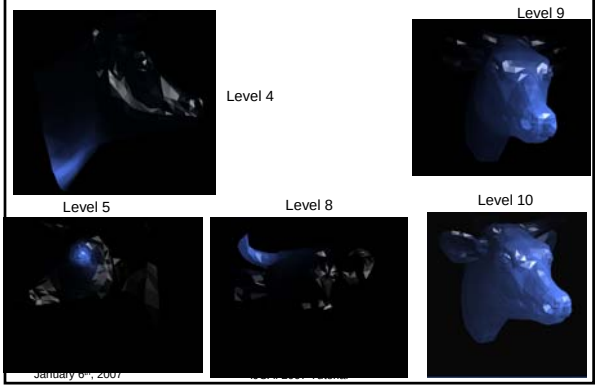
Dilation

Downsampling



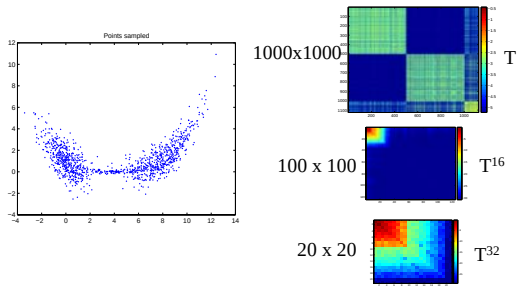
DWT Algorithm

Multiscale DWT Bases: 3D Objects



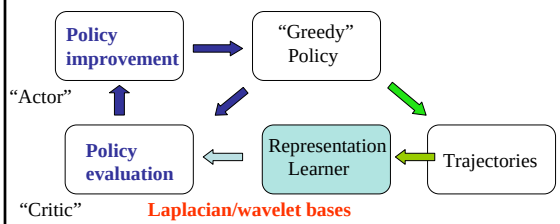
Fast Inversion of Random Walks

(Maggioni and Mahadevan, ICML 2006)



Representation Policy Iteration

(Mahadevan, UAI 2005)



Bellman Residual Method

(Munos, ICML 03; Lagoudakis and Parr, JMLR 03)

- Let us write the Bellman equation in matrix form as

$$\Phi\, w^{\pi} \approx R^{\pi} + \gamma\, P^{\pi}\, \Phi\, w^{\pi}$$

- Collecting the terms, we rewrite this as

$$(\Phi - \gamma\, P^{\pi}\, \Phi)\, w^{\pi} \approx R^{\pi}$$

- The least-squares solution is

$$w^{\pi} = [(\Phi - \gamma\, P^{\pi}\, \Phi)^{T} (\Phi - \gamma\, P^{\pi}\, \Phi)]^{-1} (\Phi - \gamma\, P^{\pi}\, \Phi)^{T} R^{\pi}$$

Bellman Fixpoint Method

- Another way to obtain a least-squares solution is to project the backed-up value function $T^\pi(V^\pi)$

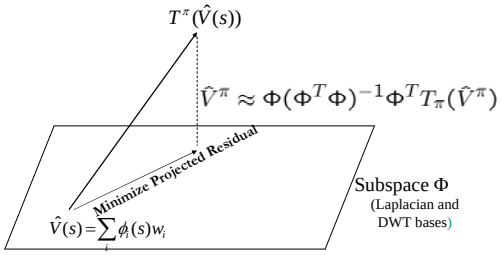
$$P = \Phi\ (\Phi^T\ \Phi)^{-1}\ \Phi^T$$

- The least-squares projected weights then becomes

$$W^\pi = (\Phi^T\ \Phi)^{-1}\ \Phi^T\ [R^\pi + \gamma\ P^\pi\ V^\pi]$$

VFA using Least-Squares Projection

(Boyan, Bradtke and Barto, Bertsekas and Nedic, Lagoudakis and Parr)



Least-Squares Policy Iteration

(Lagoudakis and Parr, JMLR 2003)

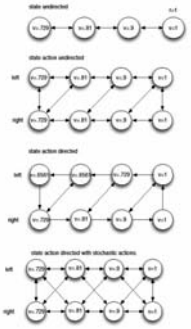
Initial random walk: $\mathcal{D} = (s_t, a_t, r, s_t')$

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

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

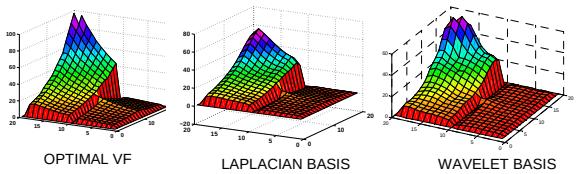
Graph Construction

- The graph is constructed using a similarity metric
  - In discrete spaces, connect state s to s' if an action led the agent from $s \rightarrow s'$
  - Action respecting embedding [Bowling, ICML 2005]
- Other distance metrics:
  - Nearest neighbor: connect an edge from s to s' if s' is one of k nearest neighbors of s
  - Heat kernel: connect s to s' if $|s - s'|^2 < \varepsilon$ with weight $w(s,s') = e^{-|s-s'|^2/2}$



Value Function Approximation using Laplacian and Wavelet Bases

(Mahadevan and Maggioni, NIPS 2005)



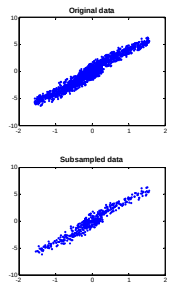
OPTIMAL VF | LAPLACIAN BASIS | WAVELET BASIS

RPI in Continuous State Spaces

(Mahadevan, Maggioni, Ferguson, Osentoski, AAAI 2006)

- RPI in continuous state spaces
  - The Nystrom extension interpolates eigenfunctions from sample points to new points
- Many practical issues are involved
  - How many samples to use to build the graph?
  - Local distance metric: Gaussian distance, k-NN
  - Graph operator: Normalized Laplacian, Combinatorial Laplacian, Random Walk, …
  - Type of graph: Undirected, directed, state-action graph

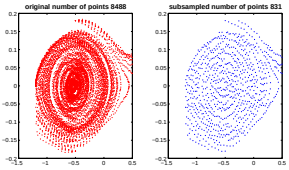
Sampling from a Continuous Manifold



Inverted Pendulum

Random subsampling

Trajectory subsampling



Mountain Car domain

Out of Sample Extension

- The testing of a learned policy requires computing the basis in novel states
- The Nystrom extension is a classical method developed in the solution of integral equations

$$\phi_m(x) = 1/\lambda_m \sum_j w_j k(x,s_j) \phi_m(s_j)$$

- Other approaches:
  - Nearest neighbor averaging (diffusion wavelets)
  - Smoothness-sensitive interpolation

The Nystrom method

(Williams and Seeger, NIPS 2001)

- The Nystrom approximation was developed in the context of solving integral equations

$$\int_D K(t,s)\,\Phi(s)\,ds = \lambda\,\Phi(t),\ t \in D$$

- A quadrature approximation of the integral:

$$\int_D K(t,s)\,\Phi(s)\,ds = \sum_j w_j\,k(x,s)\,\phi(s_j)$$

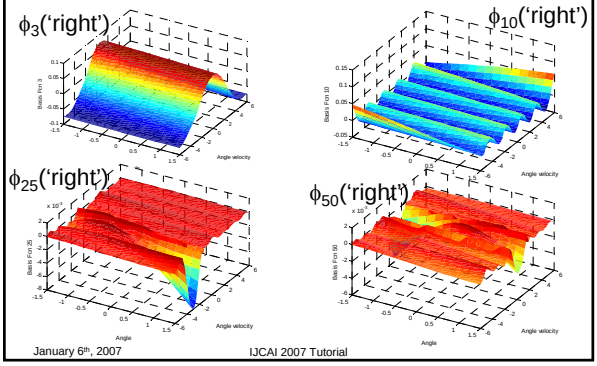
leads to the following equation

$$\sum_j w_j\,k(x,s)\,\phi(s_j) = \lambda\,\phi(x)$$

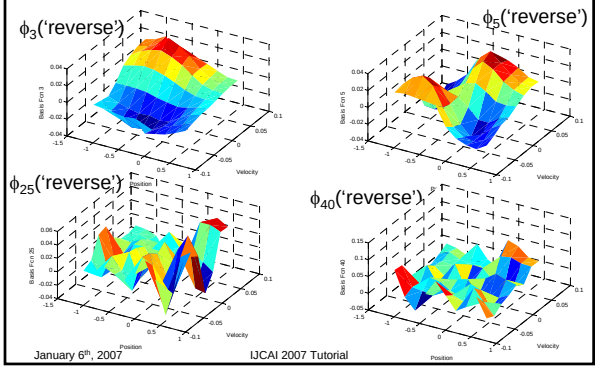
- which rewritten gives the Nystrom extension

$$\phi_m(x) = 1/\lambda_m \sum_j w_j\,k(x,s)\,\phi_m(s_j)$$

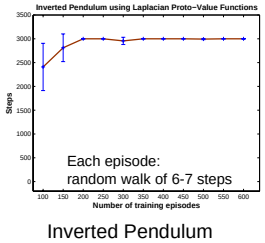
Laplacian Bases: Inverted Pendulum



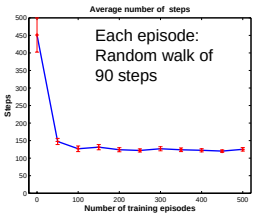
Laplacian Bases: Mountain Car



RPI in Continuous Domains

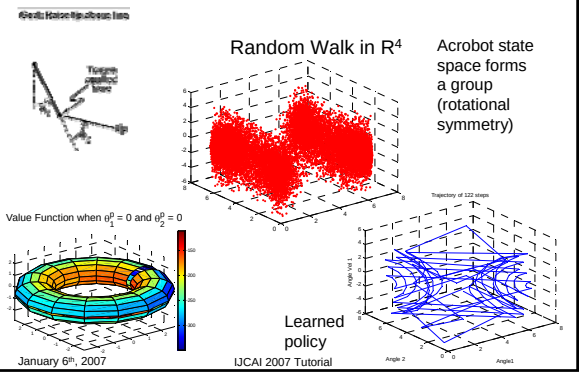


Inverted Pendulum



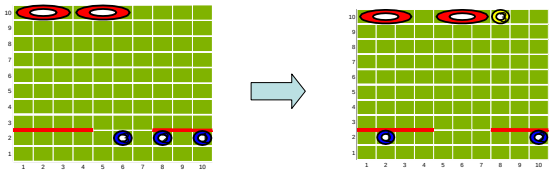
Mountain Car

Acrobot Results
(Johns, 2006)



Scaling to Large Factored MDPs: Blockers Domain
(Sallans and Hinton, JMLR 2003)

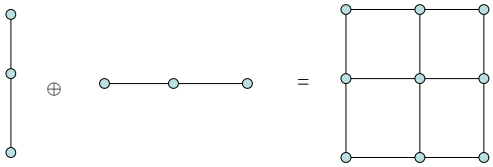
Large state space of > 10^6 states



Grid has wrap-around dynamics

Kronecker Sum Graphs

- The **Kronecker sum** of two graphs $G = G_1 \oplus G_2$ is the graph with vertex set $V = V_1 \times V_2$ and adjacency matrix $A = A_1 \otimes I_2 + I_2 \otimes A_1$
  - The Kronecker sum graph G has an edge between vertices (u,v) and (u',v') if and only if $(u,u') \in E_1$ and v=v' or (u=u') and $(v,v') \in E_2$



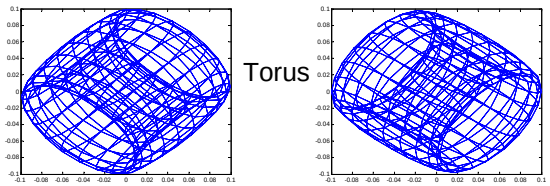
Laplacian of Kronecker Graphs

- If L_1, L_2 be the combinatorial Laplacians of graphs G_1, G_2, then the spectral structure of the combinatorial Laplacian of the Kronecker sum of these graphs $G = G_1 \oplus G_2$ is specified as

$$\sigma(L), \mathcal{X}(L)) = \{\lambda_i + \mu_j, l_i \otimes k_j\}$$

- where λ_i is the i^{th} eigenvalue of $L(G_1)$ with associated eigenvector l_i and μ_j is the j^{th} eigenvalue of $L(G_2)$ with associated eigenvector k_j.

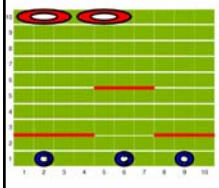
Embedding of Structured Spaces

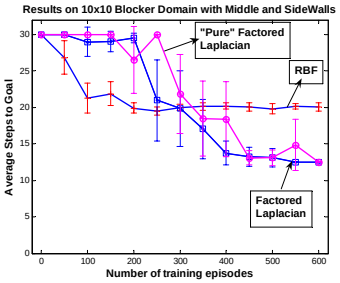


2^{nd} and 3^{rd} eigenvectors of combinatorial Laplacian

Large Factored MDP: Blockers Domain
(Mahadevan, 2006)

Topologically, this space is the tensor product of three "irregular" cylinders





Discovery of Factored Bases

- Given a weight matrix W (or Laplacian L), find matrices W_a and W_b (or L_a and L_b) such that

$$W \approx W_a \oplus W_b$$

$$L \approx L_a \oplus L_b$$

- There exists a rearrangement of W (or L) such that the best rank-k factorization can be discovered using a SVD of W' or L' [Pitsianis and van Loan, 93]

$$W' = U_W \Sigma_W V_W^T$$
$$L' = U_L \Sigma_L V_L^T$$

Decomposing Arbitrary Graphs

- Graph partitioning (Karypis and Kumar, SIAM 1998)
  - Coarsening: $|V_0| > |V_1| > \ldots |V_m|$
  - Partitioning: Divide V_m into two parts
  - Uncoarsening: Project P_m back to original graph

Laplacian Wavelet

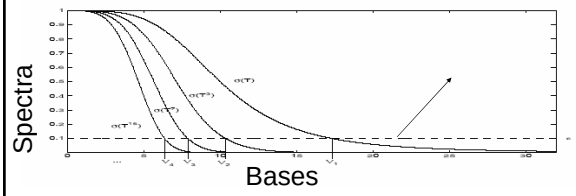
Diffusion Policy Evaluation: Fast Inversion

(Maggioni and Mahadevan, ICML 2006)

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

- Two approaches to policy evaluation:
  - DIRECT: inverting the matrix takes $O(|S|^3)$
  - ITERATIVE: successive approximation in $O(|S|^2)$
- New faster approach to policy evaluation:
  - Construct a diffusion wavelet tree from the transition matrix to invert the Green's function
  - Use the Schultz expansion to do the inversion
  - Results in a significantly faster method $\approx O(|S|)$
- Precomputation phase: task-independent
- Inversion step: reward-specific

Spectra of Powers of Diffusion Operator



If the transition matrix P is such that each random walk step takes the agent to only a few nearby states, $P^k \delta_x$ is a smooth probability cloud centered around state x, And P is contractive (has norm bounded by 1).

Weyl's theorem shows that the natural random walk on a smooth compact Riemannian manifold of dimension d has a spectral structure where the eigenvalues of powers of P decay rapidly

Schultz Expansion for Diffusion Semi-Groups

- We can model the random walk operator T as a semi-group, where the powers T^k satisfy the following conditions
  - $T^0 = I$
  - $T^{k+l} = T^k T^l$
- To compute the Green's function $(I - T)^{-1}$ associated with any diffusion semi-group, we use the Schultz expansion formula
- As a special case, we can apply this approach to policy evaluation, where T is represented by γP

$$(I-T)^{-1} f = \sum_k T^k f$$

$$S_k = \sum_{k=1}^{2^k} T^k$$

$$S_{k+1} = S_k + T^{2^k} S_k = \prod_{l=0}^{k} (I + T^{2^l}) f$$

Schultz Expansion

The starting point are the identities

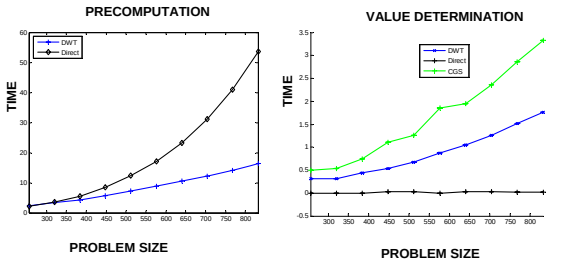
$$V^\pi = (I - \gamma P^\pi)^{-1} R = \sum_{k \geq 0} (\gamma \Pi^{-\frac{1}{2}} T^\pi \Pi^{\frac{1}{2}})^k R = \prod_{k \geq 0} (I + \gamma^{2^k} \Pi^{-\frac{1}{2}} (T^\pi)^{2^k} \Pi^{\frac{1}{2}}) R,$$

where $P^\pi = \Pi^{-\frac{1}{2}} T^\pi \Pi^{\frac{1}{2}}$, Π is the matrix whose diagonal is the asymptotic distribution of P, and R is the reward vector. The formulas hold for $\gamma \leq 1$ and R has no component in the kernel of $(I - \gamma P^\pi)$.

We have compressed in a multiscale fashion the (quasi-)dyadic powers of the operator T^π.

In many cases of interest, both the construction of the multiscale structure and the inversion take only $\mathcal{O}(|S|)$ operations!

Policy Evaluation Results



Related Work on Representation Learning in Markov Decision Processes

- Policy and reward-specific basis functions
  - "Laplacian and Krylov Basis Functions for MDPs", Petrik (IJCAI 2007)
  - "Automatic Basis Function Construction for Approximate Dynamic Programming and RL", Keller, Mannor, Precup (ICML 2006)
  - Limited to a basis for a specific policy (or value function V^{π})
- Parametric bases for linear programming
  - "Learning Basis functions in Hybrid Domains", Kveton and Hauskrecht (AAAI 2006)
  - The shape of the basis functions is human engineered
- Much ongoing research:
  - Stanford (Ng), Edinburgh (Vijaykumar), Brown (Jenkins), McGill (Precup/Mannor), Washington Univ (Smart), Texas (Stone), Duke (Parr, Carin) Rutgers (Littman), …

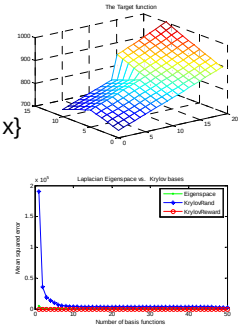
Krylov vs. Eigenvector Bases of Laplacian

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

Eigenvector Basis: {x: $Lx = \lambda x$}

Krylov Basis: {u, Lu, …, $L^n u$}

$f \sim \sum_{i \in I} \langle f, \phi_i \rangle \phi_i$



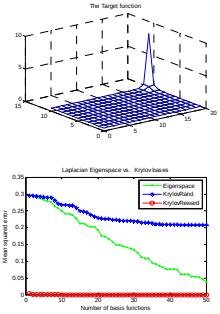
Krylov vs. Eigenvector Bases of Laplacian

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

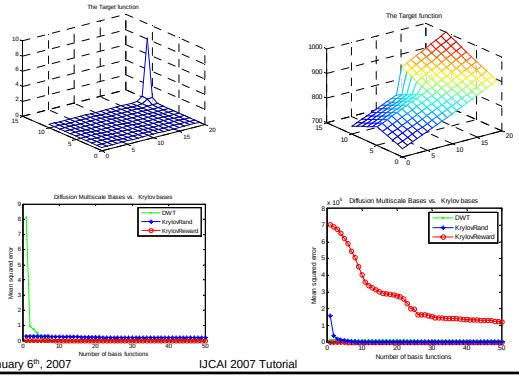
Eigenvector Basis: {x: $Lx = \lambda x$}

Krylov Basis: {u, Lu, …, L^n u}

$f \sim \sum_{i \in I} \langle f, \phi_i \rangle \phi_i$

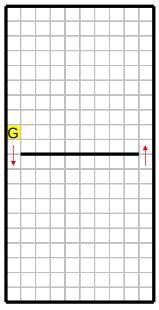


Krylov vs. Diffusion Wavelet Bases



Directed Two-Room Environment (Johns, 2006)

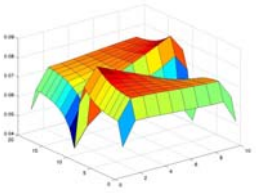
- This domain was used to compare the bases functions from the undirected Laplacian vs. the directed Laplacian
  - Two 10x10 rooms with two directed edges (all other edges are undirected)
  - Four stochastic actions, zero reward unless in goal state (+100)
  - Discount factor of 0.9



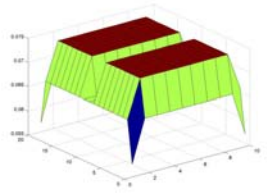
Directed vs. Undirected Laplacian

- The first eigenvector of the normalized Laplacian shows the difference directionality makes on the steady-state distribution

Directed



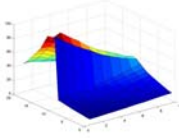
Undirected



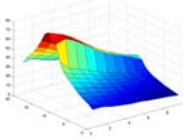
Results: Directed vs. Unidirected Laplacian (Johns, 2006)

- The undirected Laplacian results in a poorer approximation because it ignores directionality

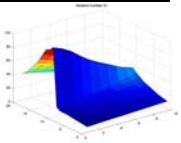
Exact VF



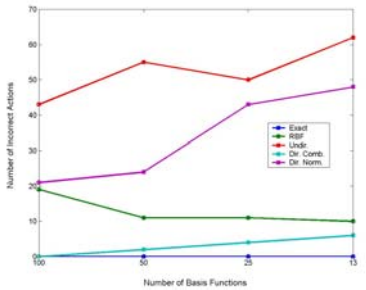
Undirected



Dir. Combinatorial



Comparison of Undirected vs. Directed Laplacians



Ongoing Extensions

- Approximation over temporally extended actions
  - Sarah Osentoski
- Laplacian bases on directed graphs
  - Jeff Johns
- Transfer of basis functions across domains
  - Kim Ferguson (ICML Workshop, 2006)
- Discovery of temporally extended actions
  - Andrew Stout
- Combining Laplacian and Krylov basis functions
  - Marek Petrik (IJCAI 2007)

Challenges and Future Directions

- Large-scale applications
  - Fourier and wavelet bases for high-dimensional continuous control tasks (e.g. humanoid robots)
  - Applications to graphics: compression, watermarking, animation
- Convergence and theoretical analysis
  - Can Laplacian and DWT bases be shown to be "optimal" in some interesting sense? [Ben-Chien and Gotsman, 2005]
  - Convergence of RPI
- Application of this approach to related problems
  - POMDPs: value function is highly compressible!
  - PSRs: low-rank approximation of dynamical systems
  - Relational and factored MDPs: exploit group structure of underlying graphs

Factored and Relational MDPs

- Much work on factored and relational MDPs and RL
  - [Koller and Parr, UAI 2000; Guestrin et al, IJCAI 2003; JAIR 2003]
  - [Fern, Yoon, and Givans, NIPS 2003]
  - ICML 2004 workshop on relational RL
- Fourier and wavelet bases over relational representations
  - Symmetries and group automorphisms
  - Non-commutative harmonic analysis

Further Reading (www.cs.umass.edu/~mahadeva)

- Fourier bases (Laplacian eigenfunctions)
  - Sridhar Mahadevan, "Samuel Meets Amarel: Automating Value Function Approximation using Global State Space Analysis", Proceedings of the National Conference on Artificial Intelligence (AAAI-2005), Pittsburgh, PA, July 9-13, 2005.
  - Sridhar Mahadevan, "Representation Policy Iteration", Proceedings of the 21st Conference on Uncertainty in AI (UAI-2005), Edinburgh, Scotland, July 26-29, 2005.
  - Sridhar Mahadevan, "Proto-Value Functions: Developmental Reinforcement Learning", Proceedings of the International Conference on Machine Learning (ICML-2005), Bonn, Germany, August 7-13, 2005.
- Wavelet bases
  - Sridhar Mahadevan and Mauro Maggioni, "Value Function Approximation using Diffusion Wavelets and Laplacian Eigenfunctions", Neural Information Processing Systems (NIPS) conference, Vancouver, December, 2005.
- Fast policy evaluation
  - Mauro Maggioni and Sridhar Mahadevan, "Fast Direct Policy Evaluation Using Multiscale Markov Diffusion Processes", University of Massachusetts, Department of Computer Science Technical Report TR-2005-39, 2005 (also accepted to ICML 2006)