Regularized Feature Selection in Reinforcement Learning

Dean Wookey

Joint work with George Konidaris

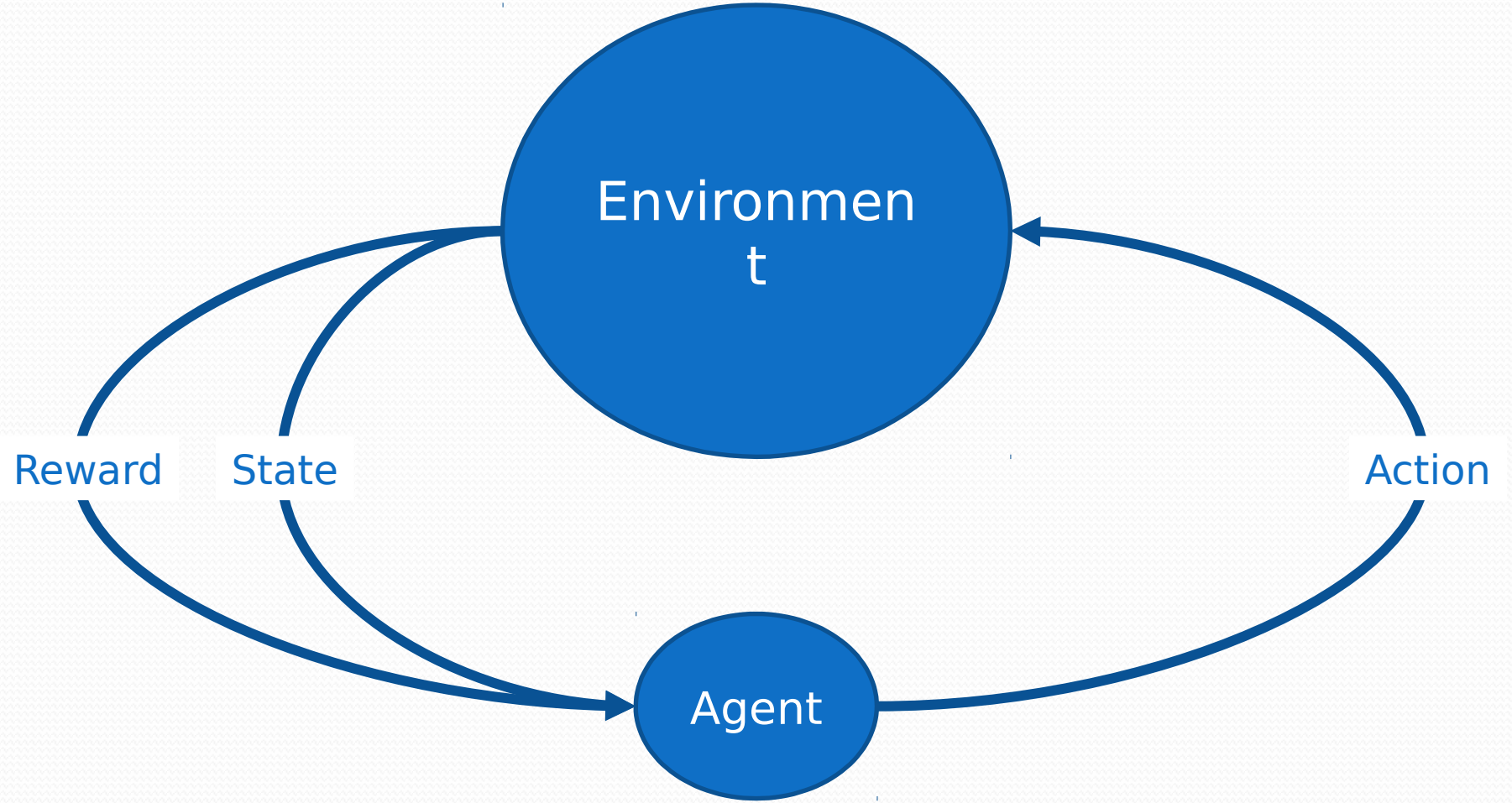
Dean S. Wookey
School of Computer Science and Applied Mathematics
University of the Witwatersrand, Johannesburg
E-mail: dean.wookey@students.wits.ac.za

George D. Konidaris
Departments of Computer Science & Electrical and Computer Engineering
Duke University, Durham, NC 27708
E-mail: gdk@cs.duke.edu

Roadmap

- Background
  - Reinforcement Learning
  - Value Functions
  - Linear Function Approximation
  - Basis Functions
  - Feature Selection
  - Bellman Error
- Value Function Smoothness
- STOMP-TD and SSOMP-TD
- Experiments
- Future Work
- Conlusion

Reinforcement Learning

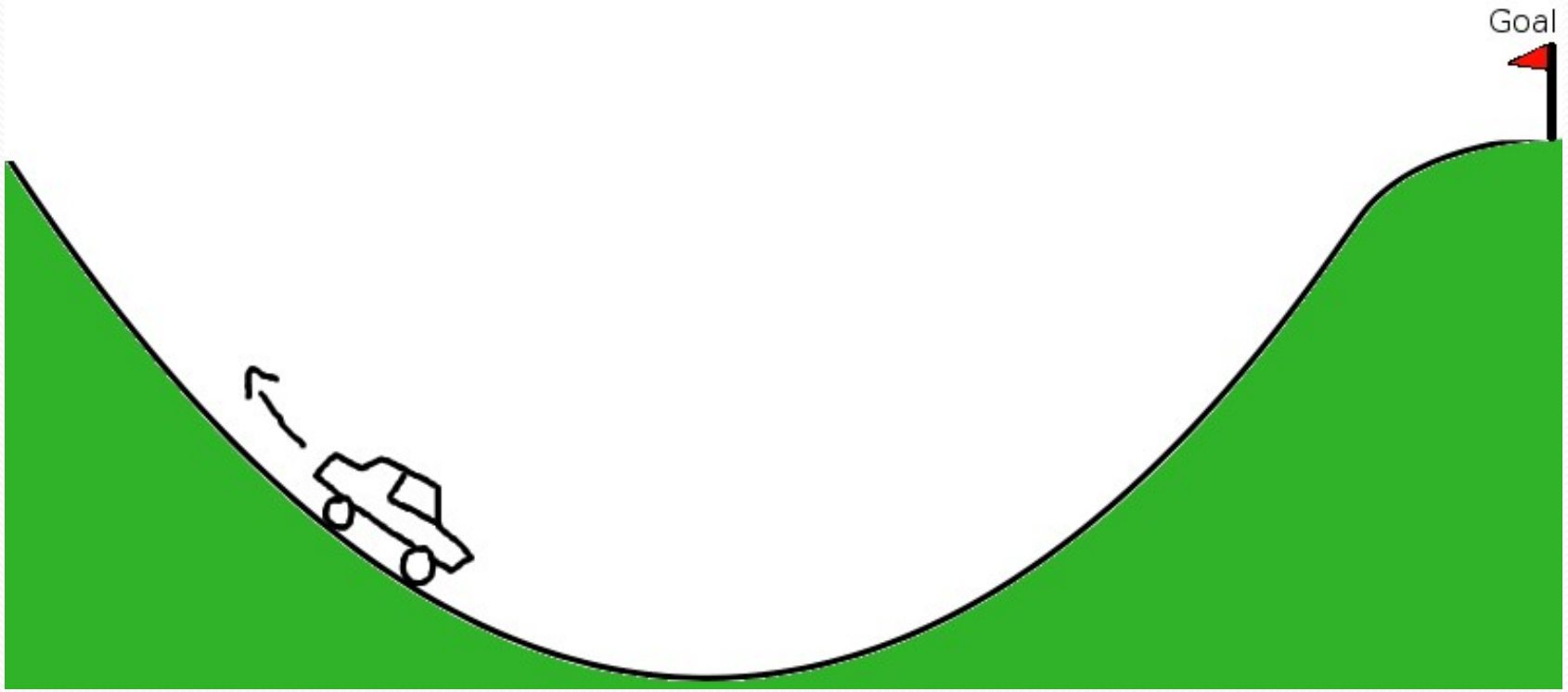


Preliminaries

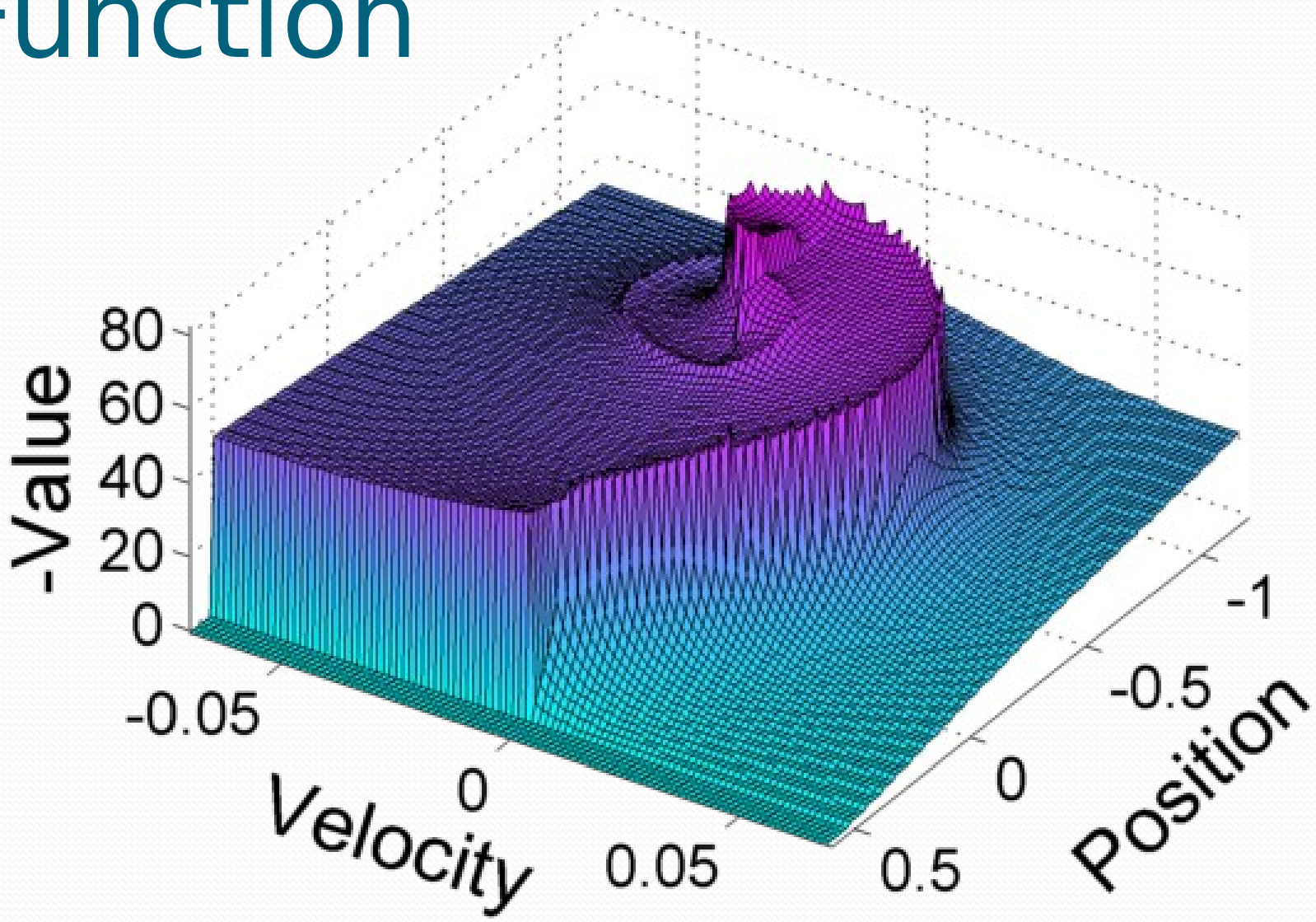
- s is the current state, s' is the following state.
- r is the reward you get when transitioning to that state.
- $\gamma \in [0,1)$ is the discount factor.
- π is the policy.

Value Function

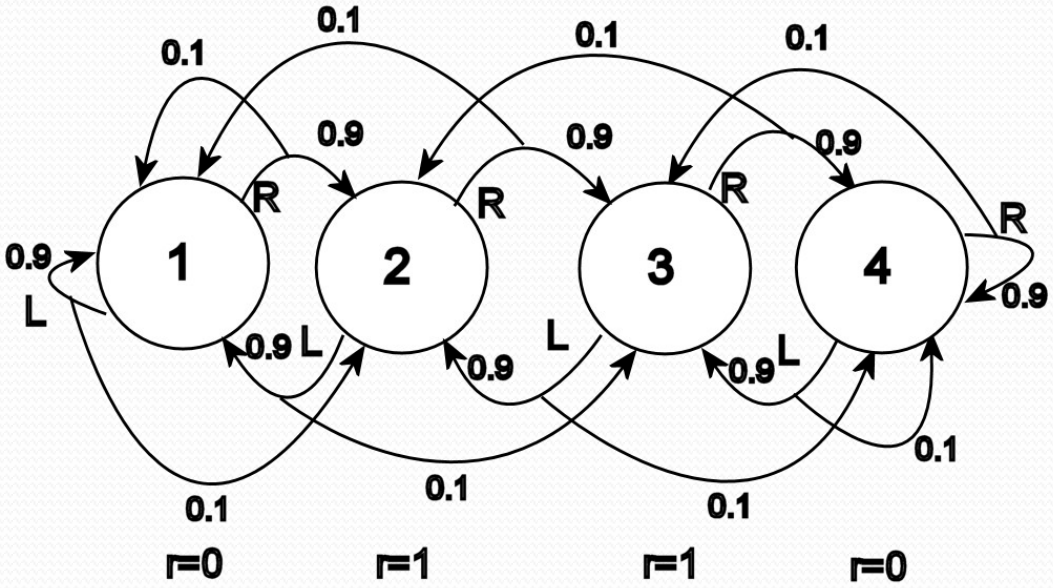
- A value function maps a state to a numerical value representing the ‘goodness’ of that state.



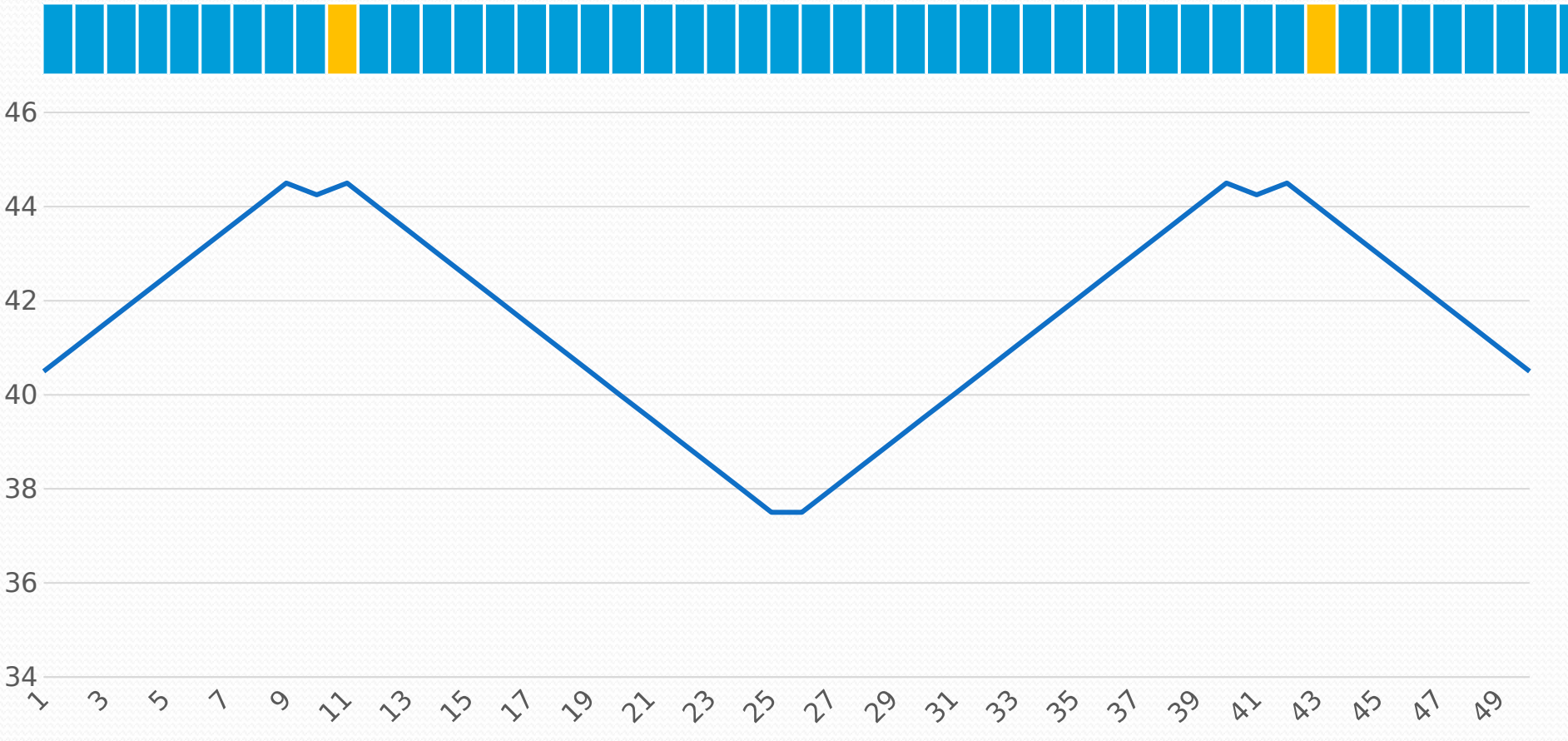
Mountain Car Value Function



50 State Chainwalk



50 State Chainwalk Value Function

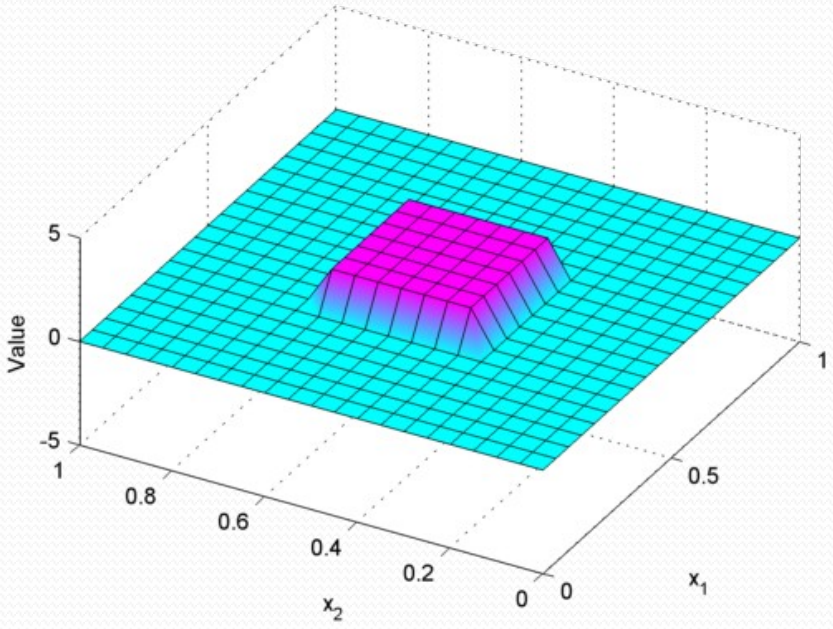


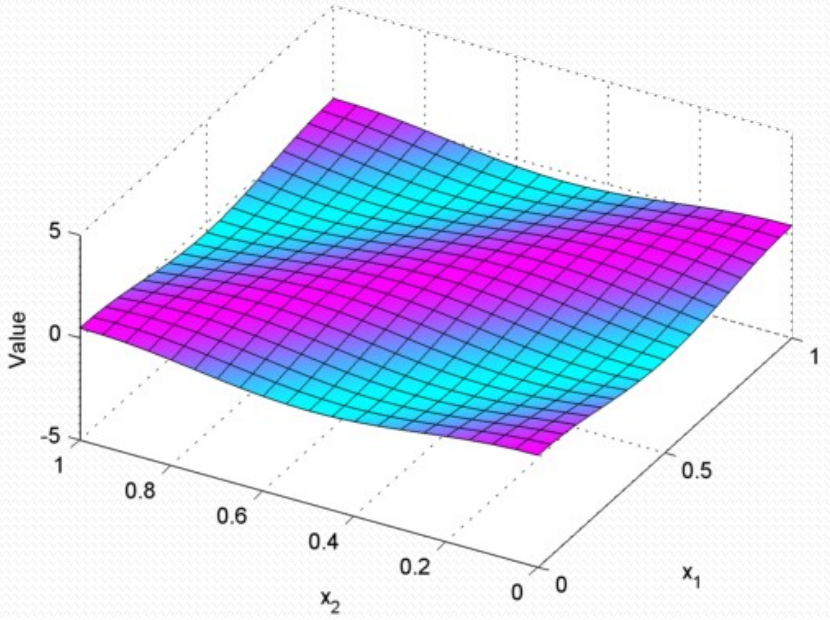
Linear Function Approximation(LFA)

- Used when state spaces are large or continuous.
- Value stored as weighted sum of basis functions:
  - $V(s) = w_0 + w_1 \varphi_1(s) + w_2 \varphi_2(s) + \ldots + w_n \varphi_n(s)$

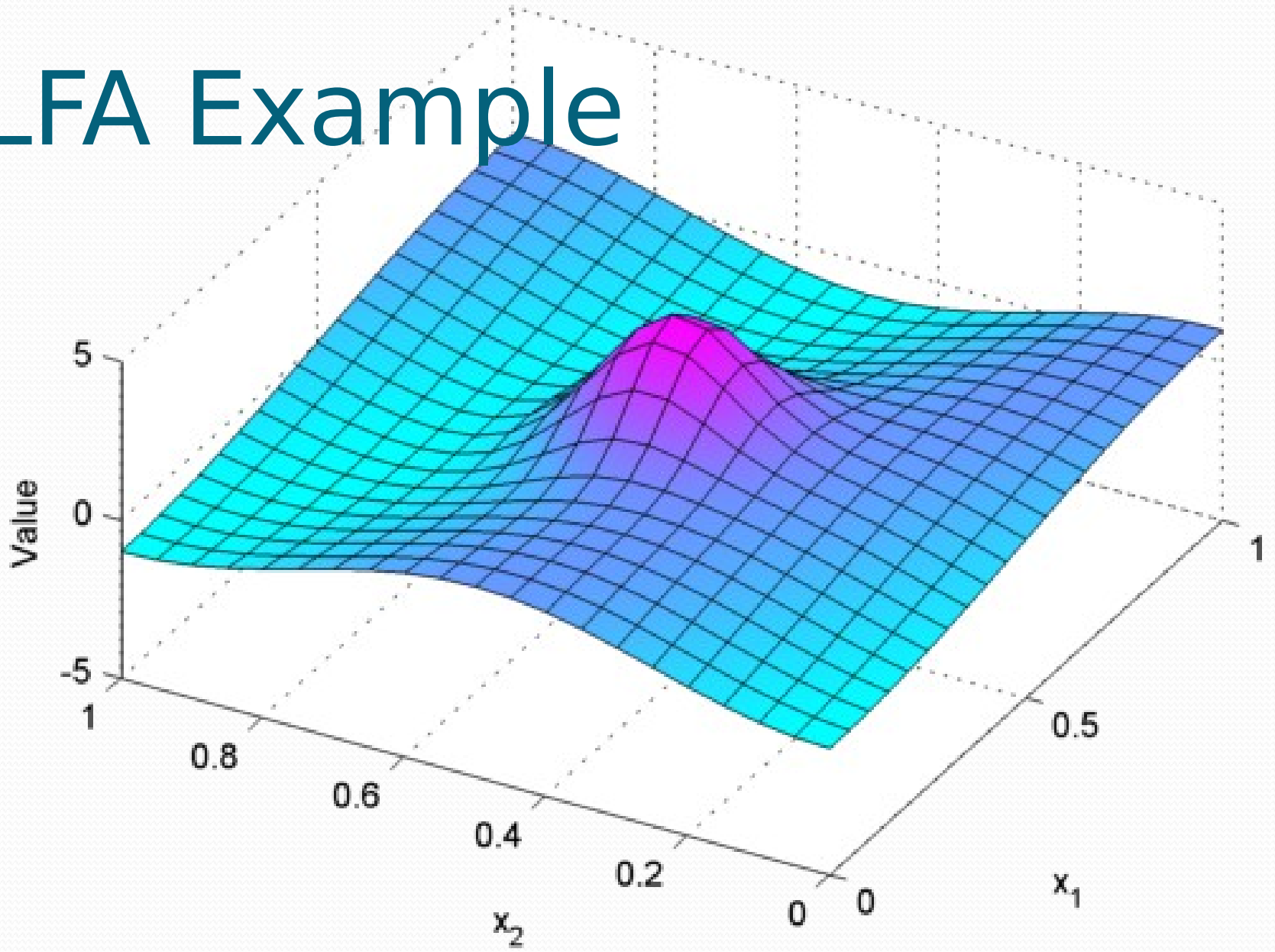
Basis Functions

- Functions of the state, φ(s).





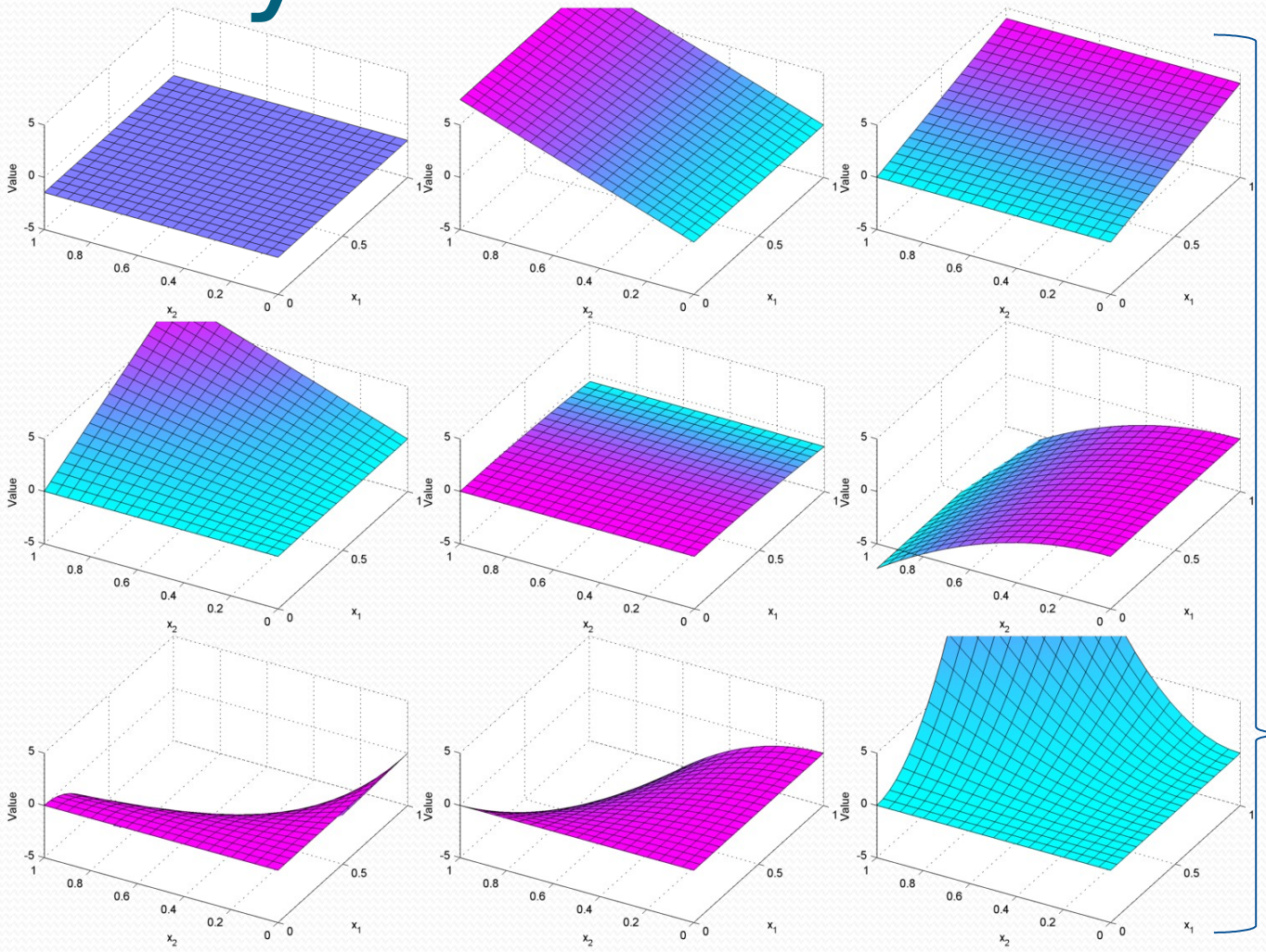
LFA Example



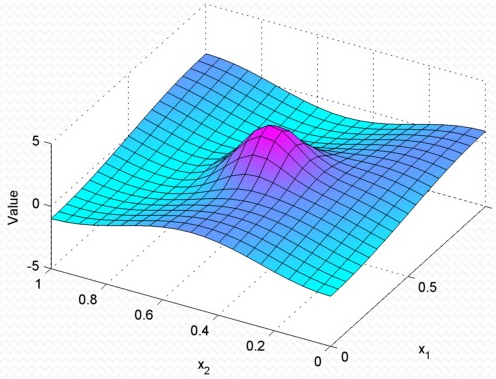
Polynomial

- Basis functions are polynomial terms.
- $\varphi(s) = x_1^a x_2^b$
- $V(s) = w_0 + w_1 \varphi_1(s) + w_2 \varphi_2(s) + \ldots + w_n \varphi_n(s)$
- Approximation of synthetic function:
  - $V(s) = -1.47 + 7.43x_2 + 3.99x_1 + 7.43x_1x_2 - 0.72 x_1^2 - 7.37x_2^2 - 27.37 x_1^2x_2 - 14.08 x_1x_2^2 + 34.02x_1^2x_2^2$

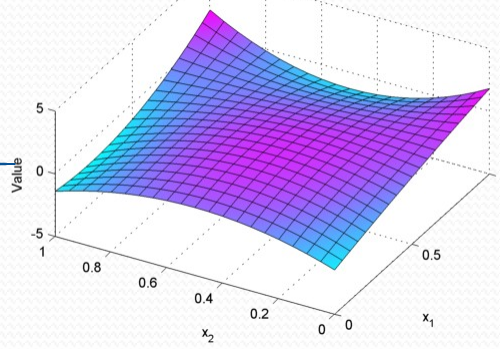
Polynomial



Target Function



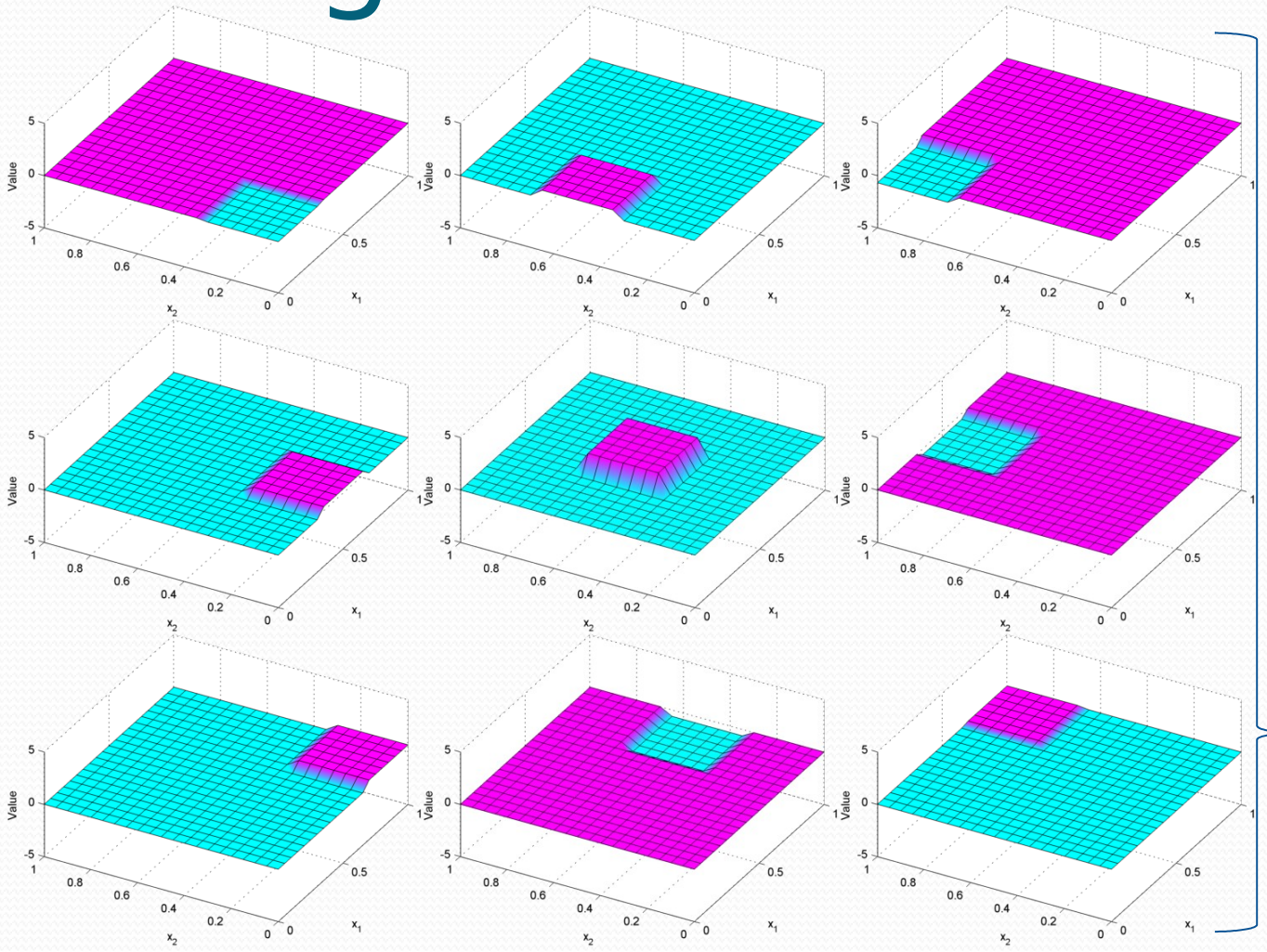
Approximation



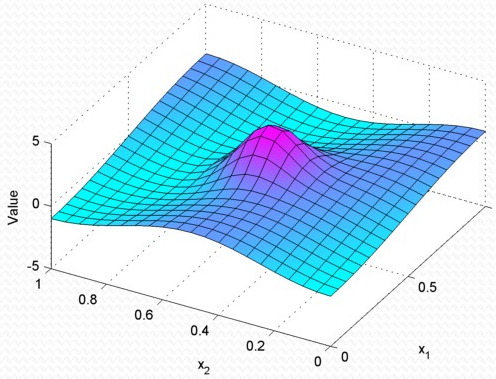
Tilings

- State is discretised into ‘tiles’.
- Each tile function φ(s) outputs a value of 1 if s lies within the tile boundaries, else it outputs 0.
- More tiles leads to more accurate value functions, but less generalisation between states.

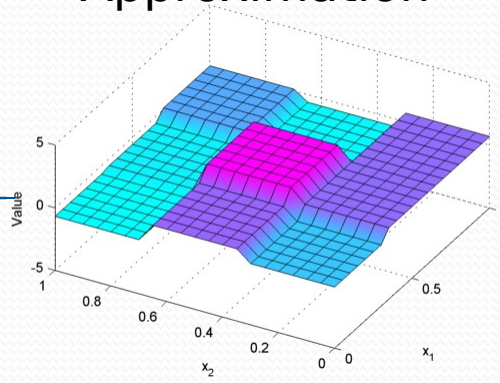
Tilings



Target Function



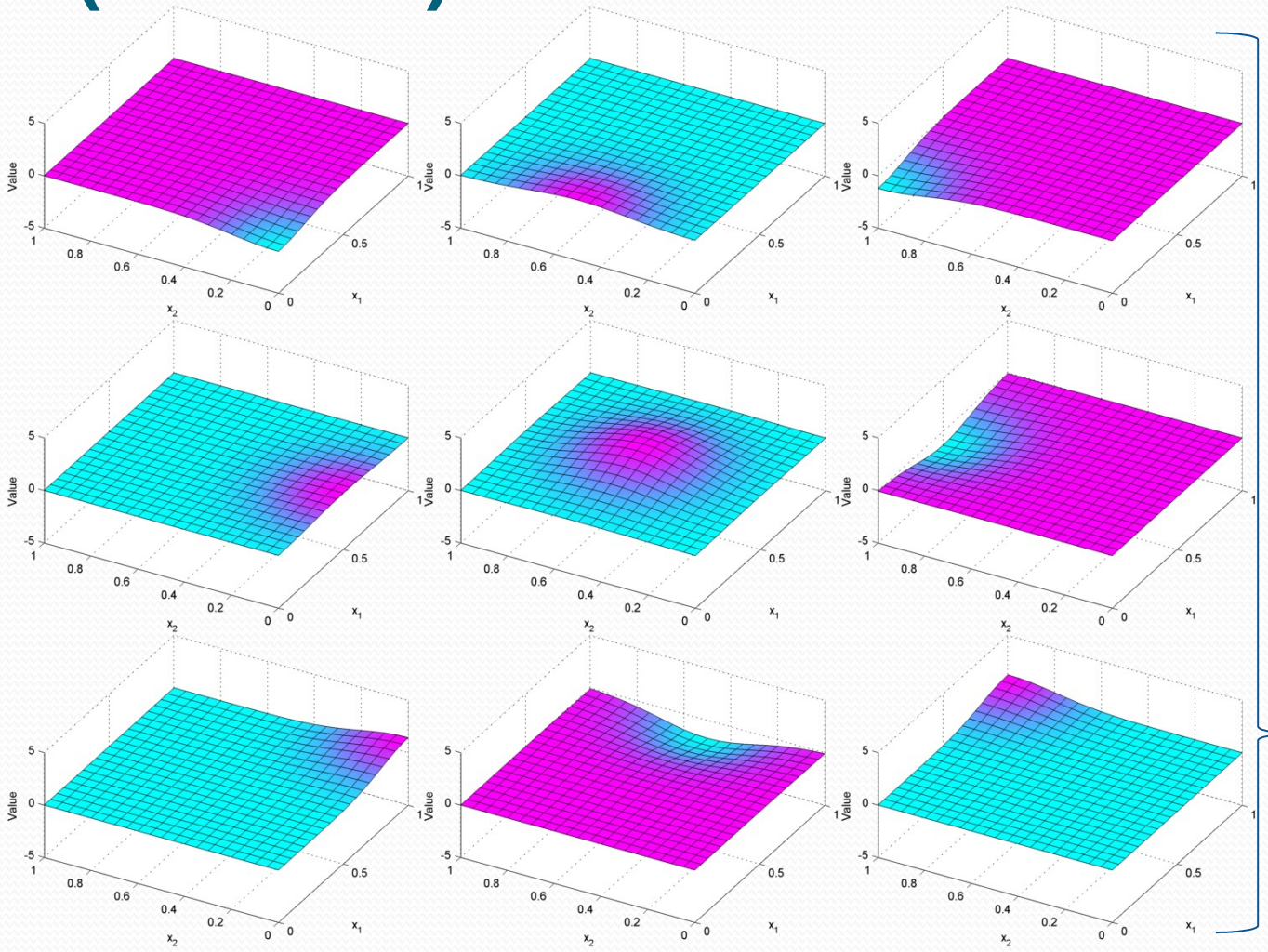
Approximation



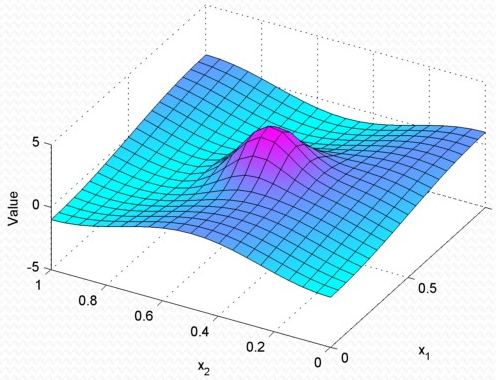
Radial Basis Functions (RBFs)

- Gaussian functions with different centres c and variance σ tiled across the state space.
- $\varphi(s) = \frac{1}{\pi^{d/4}\sigma^{d/2}} e^{-\|c-s\|^2/2\sigma^2}$

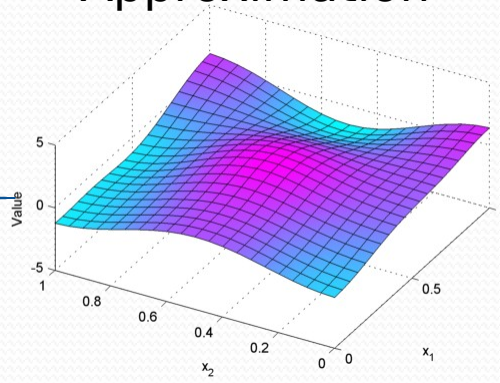
Radial Basis Functions (RBFs)



Target Function



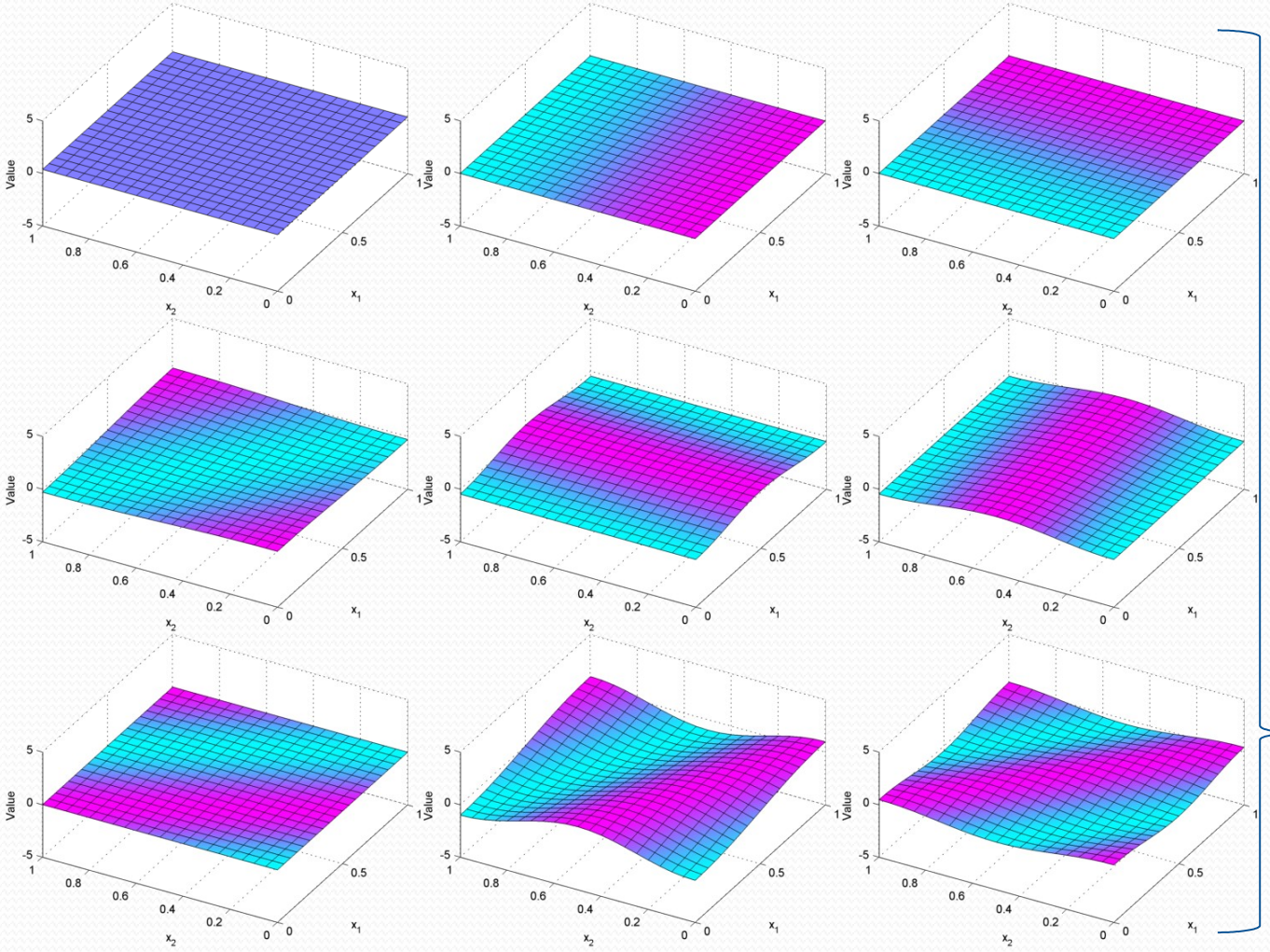
Approximation



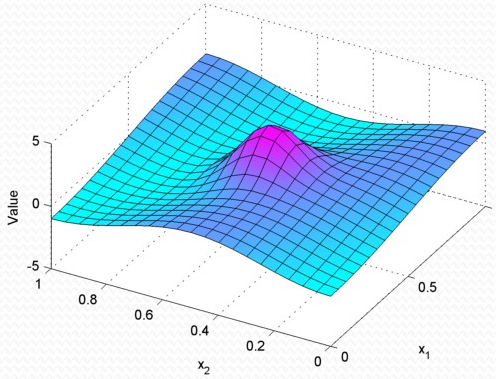
Fourier

- Cosine functions of different integer frequencies c.
- $\varphi(s) = \cos(\pi c \cdot s)$
- $V(s) = w_0 + w_1 \varphi_1(s) + w_2 \varphi_2(s) + \ldots + w_n \varphi_n(s)$
- Approximation of synthetic function:
  - $V(s) = 0.38 + 0\cos(\pi x_1) + 0\cos(\pi x_2) + 0.29\cos(\pi x_1 + \pi x_2) - 0.48\cos(2\pi x_1) - 0.48\cos(2\pi x_2) + 0\cos(2\pi x_1 + \pi x_2) - 1\cos(\pi x_1 + 2\pi x_2) + 0.47\cos(2\pi x_1 + 2\pi x_2)$

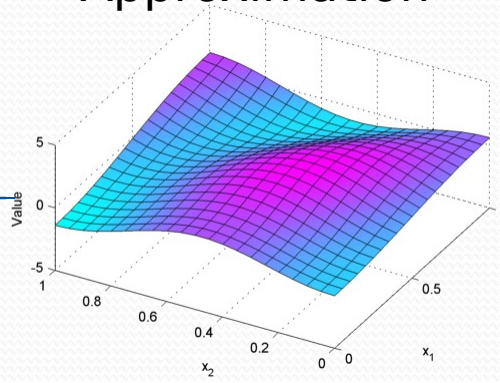
Fourier



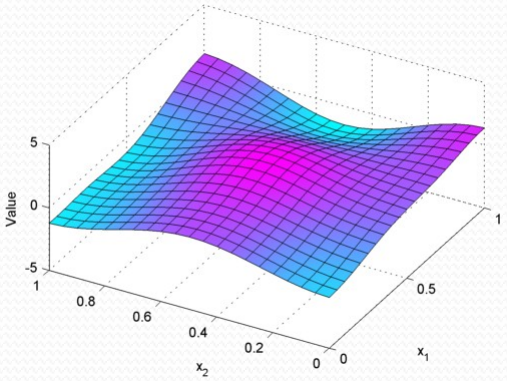
Target Function



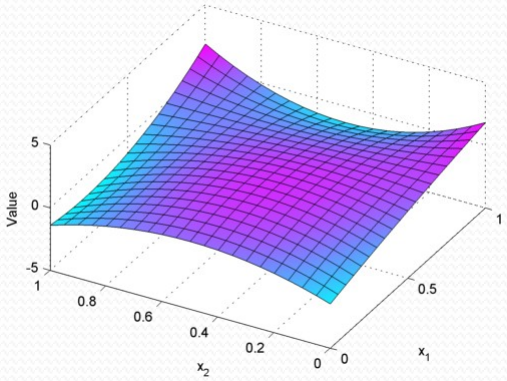
Approximation



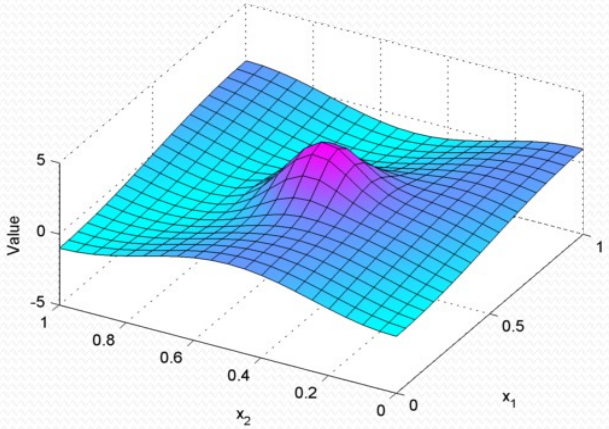
Basis Comparison



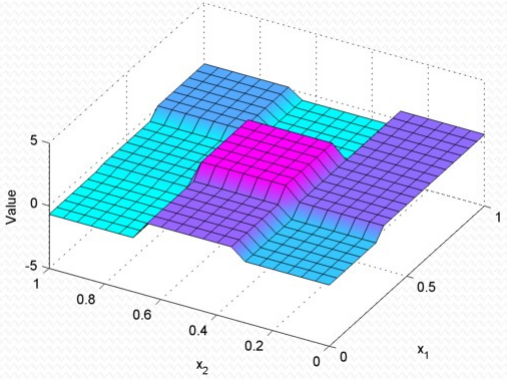
RBF



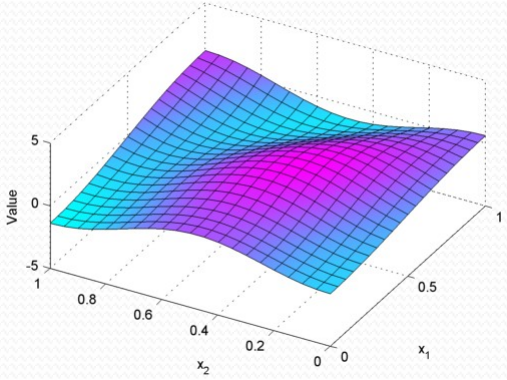
Polynomial



Target Function



Tiling



Fourier

Feature Selection

- Infinitely many basis functions to choose from for use in LFA.
- Three main strategies:
  - Construct features as needed.
  - Choose a fixed basis set.
  - Select features from a large dictionary.

Orthogonal Matching Pursuit (OMP)

- Batch algorithm for regression with feature selection.
- Samples are collected and a dictionary of candidate basis functions is created.
- At each step the basis function most correlated with the error is added to the LFA and the weights recalculated using least squares.

OMP-TD

- Similar to OMP
  - Samples (s, r, s’) are collected with a fixed policy.
  - The basis function most correlated with the Bellman error at each step is selected.
  - Least squares temporal difference (LSTD) is used to fit the weights at each step.

Bellman Error as a Selection Metric

- Residual Bellman Error: $R + \gamma\phi(s')w - \phi(s)w$
- Correlation $\rho_i = \frac{<R,\phi_i>}{||\phi_i||}$
- As $\gamma \to 1$ the value function becomes smoother and smoother and less reward function less.

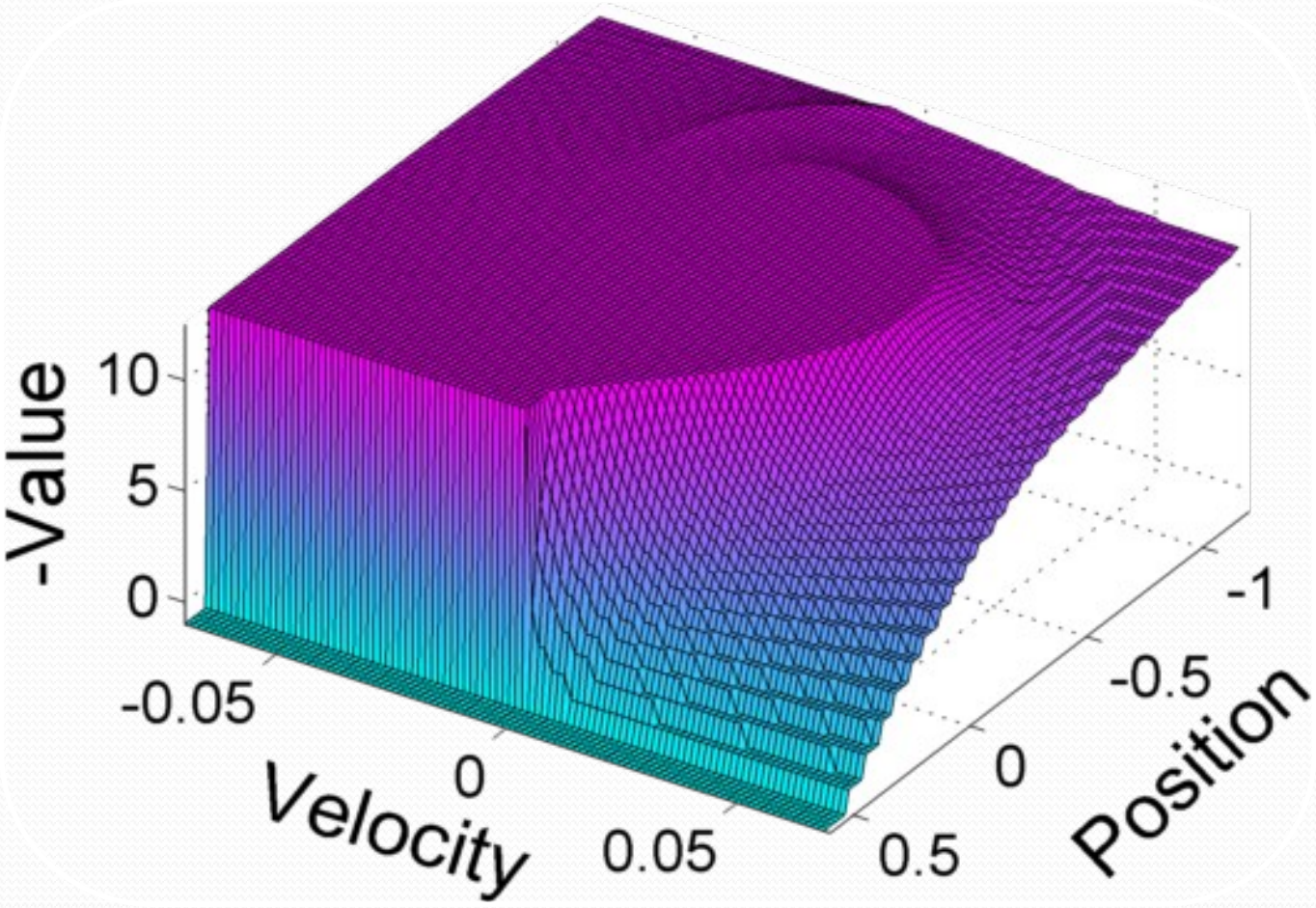
Mountain Car Smoothness

Mountain Car Smoothness

Mountain Car Smoothness

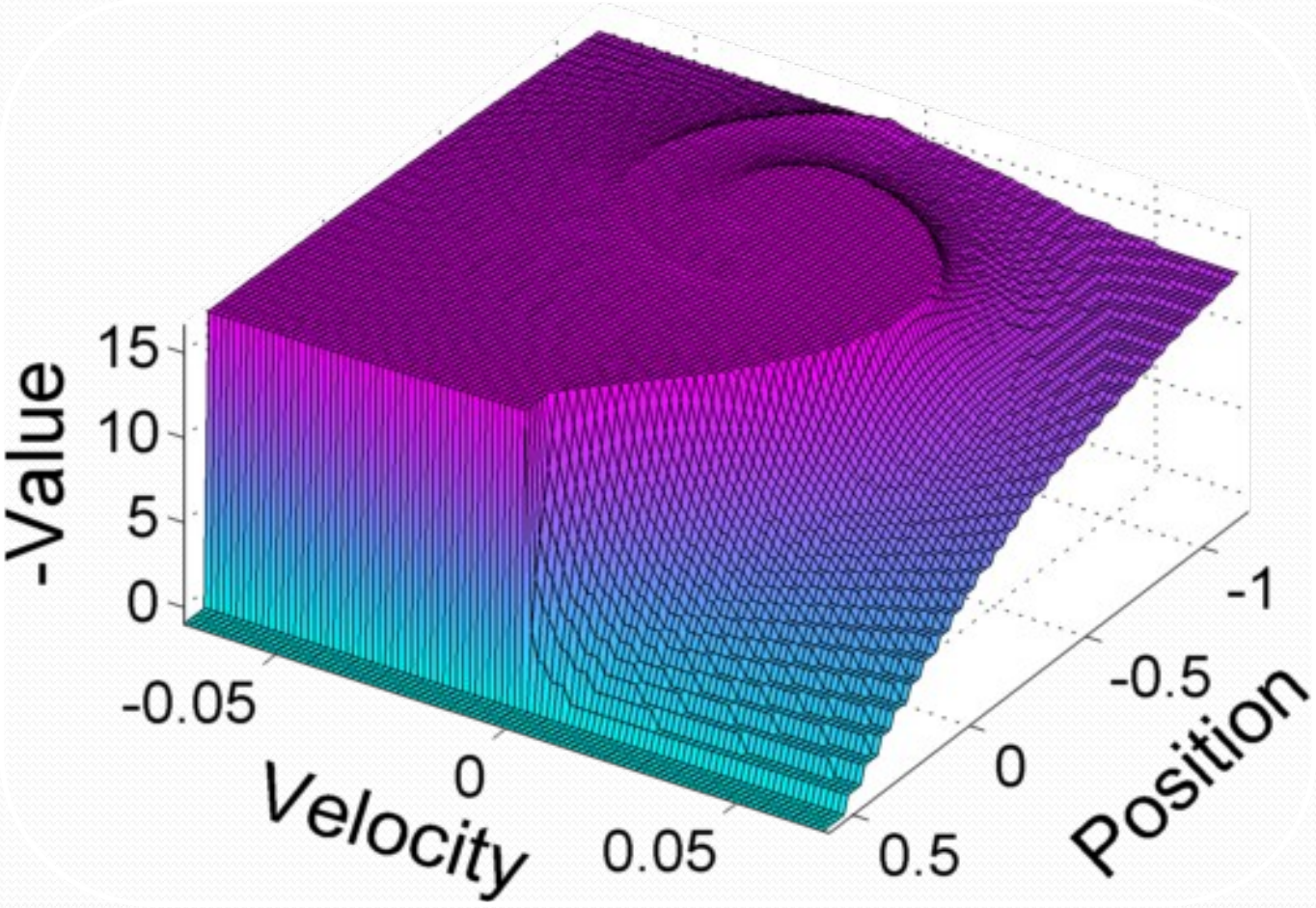
Mountain Car Smoothness

Mountain Car Smoothness



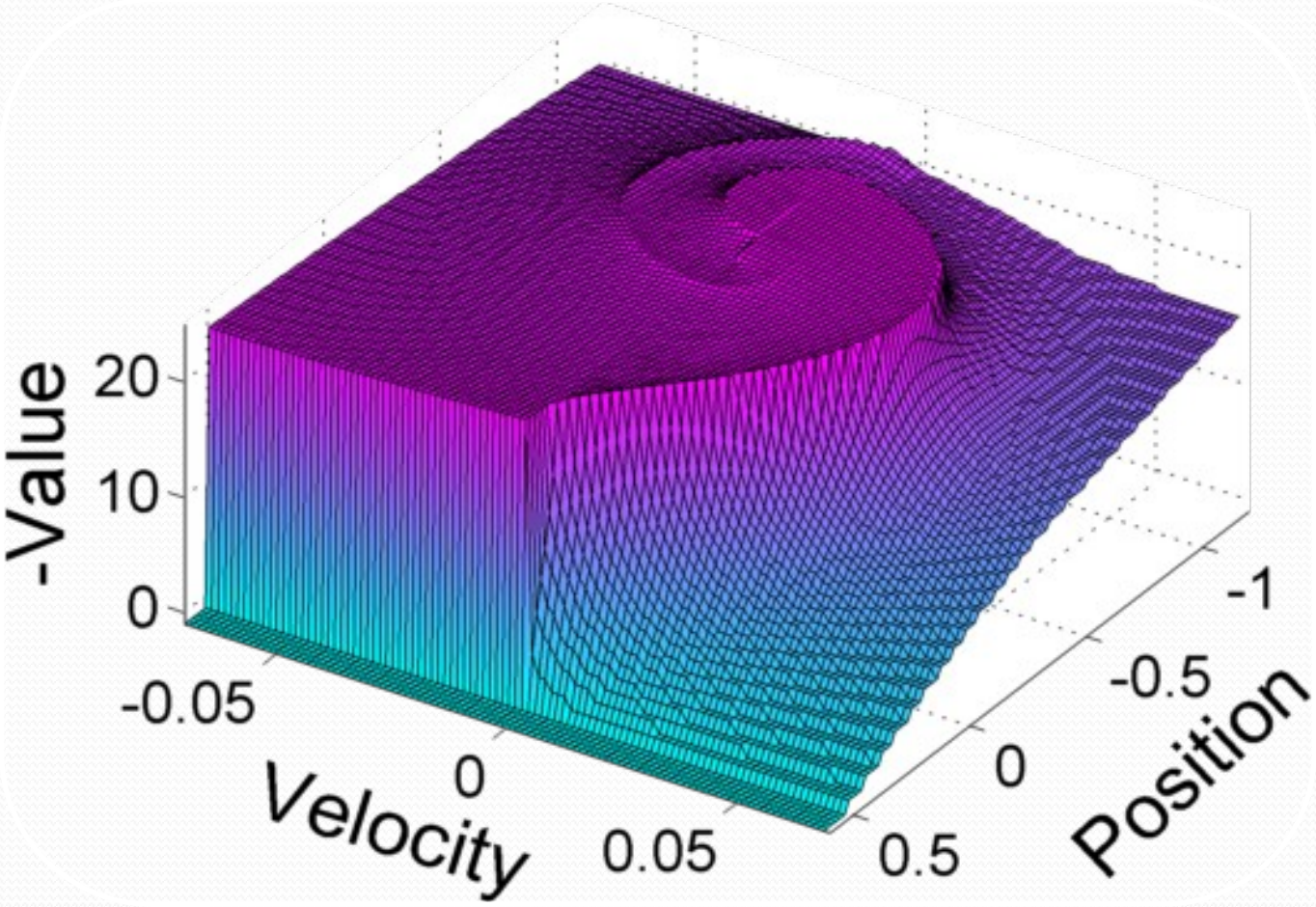
γ =0.92

Mountain Car Smoothness



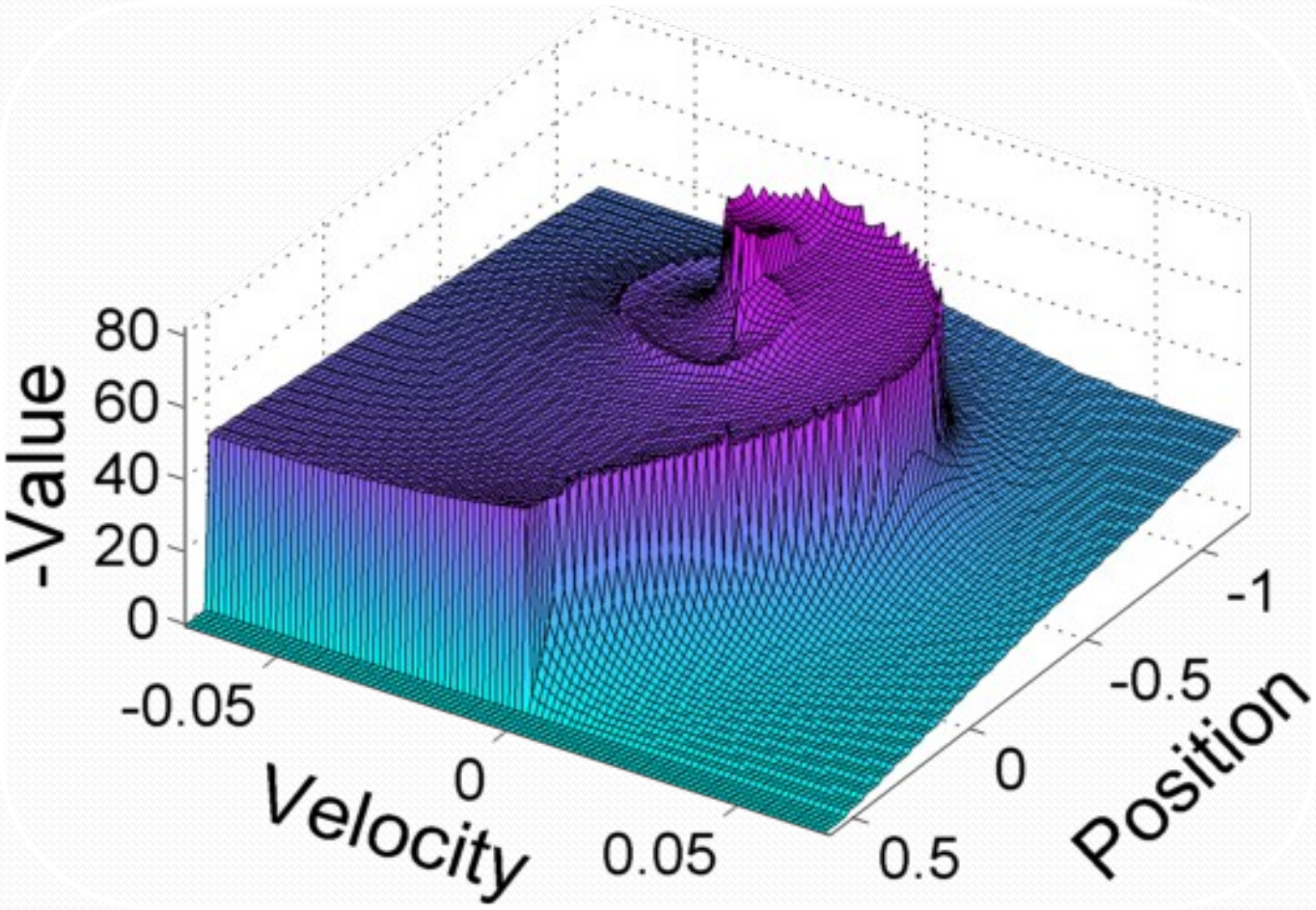
γ =0.94

Mountain Car Smoothness



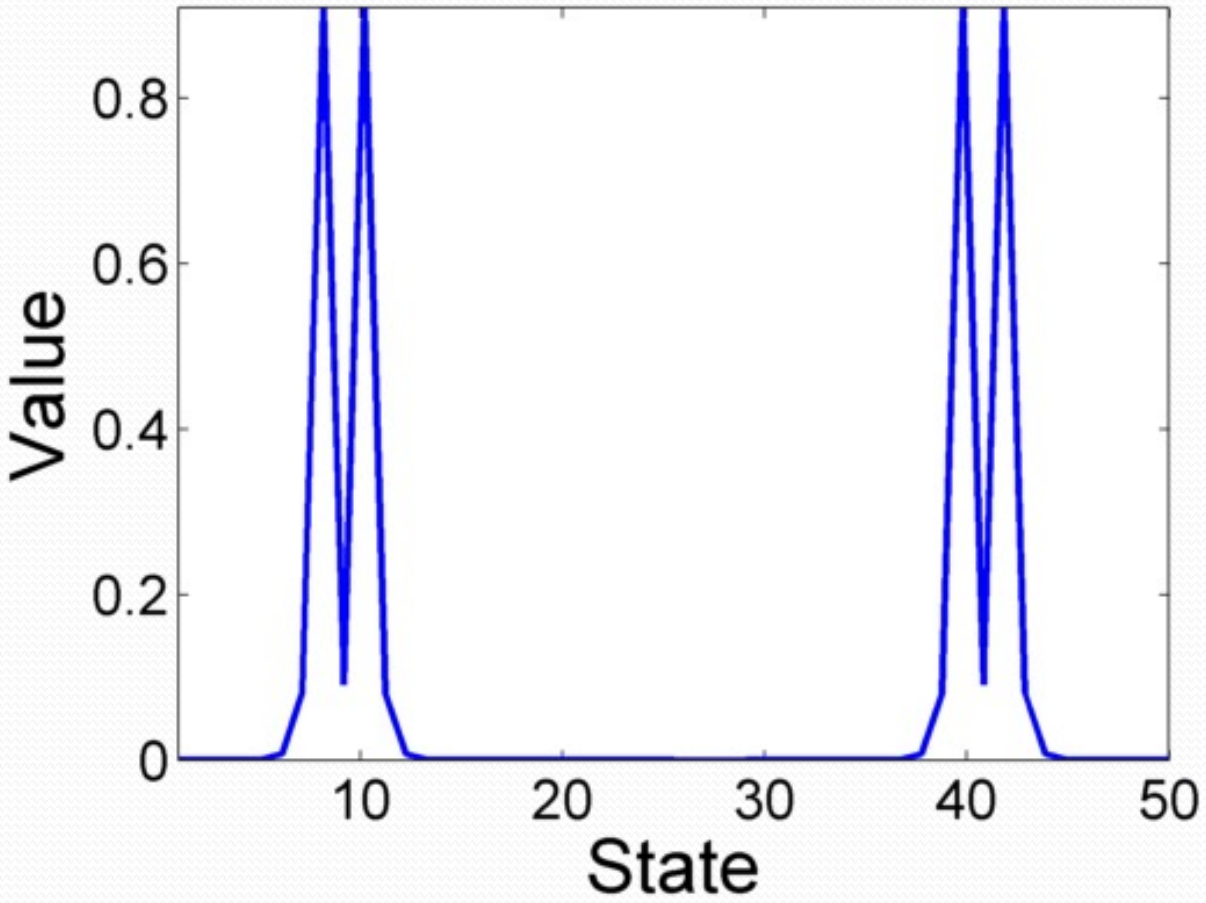
γ =0.96

Mountain Car Smoothness



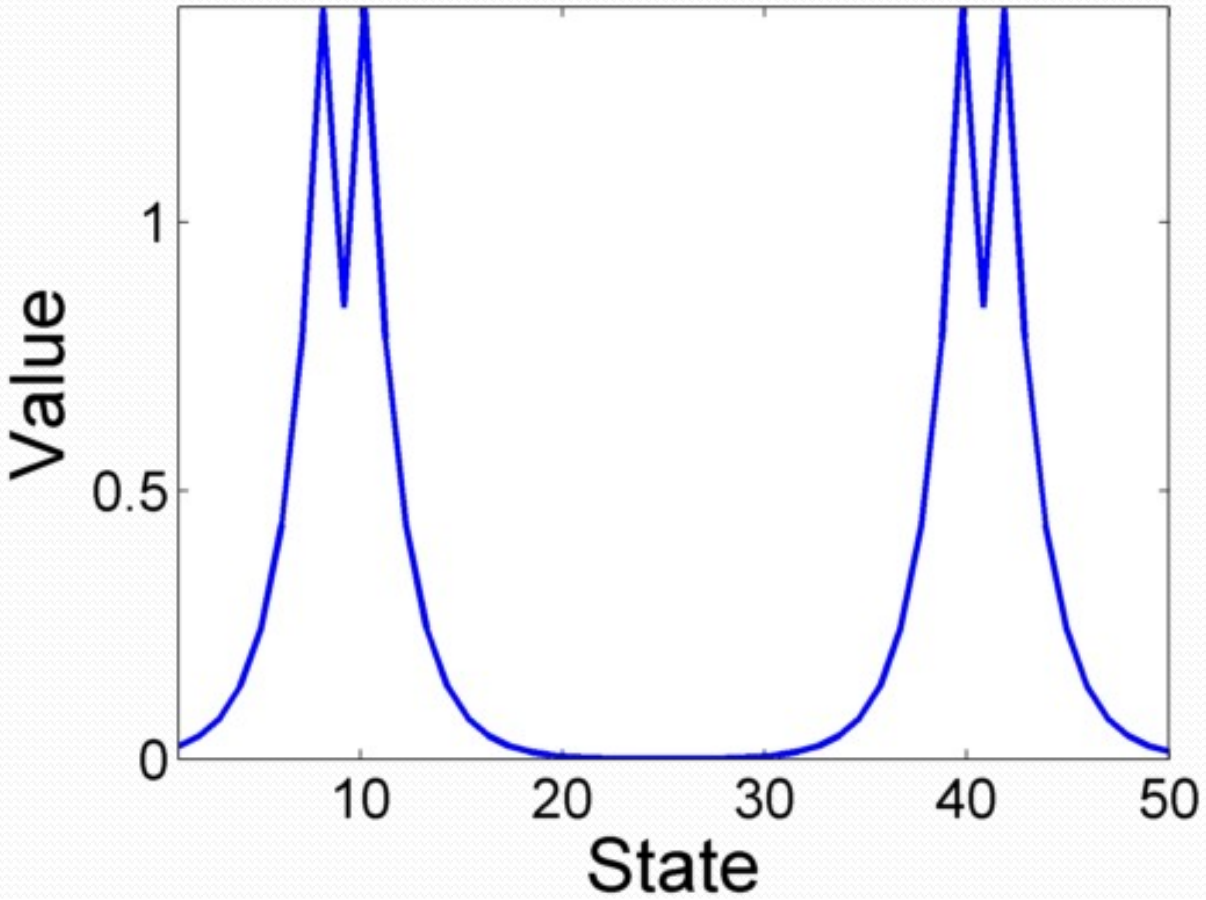
$\gamma = 0.99$

Chainwalk Smoothness



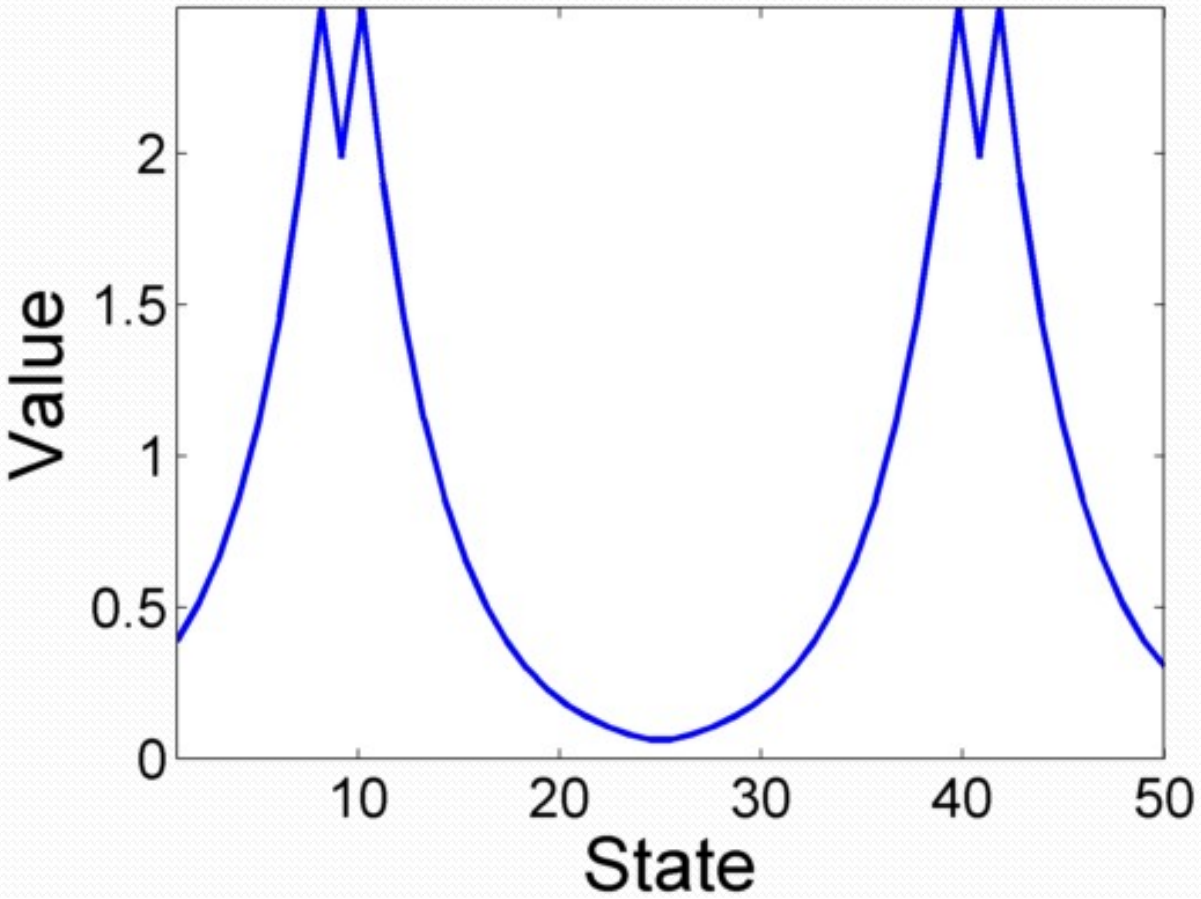
γ=0.1

Chainwalk Smoothness



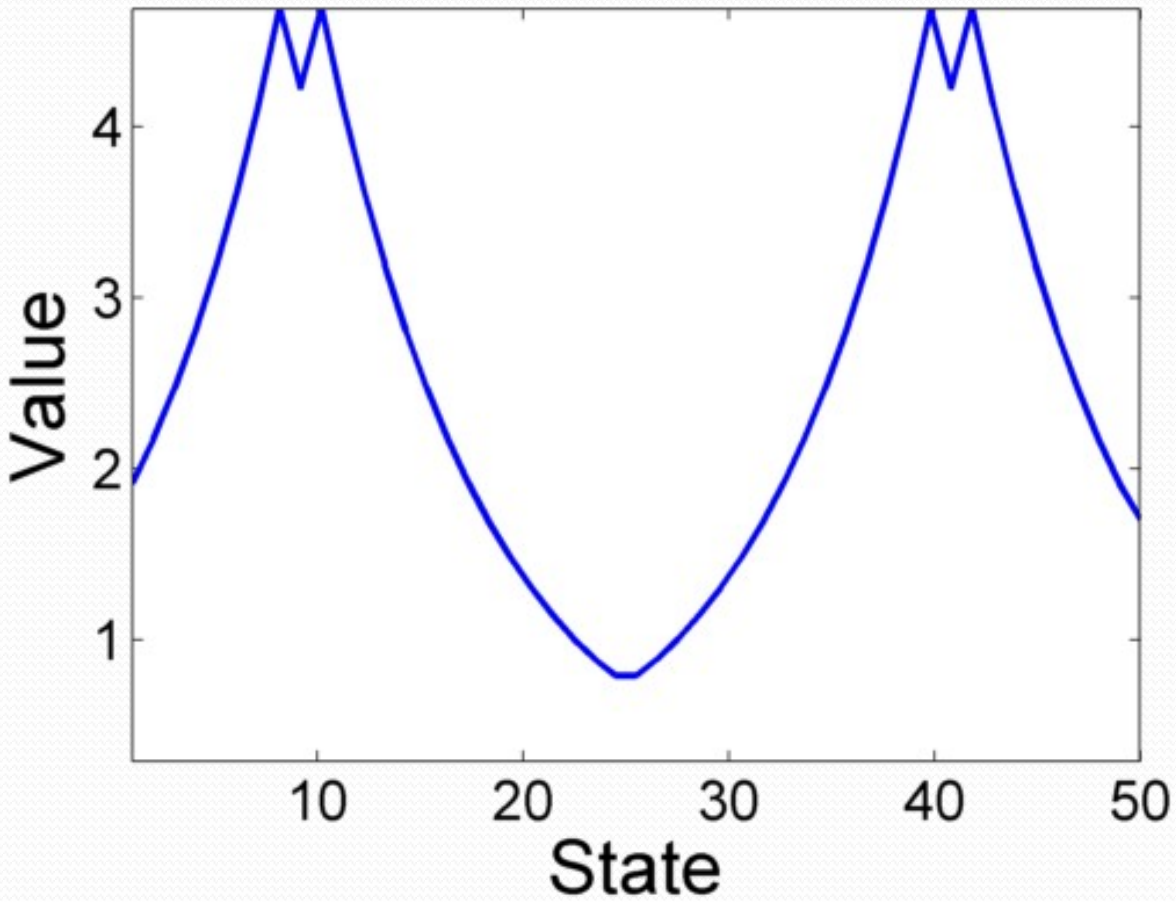
γ=0.6

Chainwalk Smoothness



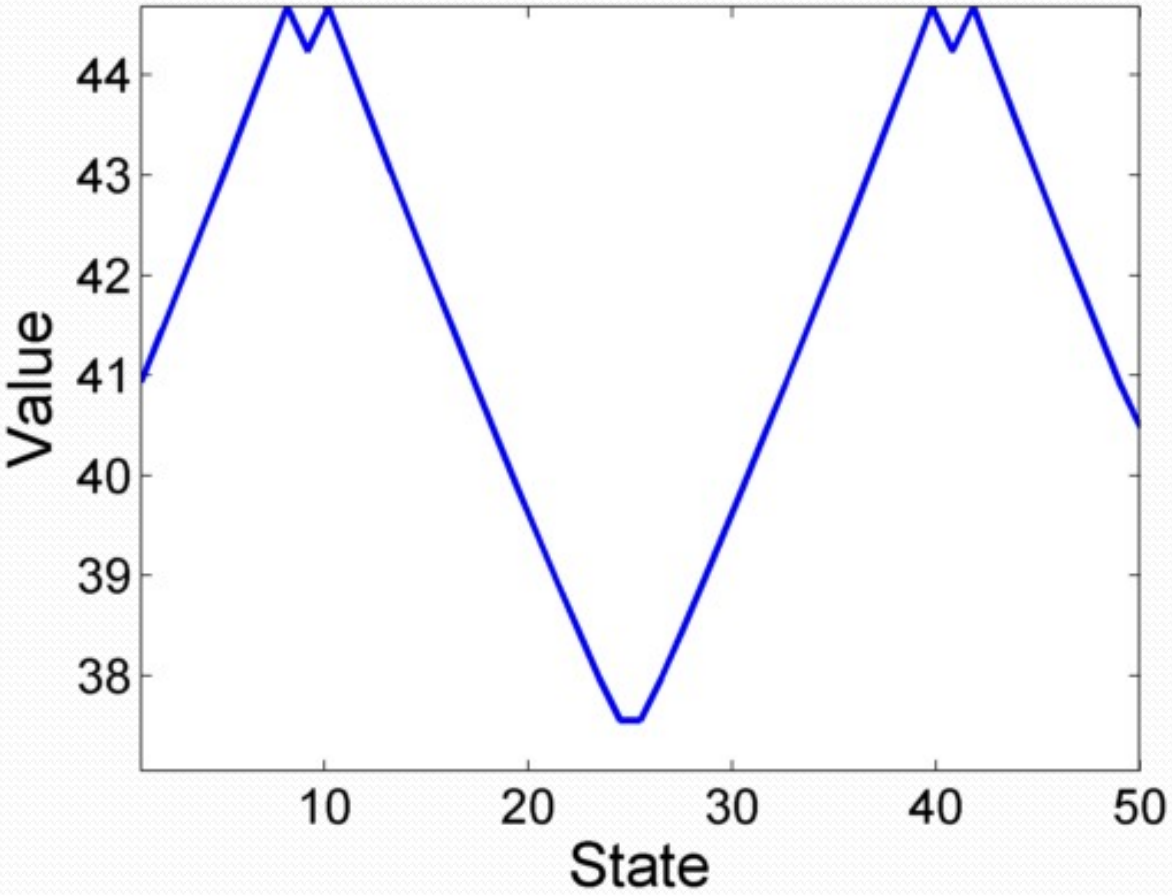
γ=0.8

Chainwalk Smoothness



γ=0.9

Chainwalk Smoothness



γ=0.99

Smoothness in Feature Selection

- Normally would be added in via regularization.
- Standard regularization techniques regularize the fit.
- We want to regularize the selection of basis functions.

Smoothness

- $U_m(f) = \int (f^{(m)}(s))^2 ds$
- $m \in \{1,2,\ldots\}$
- RBF smoothness:
- RBF smoothness: $U_m(\phi) = \frac{\prod_{j=1}^{m}(d+2(j-1))}{2\sigma^{2m}}$
- Fourier basis function smoothness:
- Fourier basis function smoothness: $U_m(\phi) = \frac{||c_i||^{2m}\pi^{2m}}{2}$

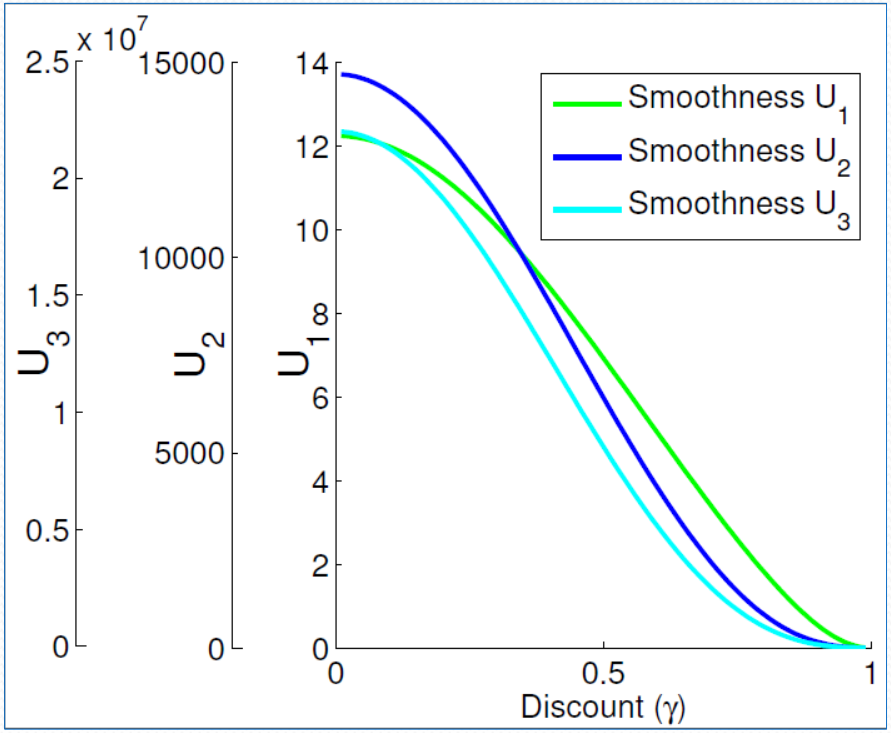
Smooth Tikhonov OMP-TD

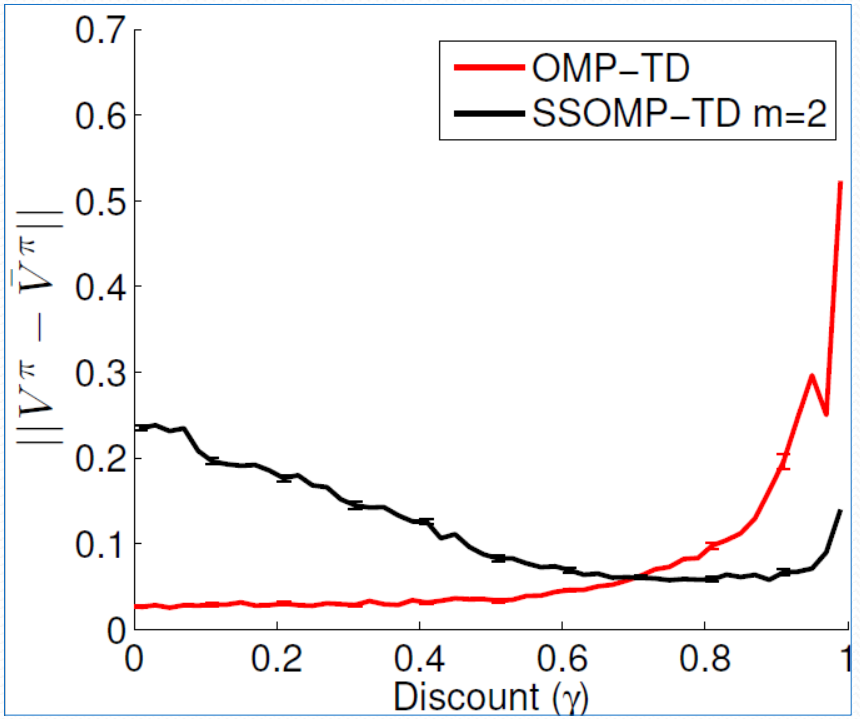
- Smoothness integrated into the selection process using Tikhonov regularization.
- $\beta_i' = \left| \frac{<R, \phi_i>}{\sqrt{\|\phi_i\|^2 + \lambda U_m(\phi_i)}} \right|$
- λ is the regularization parameter.

Smoothness Scaled OMP-TD

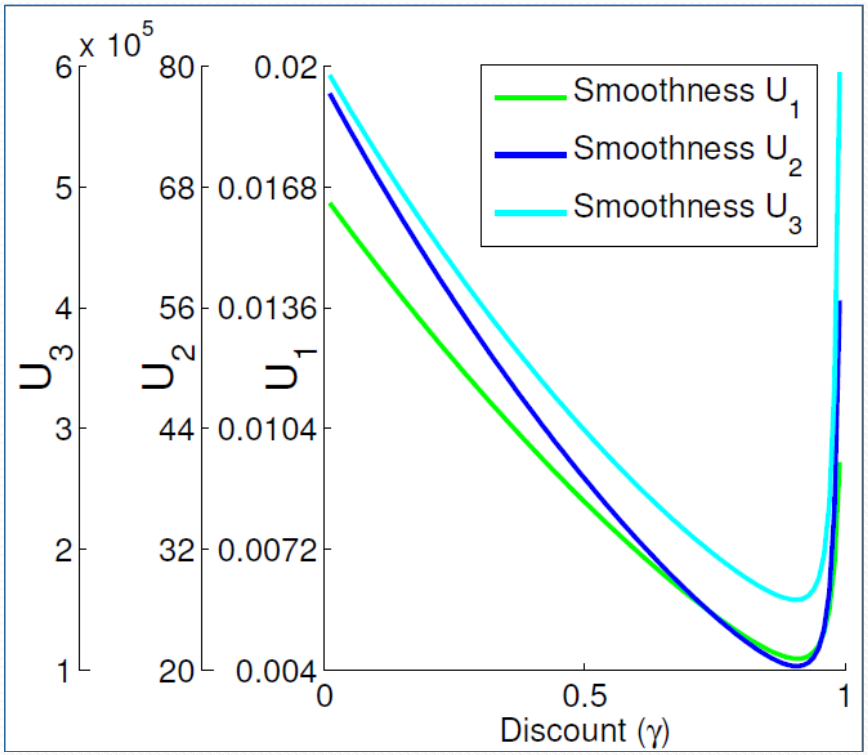
- Heuristic which modifies correlation calculation in OMP-TD to include a preference for smooth basis functions.
- $\rho_{i'} = \left| \frac{<R,\Phi_i>}{\sqrt{\|\Phi_i\|^2 U_m(\Phi_i)}} \right| = \left| \frac{\rho_i}{\sqrt{U_m(\Phi_i)}} \right|$
- Only parameter is m. Usually set to 2.

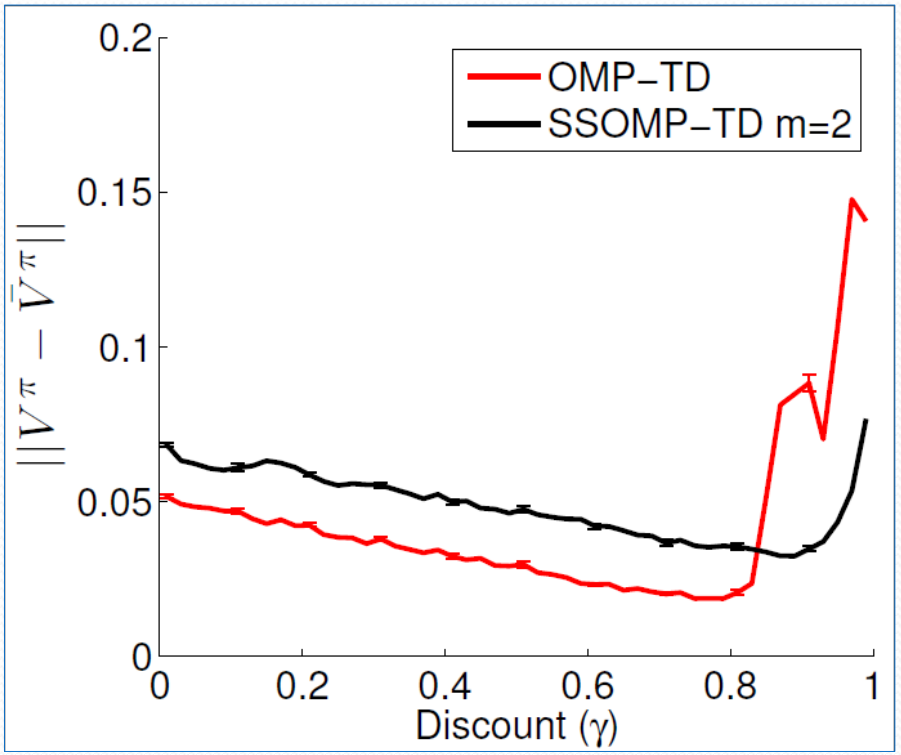
Chainwalk Smoothness



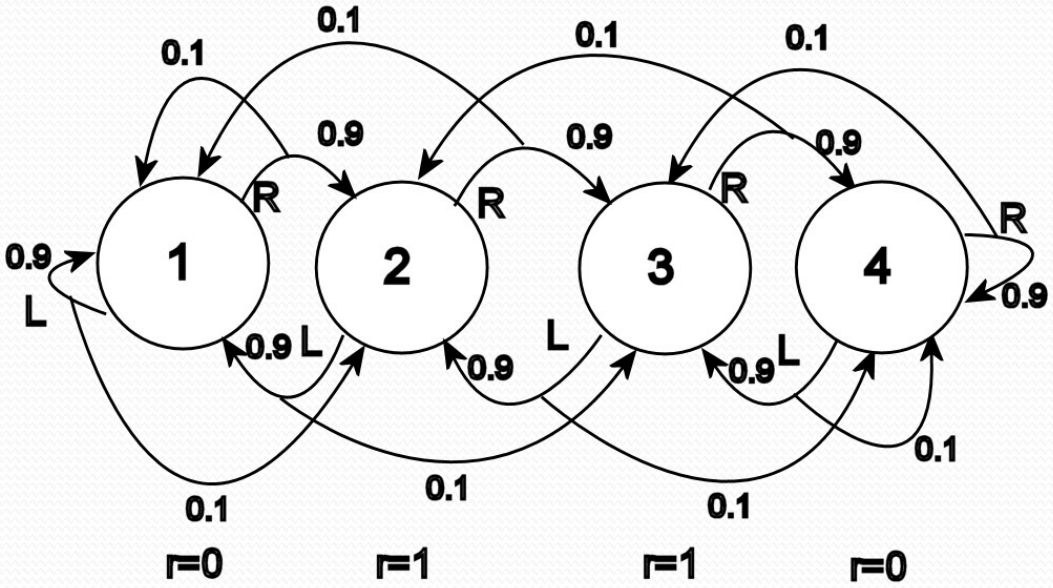


Mountain Car Smoothness

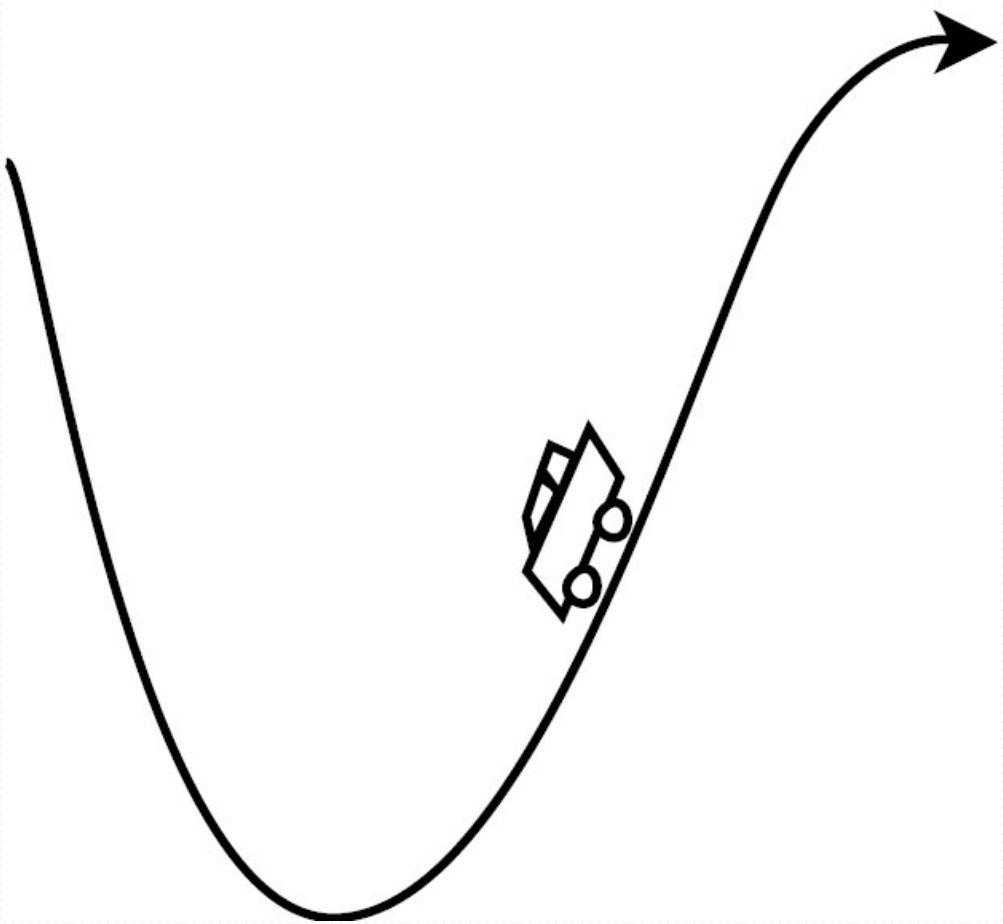




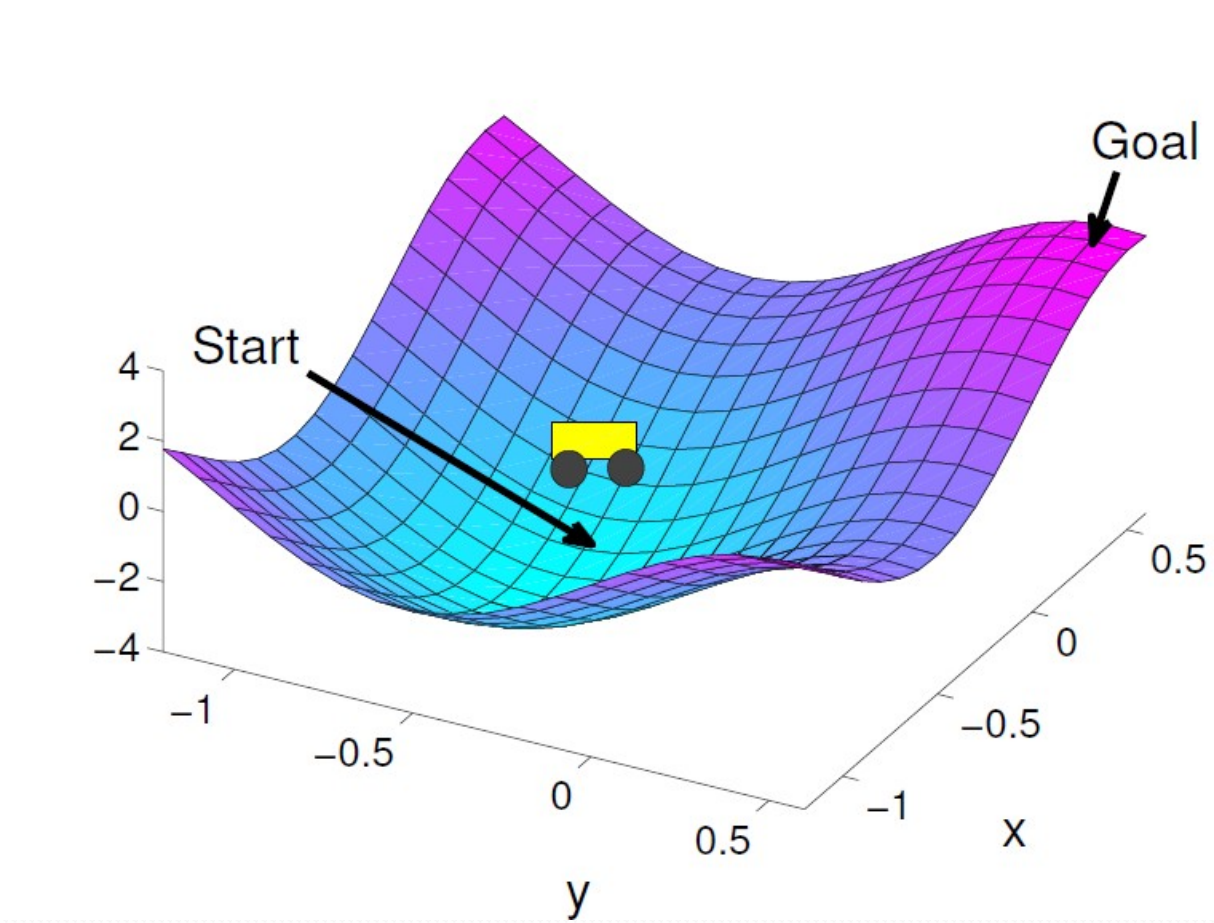
50 State Chainwalk



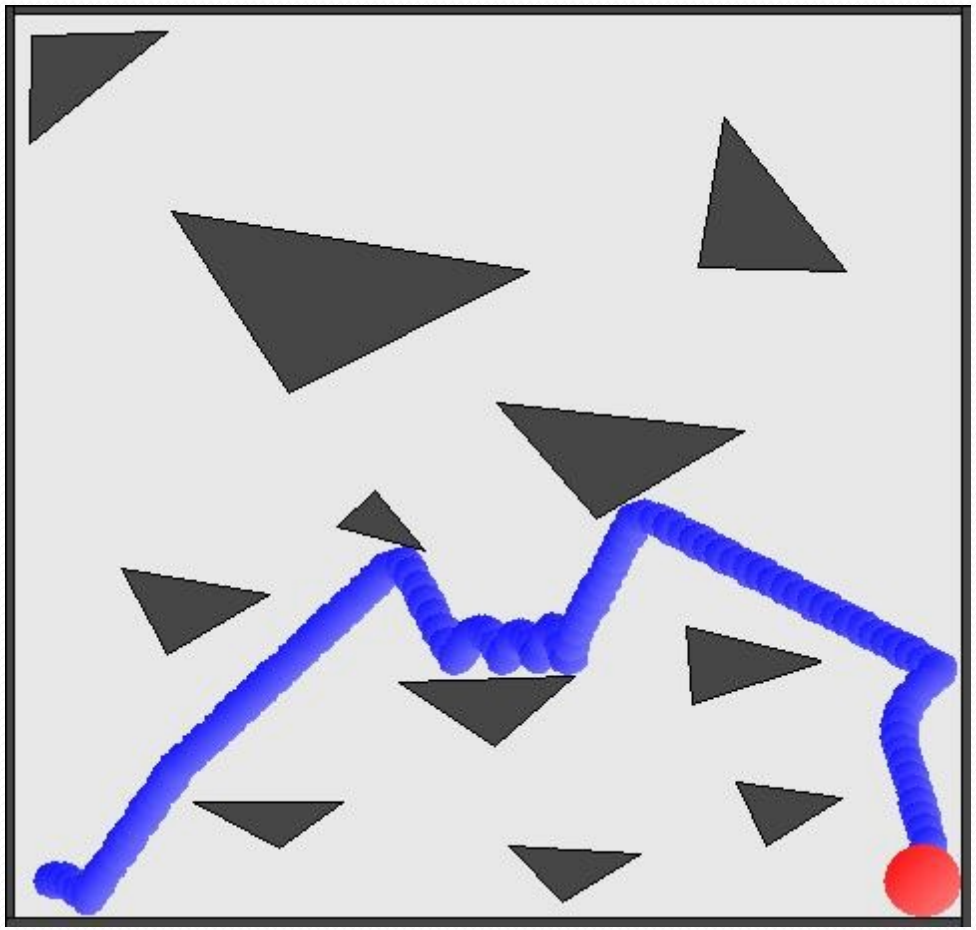
Mountain Car



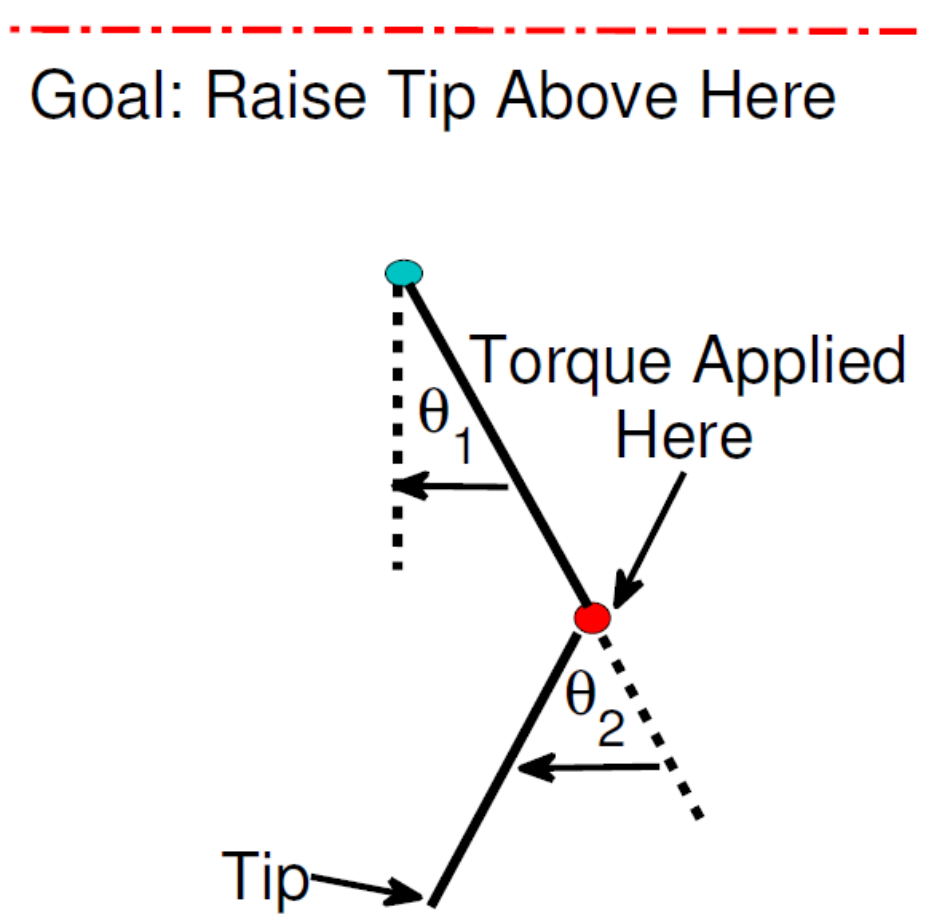
Mountain Car 3D



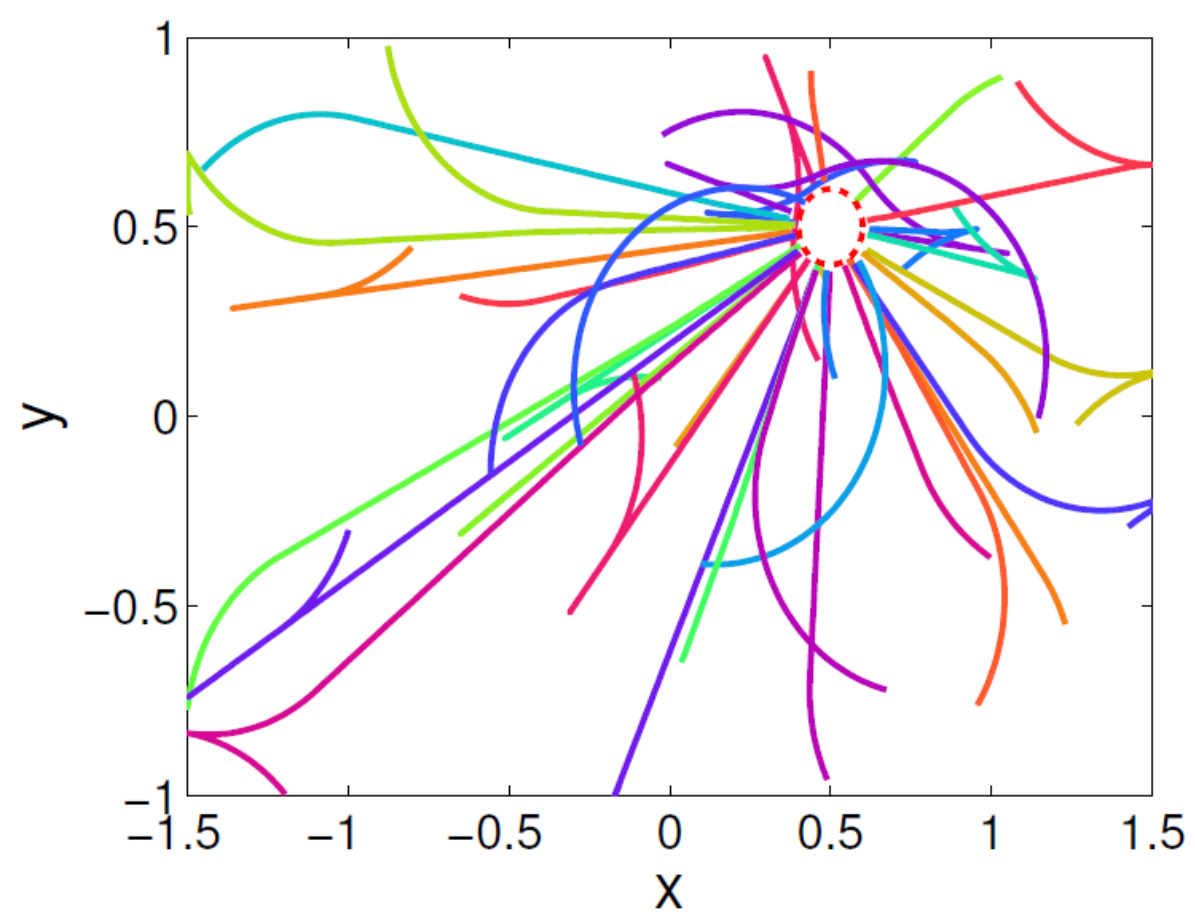
Pinball



Acrobot



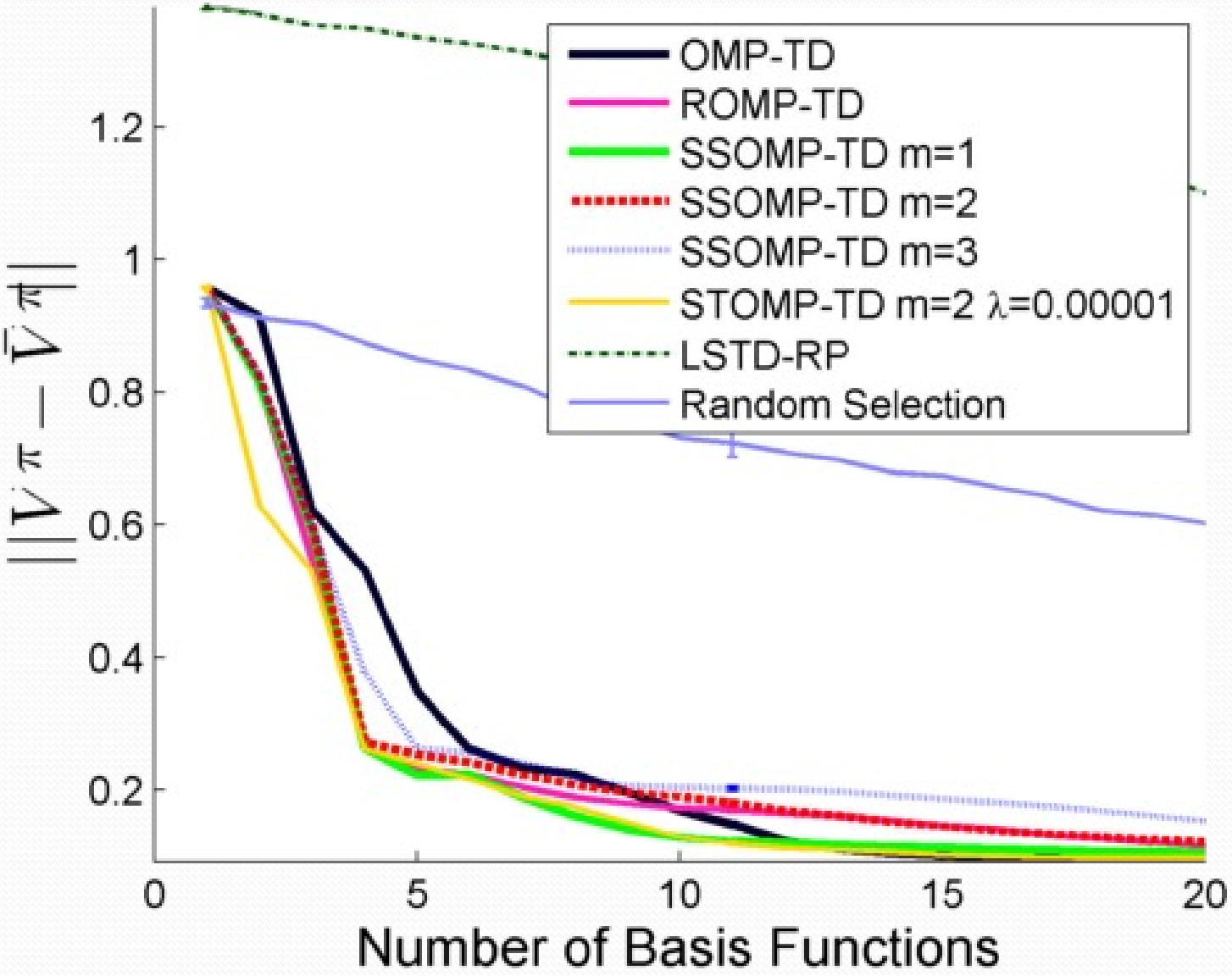
RC Car



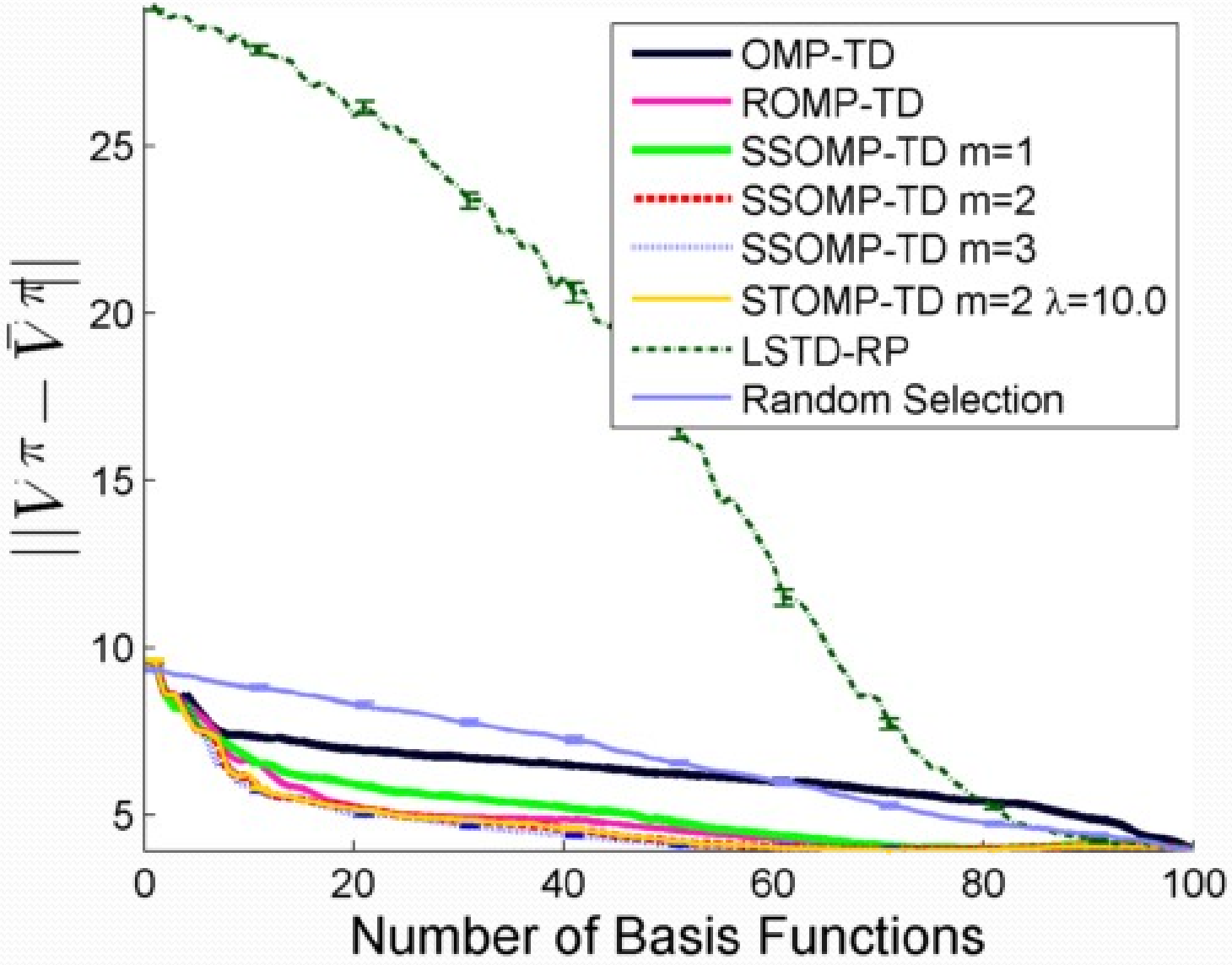
Experiments

- Started with a fixed basis set. Only constant for RBFs and the independent functions for the Fourier experiments.
- Collected samples with a fixed policy.
- Selected a fixed number of basis functions, after each selection recorded the error.
- Goal was to compare each method on the quality of basis function selected.

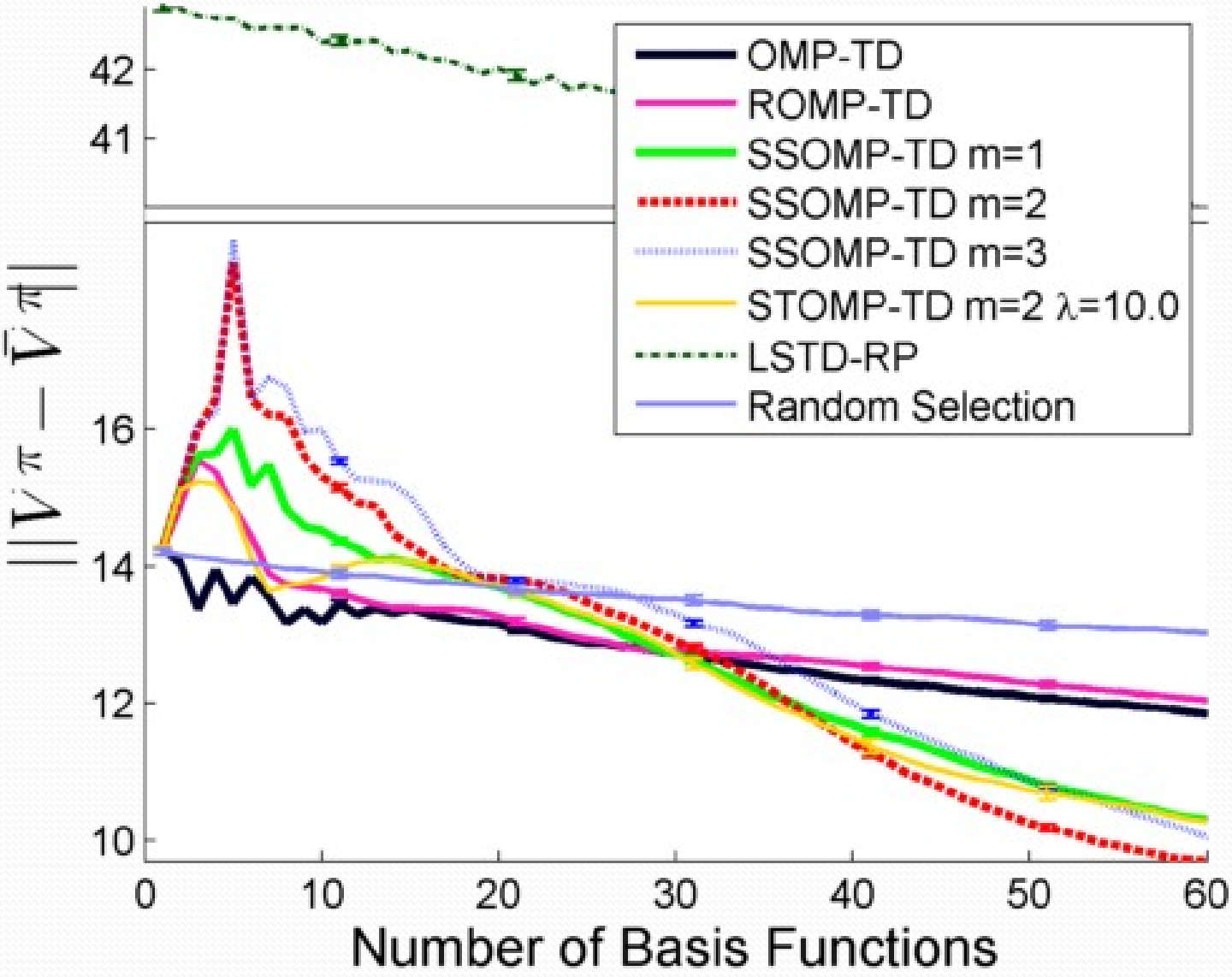
Fourier Chainwalk Results



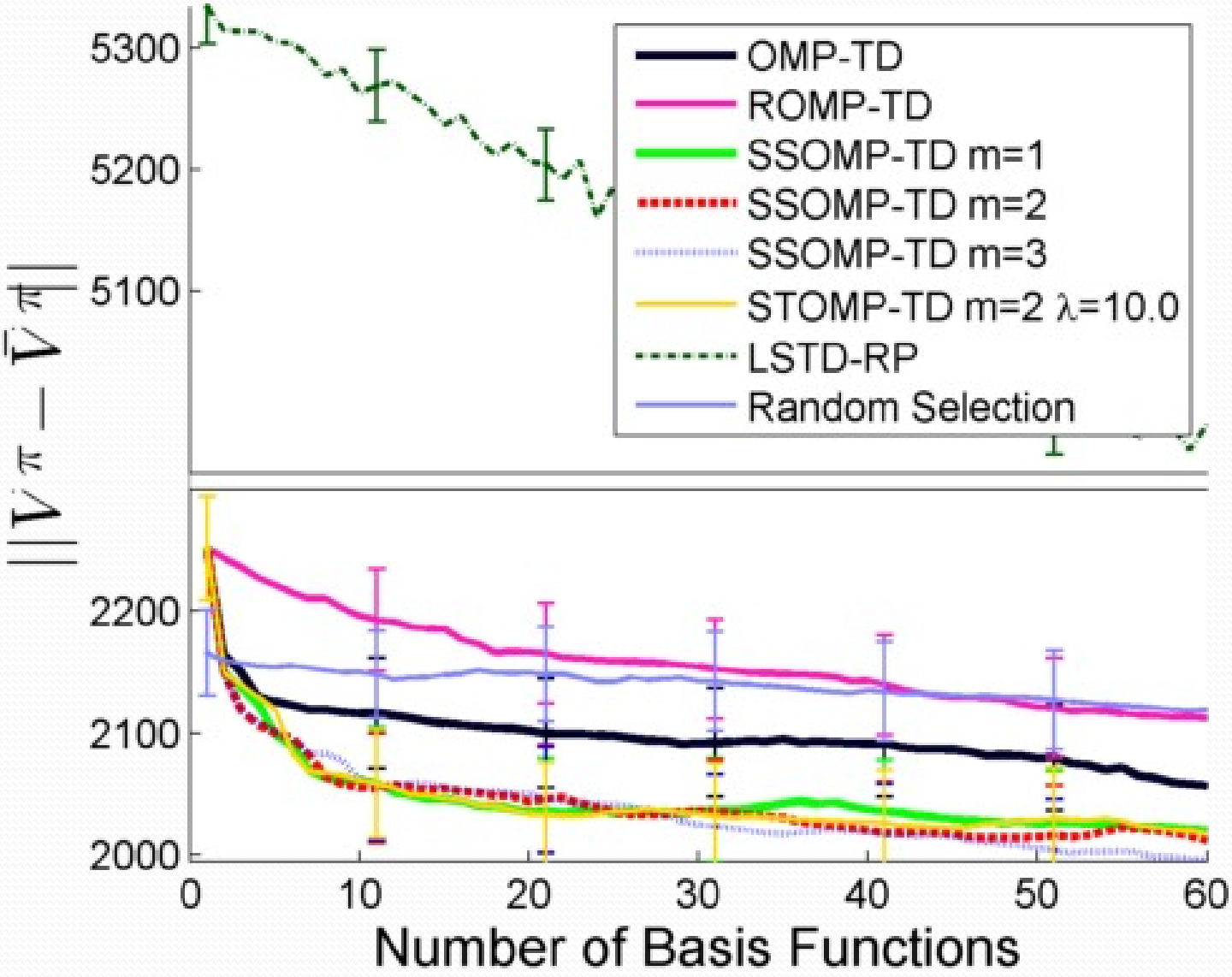
Fourier Mountain Car Results



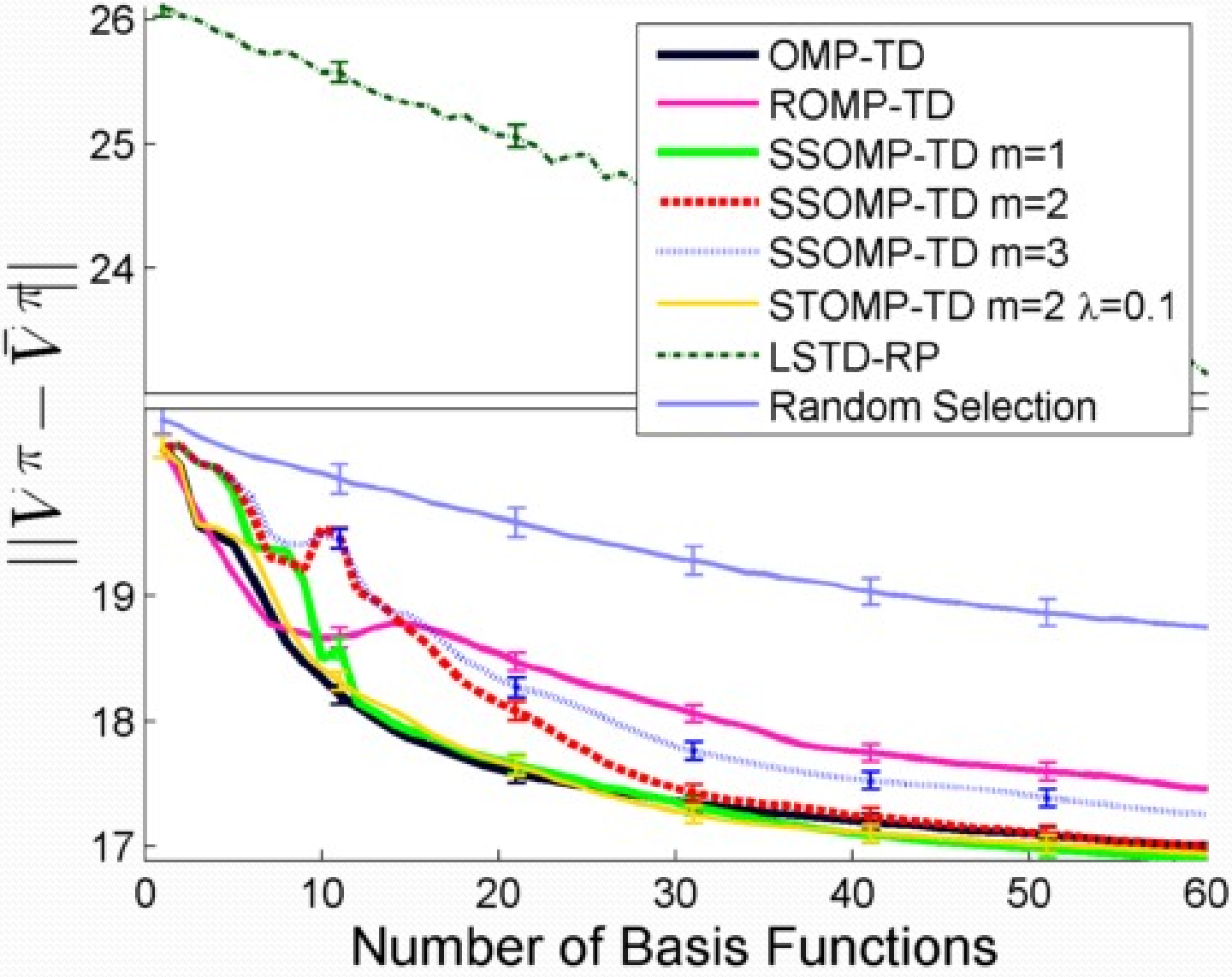
Fourier Mountain Car 3D Results



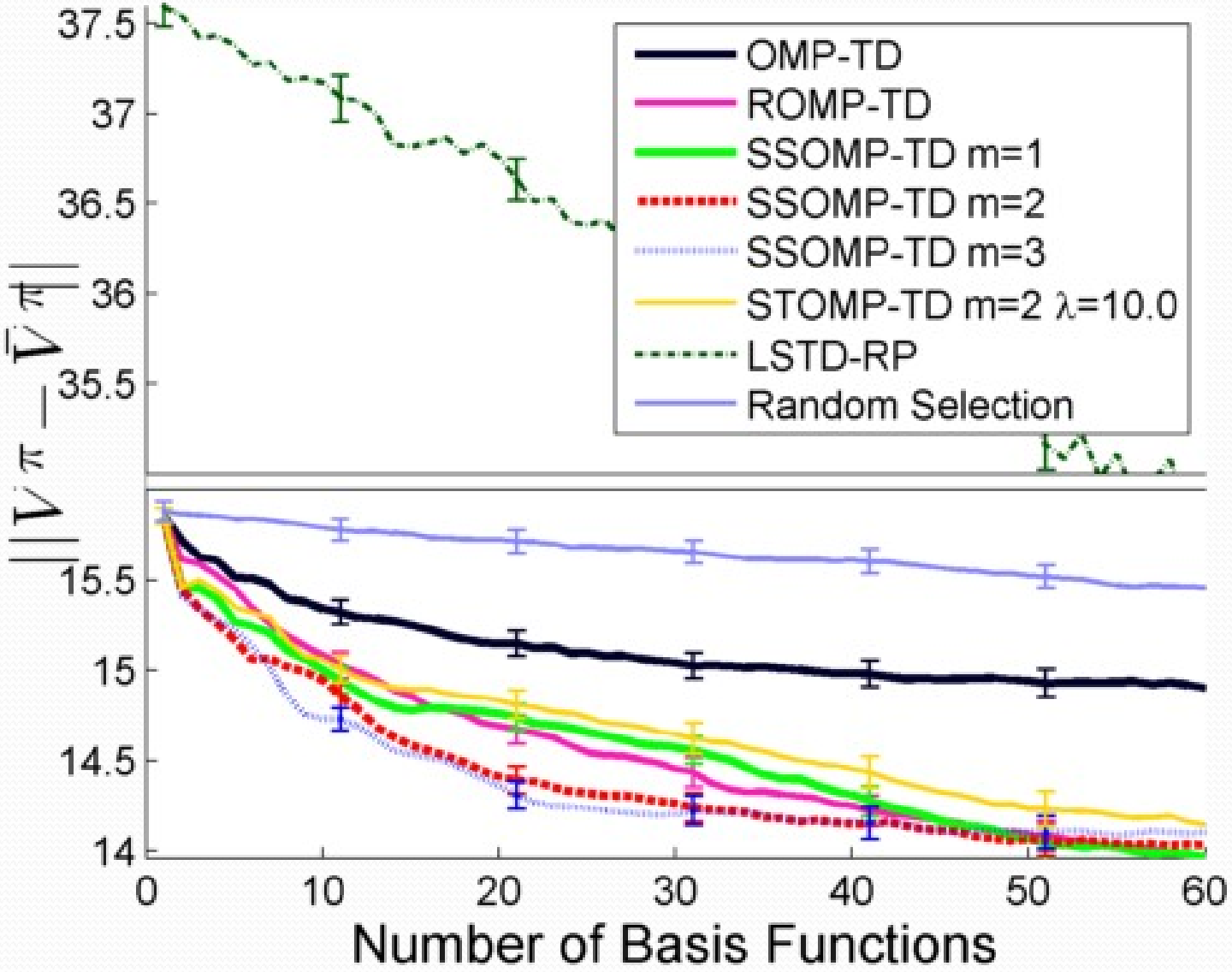
Fourier Pinball Results



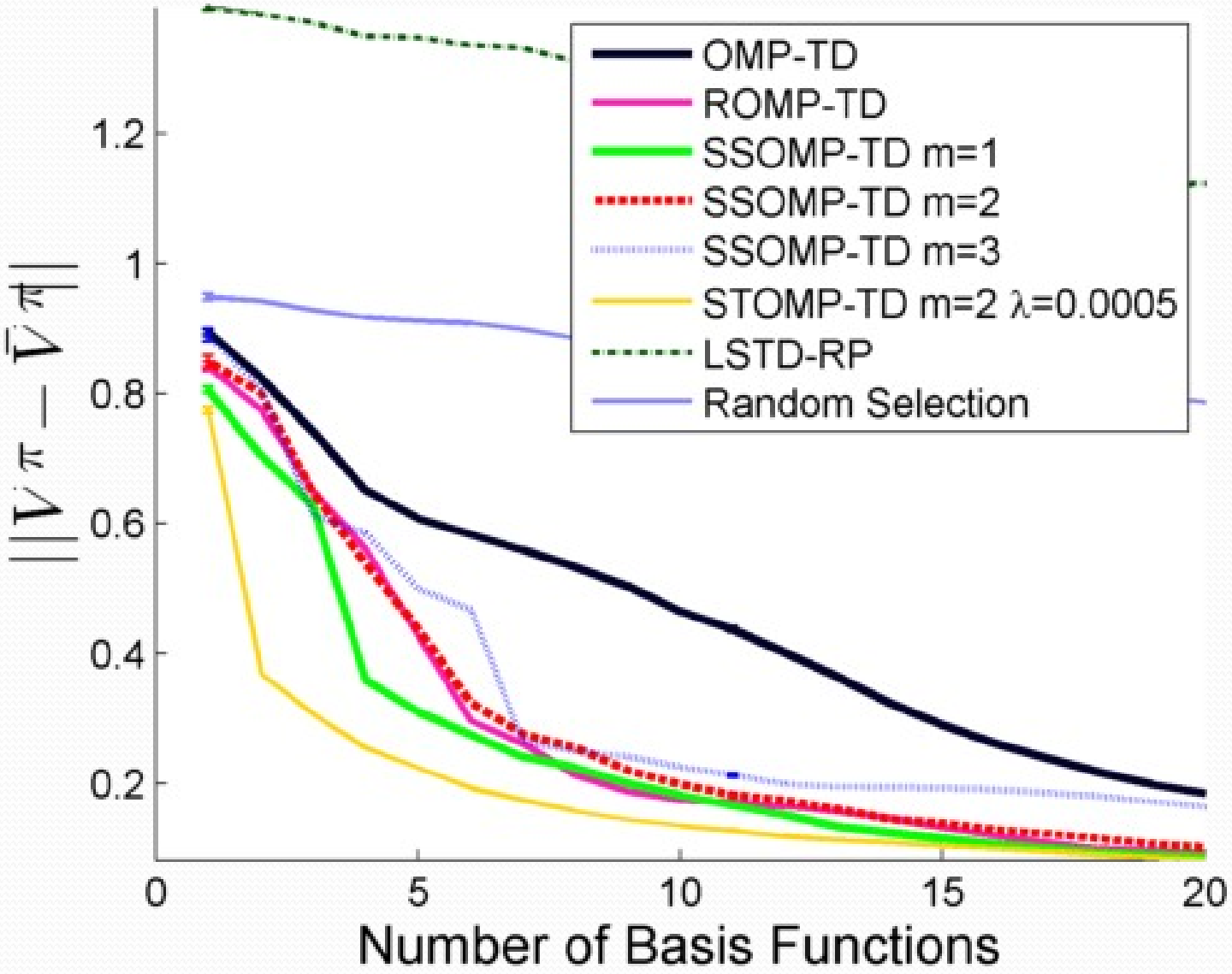
Fourier Acrobot Results



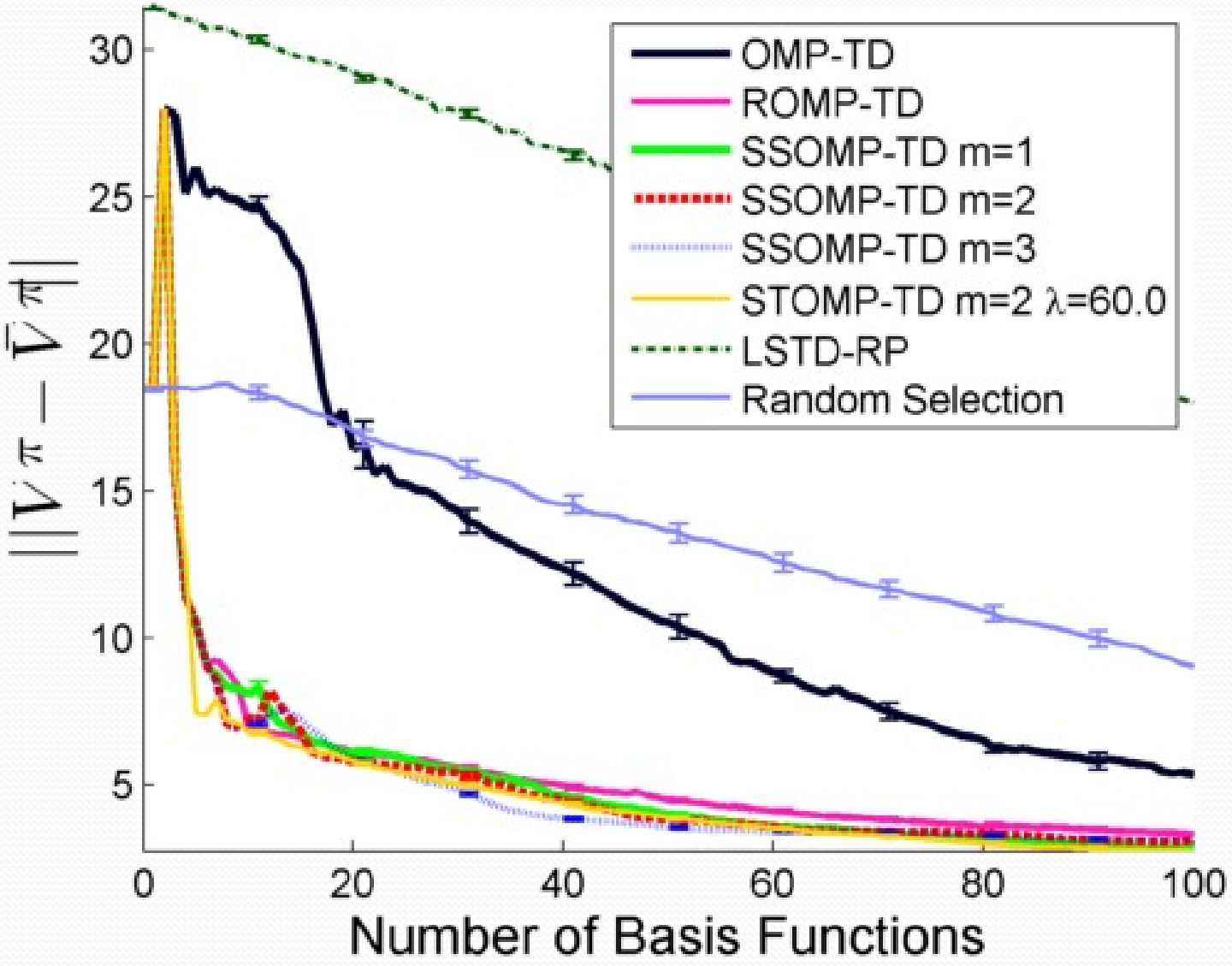
Fourier RC Car Results



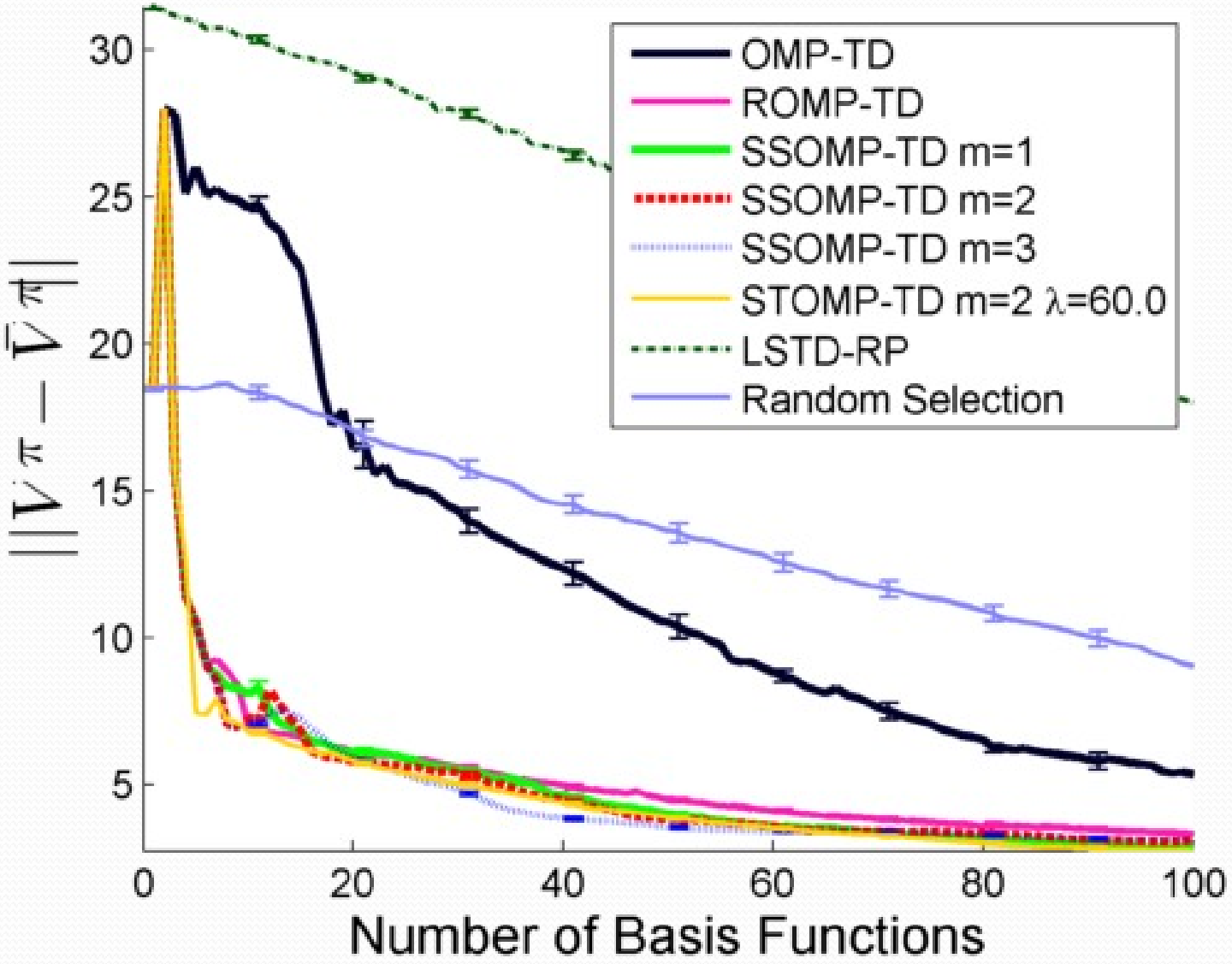
RBF Chainwalk Results



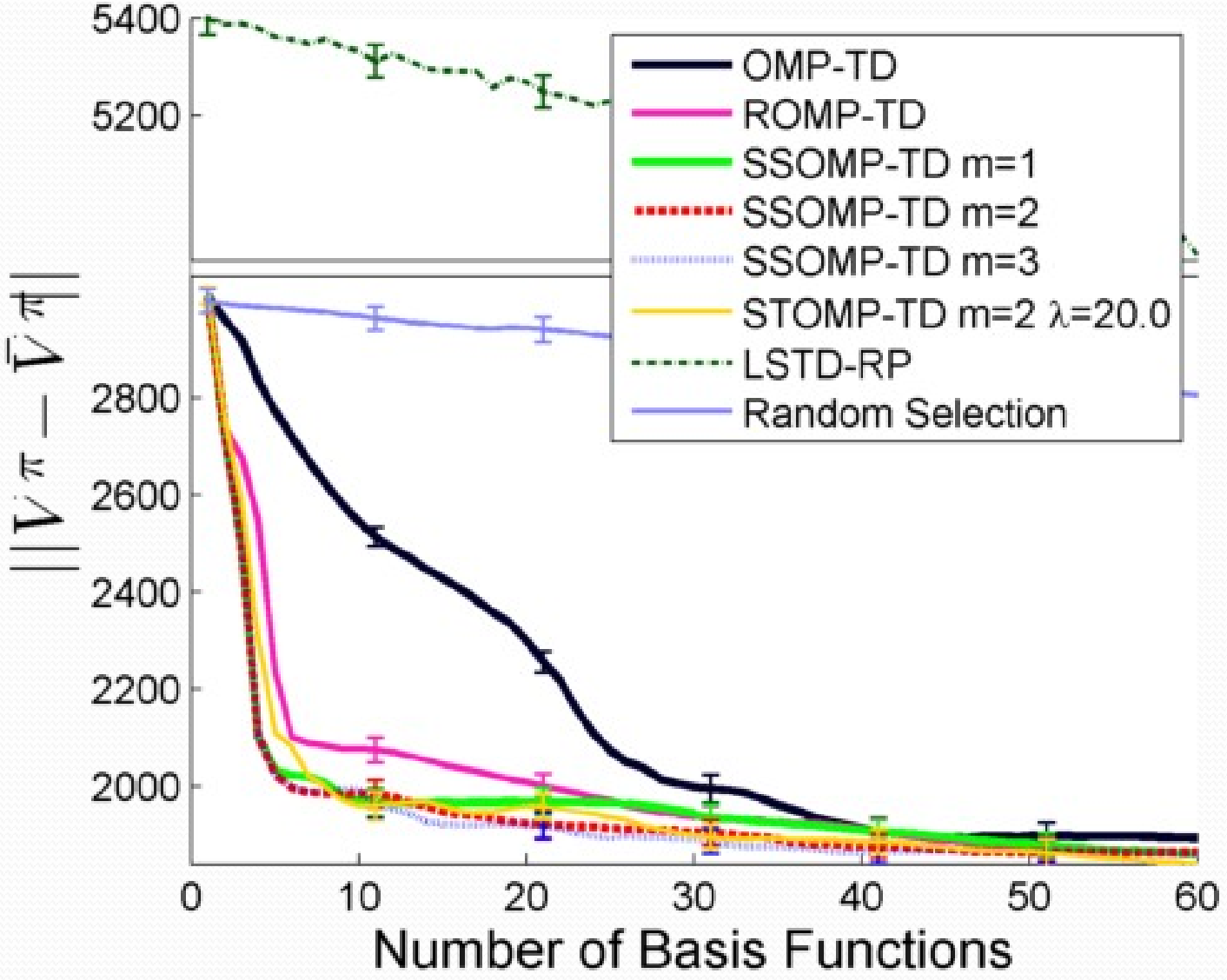
RBF Mountain Car Results



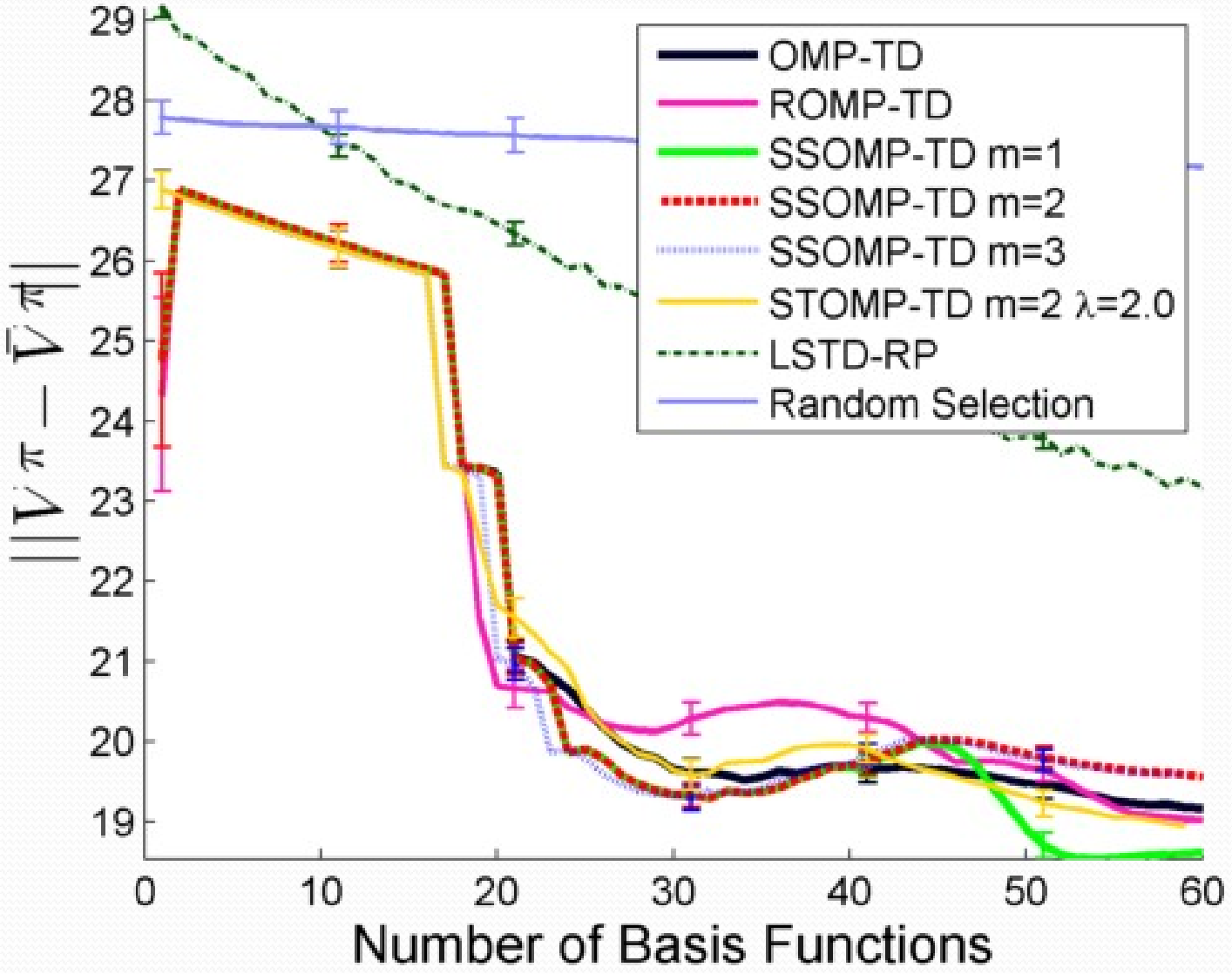
RBF Mountain Car 3D Results



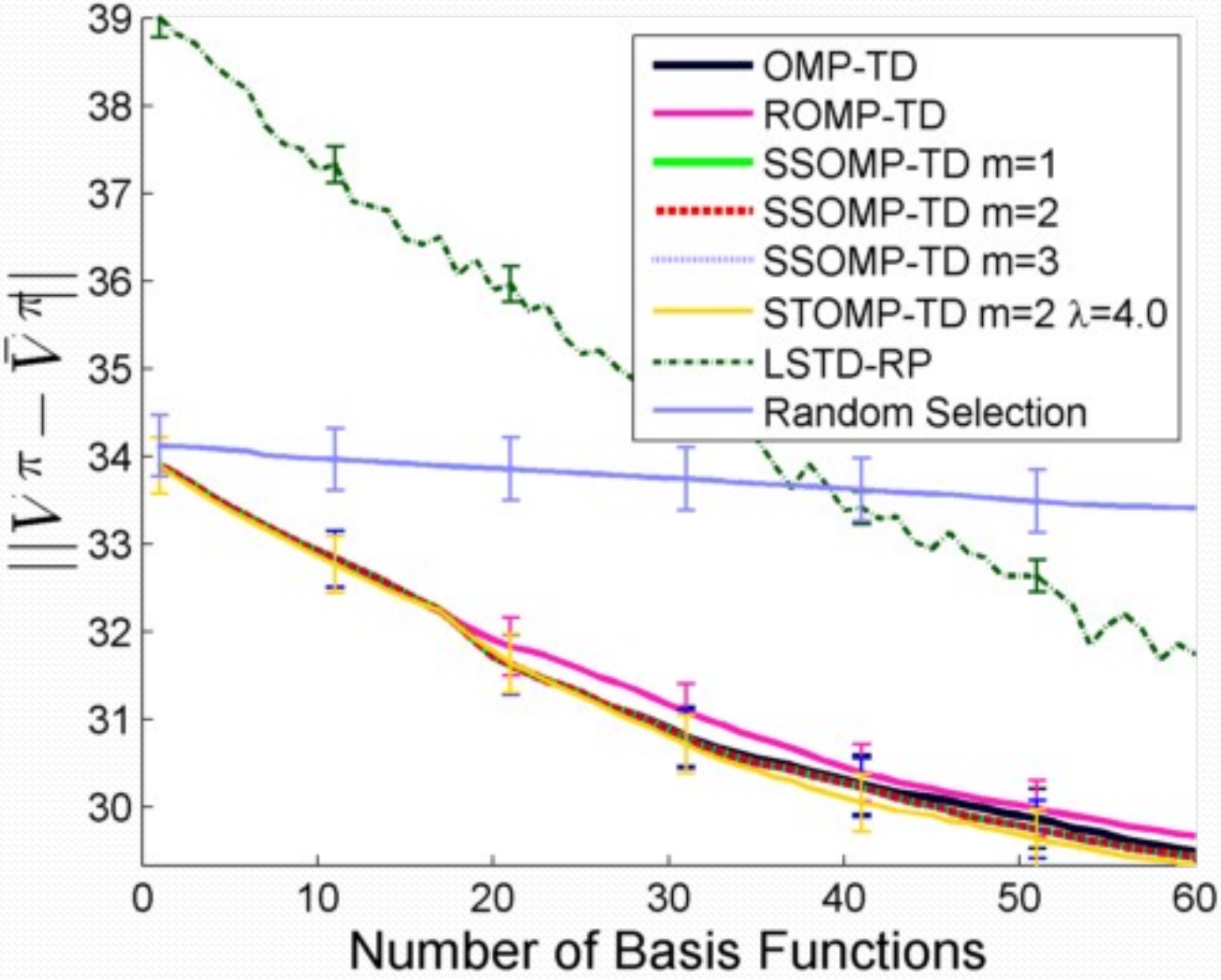
RBF Pinball Results



RBF Acrobot Results



RBF RC Car Results



Future Work

- Wavelets.
- Sublinear Fourier basis function selection.
- Improving the feature selection metric.
  - Selection using LSTD.
  - Treat each state in the sample set as a basis function

Conclusion

- Feature selection is important, especially in RL.
- Using the correlation to the Bellman error as a feature selection metric works well, but there is room for improvement.
- Value functions tend to be smooth.
- STOMP-TD and SSOMP-TD compensate for the weaknesses in OMP-TD when γ is large.
- Possible future avenues include remapping the state space and intelligent selection of dictionary functions.