Nonparametric Representation of Policies and Value Functions: A Trajectory-Based Approach

by Christopher G. Atkeson, Jun Morimoto
held by Rudolf Lioutikov

TECHNISCHE UNIVERSITÄT DARMSTADT

Nonparametric

TECHNISCHE UNIVERSITÄT DARMSTADT

- no assumptions about underlying data distribution
- usually needs more samples than parametric approaches
- computationally expensive

Nonparametric

TECHNISCHE UNIVERSITÄT DARMSTADT

- Examples are
  - K-Nearest Neighbors
  - Kernel Ridge Regression
  - Locally Weighted Regression

Goals

TECHNISCHE UNIVERSITÄT DARMSTADT

- Policies and Value-Functions are represented nonparametrically along trajectories
- minimizing costs (representational resources, computation time, amount of training data)
- reducing amount of training data needed by learning models
- More powerful updates of first and second derivatives of value functions and first derivatives od policies
- reducing representational resources by representing value functions and policies along carefully chosen trajectories

Key Ideas

TECHNISCHE UNIVERSITÄT DARMSTADT

- coordinate many trajectories
- more global function created by combining value functions for the trajectories
- as long as value functions are consistent between trajectories, and cover the appropriate space, the global value function created, will be correct
- supports accurate updating since any update must occur along densely represented optimized trajectories and an adaptive resolution representation that allocates resources to where optimal trajectories tend to go.

Key Ideas

TECHNISCHE UNIVERSITÄT DARMSTADT

- segment trajectories at discontinuities of system dynamics
- reducing amount of discontinuity in the value function of each segment
- assume smooth dynamics and criteria, so that first and second derivatives exist
- using LWR to represent value functions at discontinuities

Bellman's Principle of Optimality

TECHNISCHE UNIVERSITÄT DARMSTADT

- Richard Bellman, Dynamic Programming, 1957
- "An optimal sequence of controls in a multistage optimization problem has the property that whatever the initial stage, state and controls are, the remaining controls must constitute an optimal sequence of decisions for the remaining problem with stage and state resulting from previous controls considered as initial conditions"

Bellman's Principle of Optimality

TECHNISCHE UNIVERSITÄT DARMSTADT

$$\mathbf{u} = \arg\min_{\mathbf{u}}(L(\mathbf{x}, \mathbf{u}) + \lambda V(f(\mathbf{x}, \mathbf{u}))) \tag{1}$$

- $\mathbf{x}$ is a state
- $\mathbf{u}$ are actions
- f represents the dynamics of the system i.e. $\mathbf{x}_{i+1} = f(\mathbf{x}_i, \mathbf{u}_i)$
- L is a one step cost function
- V is the value function
- λ is a discount factor

Taylor approximation

TECHNISCHE UNIVERSITÄT DARMSTADT

- given the point $(\mathbf{x}_p, \mathbf{u}_p)$ we define $\tilde{\mathbf{x}} = \mathbf{x} - \mathbf{x}_p$ and $\tilde{\mathbf{u}} = \mathbf{u} - \mathbf{u}_p$

$$V(\mathbf{x}) \approx V_0 + V_{\mathbf{x}}\tilde{\mathbf{x}} + \frac{1}{2}\tilde{\mathbf{x}}^T V_{xx}\tilde{\mathbf{x}}$$

$$f(\mathbf{x}, \mathbf{u}) \approx f_0 + f_{\mathbf{x}}\tilde{\mathbf{x}} + f_{\mathbf{u}}\tilde{\mathbf{u}} + \frac{1}{2}\tilde{\mathbf{x}}^T f_{xx}\tilde{\mathbf{x}} + \tilde{\mathbf{x}}^T f_{xu}\tilde{\mathbf{u}} + \frac{1}{2}\tilde{\mathbf{u}}^T f_{uu}\tilde{\mathbf{u}}$$

$$L(\mathbf{x}, \mathbf{u}) \approx L_0 + L_{\mathbf{x}}\tilde{\mathbf{x}} + L_{\mathbf{u}}\tilde{\mathbf{u}} + \frac{1}{2}\tilde{\mathbf{x}}^T L_{xx}\tilde{\mathbf{x}} + \tilde{\mathbf{x}}^T L_{xu}\tilde{\mathbf{u}} + \frac{1}{2}\tilde{\mathbf{u}}^T L_{uu}\tilde{\mathbf{u}}$$

Backwards sweep

TECHNISCHE UNIVERSITÄT DARMSTADT

- computing the derivatives of the Q-function

$$
\begin{aligned}
Q_{\mathbf{x}} &= L_{\mathbf{x}} + \lambda V_{\mathbf{x}} f_{\mathbf{x}} \\
Q_{\mathbf{u}} &= L_{\mathbf{u}} + \lambda V_{\mathbf{x}} f_{\mathbf{u}} \\
Q_{\mathbf{xx}} &= \lambda f_{\mathbf{x}}^T V_{\mathbf{xx}} f_{\mathbf{x}} + \lambda V_{\mathbf{x}} f_{\mathbf{xx}} + L_{\mathbf{xx}} \\
Q_{\mathbf{ux}} &= \lambda f_{\mathbf{u}}^T V_{\mathbf{xx}} f_{\mathbf{x}} + \lambda V_{\mathbf{x}} f_{\mathbf{ux}} + L_{\mathbf{ux}} \\
Q_{\mathbf{uu}} &= \lambda f_{\mathbf{u}}^T V_{\mathbf{xx}} f_{\mathbf{u}} + \lambda V_{\mathbf{x}} f_{\mathbf{uu}} + L_{\mathbf{uu}}
\end{aligned}
$$

Backwards sweep

TECHNISCHE UNIVERSITÄT DARMSTADT

- defining additional terms

$$\Delta \mathbf{u} = Q_{\mathbf{uu}}^{-1} Q_{\mathbf{u}}$$
$$\mathbf{K} = Q_{\mathbf{uu}}^{-1} Q_{\mathbf{ux}}$$
$$V_{\mathbf{x}_{i-1}} = Q_{\mathbf{x}} - Q_{\mathbf{u}} \mathbf{K}$$
$$V_{\mathbf{xx}_{i-1}} = Q_{\mathbf{xx}} - Q_{\mathbf{ux}} \mathbf{K}$$

Forward sweep

TECHNISCHE UNIVERSITÄT DARMSTADT

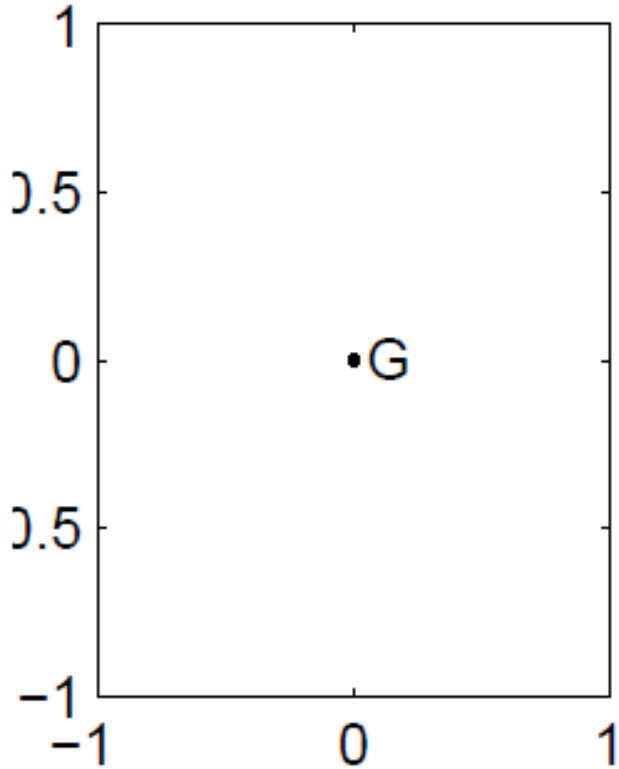
- to update the trajectory itself, forward integration can be used

$$\mathbf{u}_{new} = \mathbf{u} - \Delta\mathbf{u} - \mathbf{K}(\mathbf{x}_{new} - \mathbf{x}) \tag{2}$$

Regulator task

TECHNISCHE UNIVERSITÄT DARMSTADT

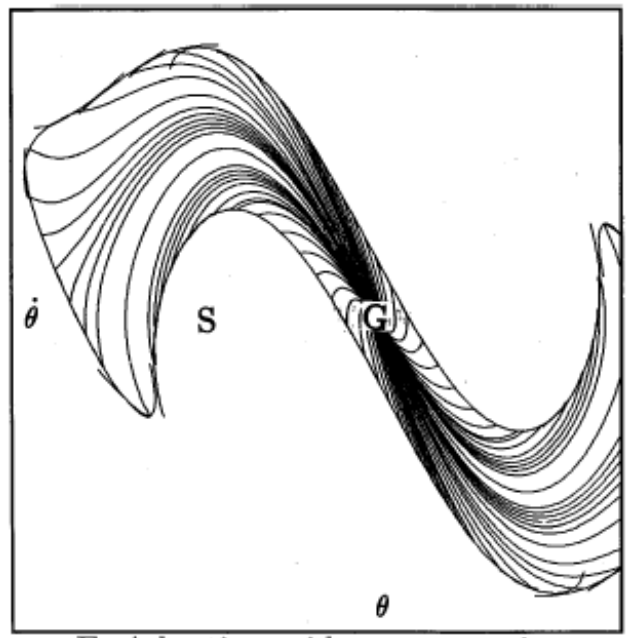
- aka steady state control
- trajectory starts and ends at the goal
- value function is in the vicinity of the goal
- constanc policy $\tilde{\mathbf{u}} = \mathbf{K}\tilde{\mathbf{x}}$



Task with point goal

TECHNISCHE UNIVERSITÄT DARMSTADT

- value function for a swing up problem
- regulation about the unstable equilibrium
- nonlinearities limit the region of applicability
- non-trivial trajectories needed for a larger region
- here: thresholded region is filled with trajectories
- consistent value functions among neighbors
- optimal value function / policy are in this region

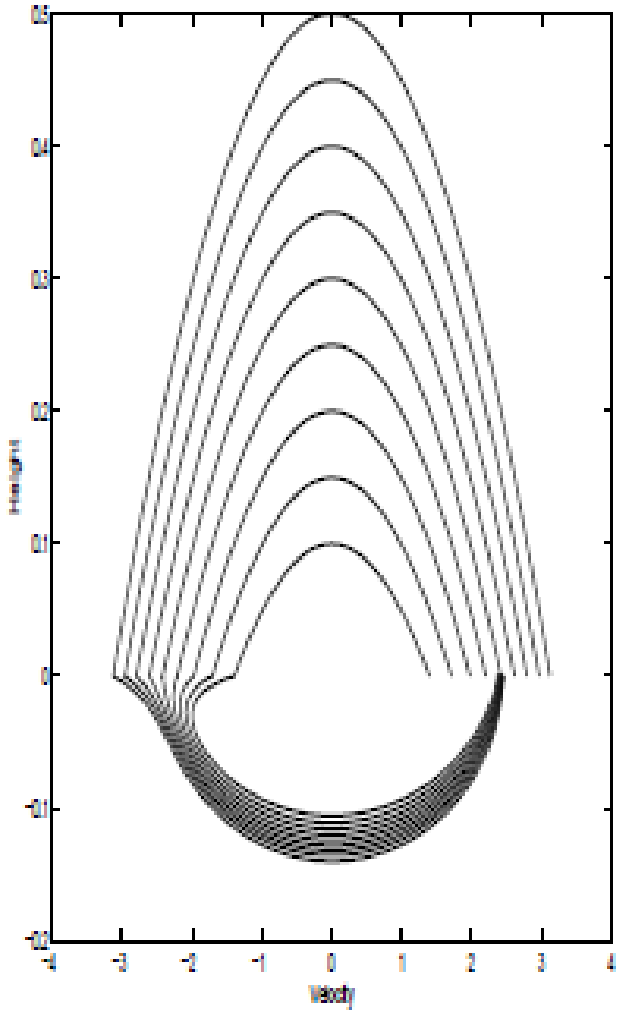


E: Adaptive grid representation

Periodic Task

TECHNISCHE UNIVERSITÄT DARMSTADT

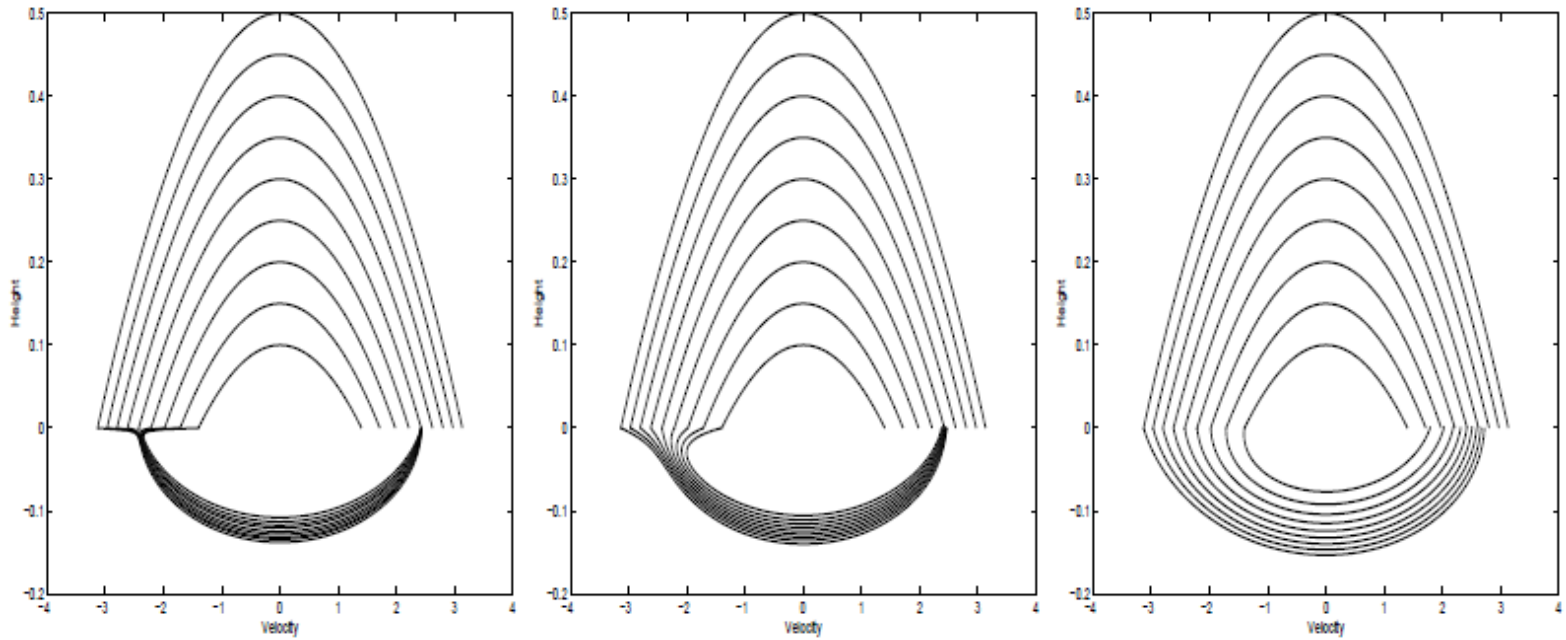
- control of vertical hopping
- goal state replaced by desired hopping height
- counterclockwise movement direction



Periodic Task

TECHNISCHE UNIVERSITÄT DARMSTADT

- discontinuity at horizontal axis
- funneling effect
- funnel width controlled by penalty on $\mathbf{u}$ usage
- optimal hopper with different penalties



How to get initial trajectories ?

TECHNISCHE UNIVERSITÄT DARMSTADT

- for regulator tasks its trivial
- for point goal tasks trajectories can be extended backwards away from the goal
- for periodic tasks, crude trajectories must already exist
- created by other approaches

What approaches are used ?

TECHNISCHE UNIVERSITÄT DARMSTADT

- manually designed controllers
- imitation learning
- parameterized trajectory from policy search

Trajectory refinement

TECHNISCHE UNIVERSITÄT DARMSTADT

- global policy by policy optimization
- local policy from nearest point with same type of dynamics
- local value function estimate from the nearest point with the same type of dynamics
- use the policy from the nearest trajectory

Trajectory refinement

TECHNISCHE UNIVERSITÄT DARMSTADT

- all four methods are used parallel
- locally optimize each produced trajectory
- only the best trajectory is stored

Walking Robot

TECHNISCHE UNIVERSITÄT DARMSTADT

- simple planar biped, that walks along a bar
- no knees, but can grab the bar as leg swings by (like a monkey)
- 5 dimensional state space
  - left and right leg angle θ_l, θ_r
  - left and right leg angular velocity $\dot{\theta}_l, \dot{\theta}_r$
  - stance foot location
- controlled action is torque τ at the hip

Cost function

TECHNISCHE UNIVERSITÄT DARMSTADT

- goal described by “shaping” terms
  - keeping the hips at the right altitude
  - with minimal vertical velocity
  - keeping leg amplitude within reason
  - maintaining a symmetric gait
  - maintaining the desired hip forward velocity

Cost function

TECHNISCHE UNIVERSITÄT DARMSTADT

- results in the following cost function

$$cost = w_y(y-1)^2 + w_{\dot{y}}\dot{y}^2 + w_l leg_r^2 + w_l leg_l^2 + w_{lr} leg_{lr} + w_{\dot{x}}(\dot{x} - \dot{x}_d)^2 + \tau^2$$

  - weighting factors $w_y = 100$, $w_{\dot{y}} = 100$, $w_l = 100$, $w_{lr} = 100000$, $w_{\dot{x}=100}$
  - desired leg velociy $\dot{x}_d = 0.4m/s$
  - leg_l, leg_r describe how far leg has gone past its limits ± 0.1 in radians
  - leg_{lr} is the product of the leg angles

Initial trajectory

TECHNISCHE UNIVERSITÄT DARMSTADT

- Initial trajectory generated by optimizing coefficients of a linear policy

$$\tau = \alpha_0 + alpha_1\theta_{lr} + alpha_2\theta_l + \alpha_3 y + \alpha_4\dot{\theta}_{lr} + \alpha_5\dot{\theta}_l + \alpha_6\dot{(x)} + alpha_7\dot{y}$$

- where θ_{lr} is the angle between the legs
- used when left leg was in stance
- negate appropriate signs for right leg

Result cost

TECHNISCHE UNIVERSITÄT DARMSTADT

- cheaper than parametric policy optimization approach
- measured undiscounted cost over 1 second
- starting in a state along the lowest cost trajectory
- optimized parametric policy: 4316
- trajectory-based approach: 3502

Result robustness

TECHNISCHE UNIVERSITÄT DARMSTADT

- more robust than parametric policy optimization approach
- adding offset to the starting point until policy fails
- maximum offset for parametric policy
  - $-0.02 \leq \theta_l \leq 0.06$
  - $-0.45 \leq \dot{\theta}_l \leq 0.1$
  - $-0.2 \leq \theta_{lr} \leq 0.03$
  - $-0.78 \leq \dot{\theta}_{lr} \leq 0.2$
- maximum offset of trajectory-based approach was greater or equal in each case
- not surprising. trajectory based approach uses parametric policy to initialize trajectories.

Result interference

TECHNISCHE UNIVERSITÄT DARMSTADT

- optimized parametric policy from a distribution of starting states.
- original states and states with positive offset
- new cost was 14.747 (before 4316)
- cost for trajectory approach stayed the same (3502)

Robustness to modelling error, probabilistic

TECHNISCHE UNIVERSITÄT DARMSTADT

- state is augmented with unknown parameters (masses, friction, etc.)
- Kalman filter is used as the new dynamics equation
- cost and value function are modified
- reward for moving into regions where value function becomes planar

Robustness to modelling error, game based

TECHNISCHE UNIVERSITÄT DARMSTADT

- minimax optimization
- dynamics equation is augmented with a disturbance term
- “opponent" controls the disturbance
- expect worst possible disturbance
- trajectories should be robust against most-likely disturbance
- cost function contains trade-off between robustness and original reward

Critic reception

TECHNISCHE UNIVERSITÄT DARMSTADT

- the approach is not sufficiently explained
- no performance benchmarks (neither time nor resources)
- what parametric approach is used?