Practical numerical methods for stochastic optimal control of biological systems in continuous time and space

Alex Simpkins

and

Emanuel Todorov

4/1/2009

2009 IEEE International Symposium on Adaptive Dynamic Programming and Reinforcement Learning

Introduction (I)

- Modeling sensorimotor control/learning with stochastic optimal control leads to PDE's which are Nonlinear, Stochastic, 2nd order, high dimensional, no certainty equivalence, potentially non-quadratic costs
- Two solution approaches
  - **Approximate the *problem*, solve approximation exactly**
    - Create an MDP, solve it with dynamic programming
  - **Approximate a *solution* to the original problem with continuous time/space function approximators which minimizes the Bellman error (diff. Between left and right sides of HJB equation)**
    - Minimizing
    - Can also work directly with policy
    - Actor-critic method

Introduction (II)

- Discretization
  - **Approximate problem then solve approximation**
  - **Only works in low dimensional spaces**
  - **Will be used here to validate function approximators**
  - **Guaranteed to converge to the optimal solution of the approximating MDP**
- Continuous function approximation
  - **Approximate a global solution to the original problem**
    - Many ways to do this, we focus on one
  - **Works in higher dimensional spaces**
  - **No guarantees but can work well**

Introduction (III)

- In practice many things can interfere with a successful solution
  - **Number of parameters required**
  - **Overfitting/generalization**
  - **Numerical errors (sparsity, differentiation, condition number)**
  - **How to determine the constant parameters (centers and widths of Gaussians in our case)**
  - **How to know if good fit?**
  - **Numerical stability**
  - **Visualizing the solution**
  - **Repeated measures/one time solution**

The problem formulation (I)

- Dynamics

$$dx = \big(a(x) + B(x)u(x)\big)dt + C(x)d\omega$$

- Immediate cost

$$\ell(x, \pi) = q(x) + \frac{1}{2}||\pi||^2$$

The problem formulation (II)

- Formulate the cost/reward function
  - **In this case discounted cost infinite horizon (no expected final time to the pendulum problem**

$$V^{\pi}(x) = \int_{t}^{\infty} e^{-\alpha(s-t)} \ell[x(s), u(s)] ds$$

- The HJB equation for infinite horizon discounted cost stochastic problems is found by the principle of optimality

$$\alpha V^*(x) = \min_u \left\{ \ell(x,u) + (a(x) + B(x)u(x)) V_x^*(x) + \frac{1}{2} Tr(C(x)C(x)^T V_{xx}^*(x)) \right\},$$

The problem formulation (III)

- Compute the optimal control in closed form in terms of the value function

$$\pi(x) = -B(x)V_x(x)$$

- Substitute above into the HJB equation, drop *min* operator, we arrive at the problem

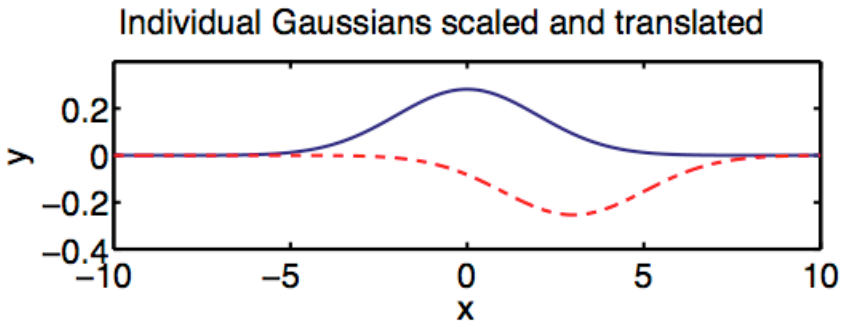
$$\alpha V(x) = q(x) + a(x)^T V_x(x) + \frac{1}{2} Tr(C(x)C(x)^T V_{xx}(x)) - \frac{1}{2}||\pi||^2$$

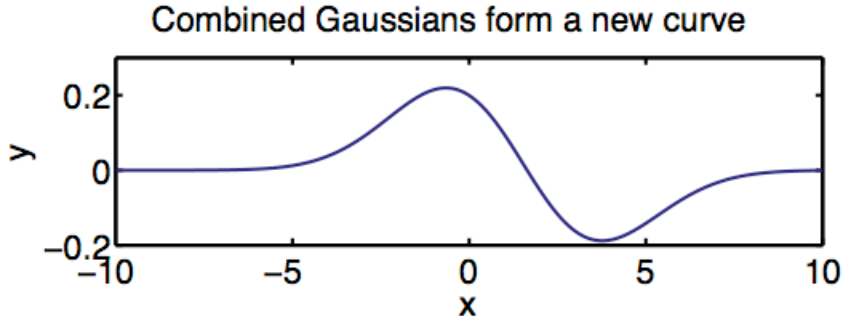
The function approximator

- Consider a global continuous function approximator which is linear in the parameters (w):

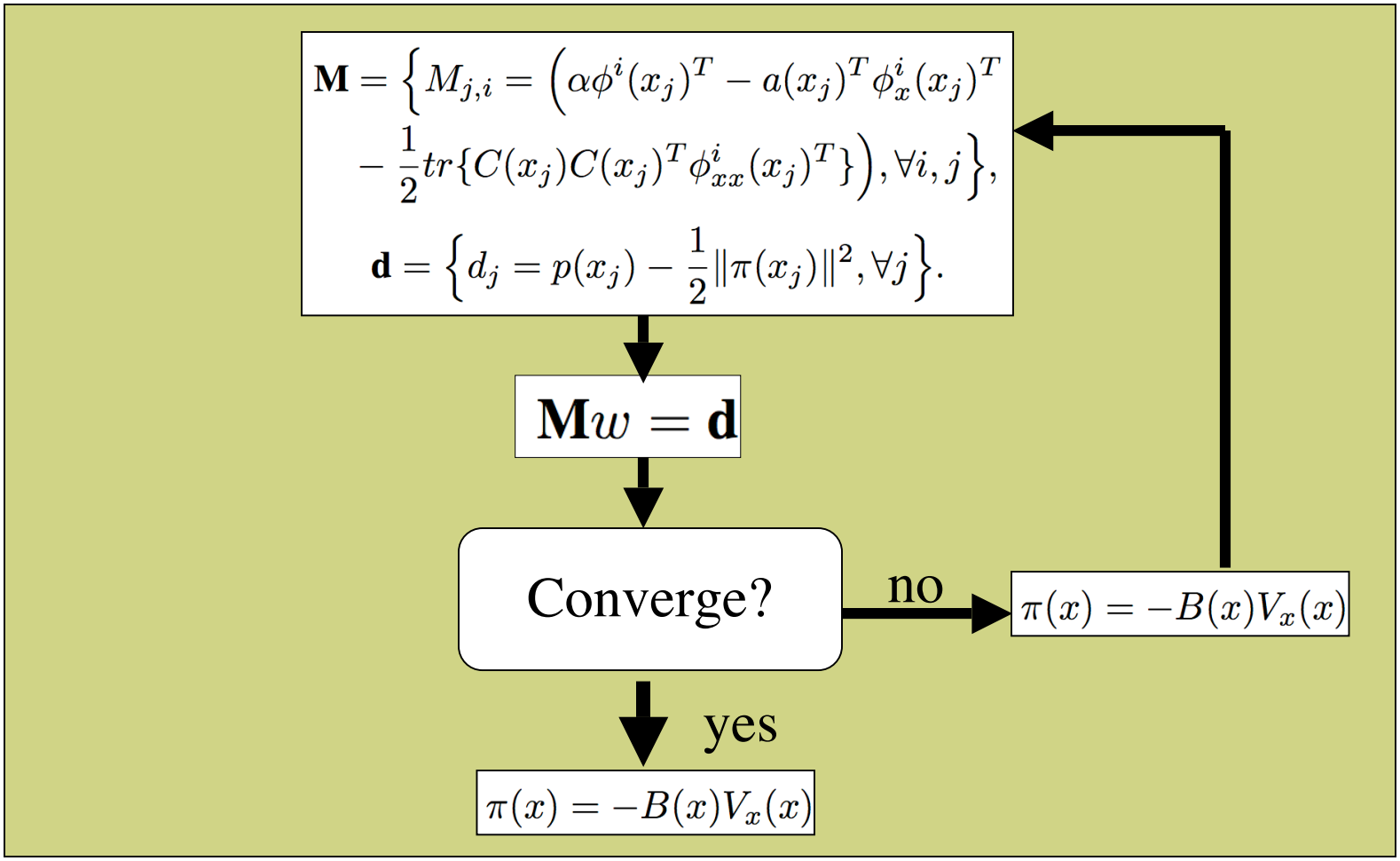
$$V(x;w) = \sum_i w_i \phi^i(x) \qquad V_x(x;w) = \sum_i w_i \phi_x^i(x) \qquad V_{xx}(x;w) = \sum_i w_i \phi_{xx}^i(x)$$

- Where ϕ is a set of predefined features and w is a vector of parameters to compute
- Typical features in this case are chosen to be a large number of Gaussians whose centers are uniformly sampled



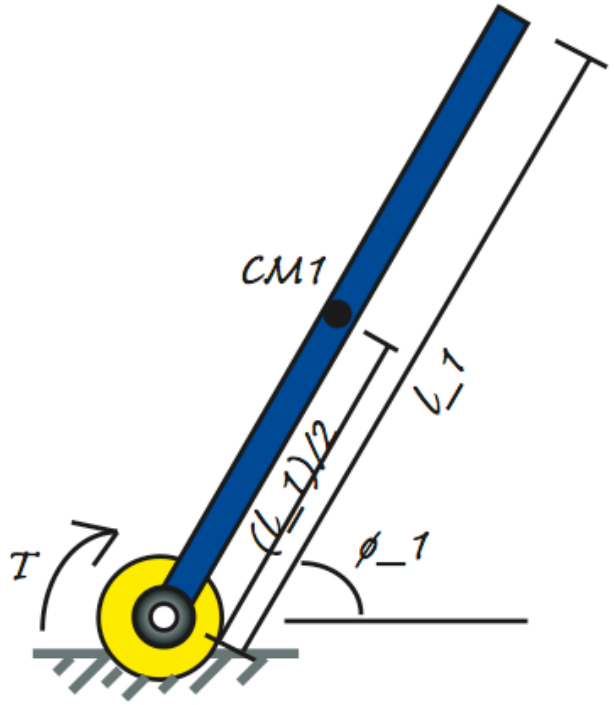


The iLS scheme - collocation



The well-known example problem - 1-DOF pendulum

- Inverted pendulum torque-limited swing-up task
- Dynamics and cost rate



$$J\ddot{\theta} + H\dot{\theta} + G(\theta) = \tau$$

$$J = ml^2$$

$$G = mgl\cos(\theta)$$

$$H = 0$$

$$m = 1kg,\ l = 1m,\ g = 9.81m/s^2$$

$$\ell(x,u) = k_\theta(\theta - \pi/2)^2 + k_{\dot{\theta}}(\dot{\theta})^2 + \frac{1}{2}u^2$$

Write dynamics in the standard form

$$x = \begin{bmatrix} \theta & \dot{\theta} \end{bmatrix}^T,$$

$$a(x) = \begin{bmatrix} \dot{\theta} \\ -J^{-1}(H\dot{\theta} + G) \end{bmatrix},$$

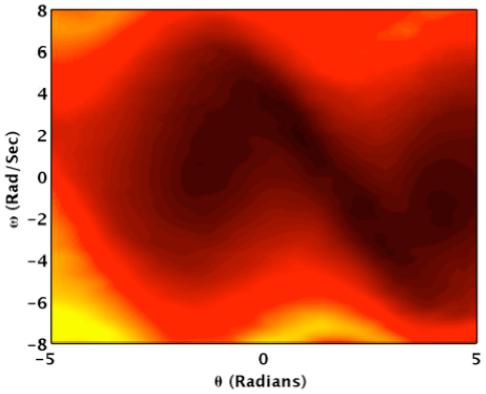
$$B = \begin{bmatrix} 0 \\ J^{-1} \end{bmatrix},$$

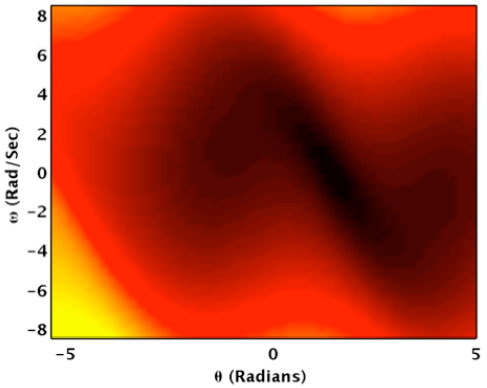
$$u = \tau.$$

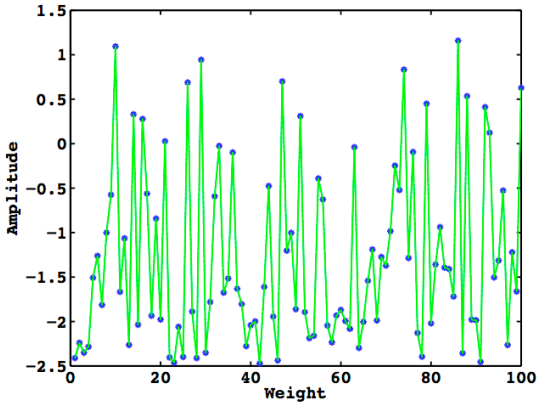
$$\dot{x} = a(x) + Bu(x)$$

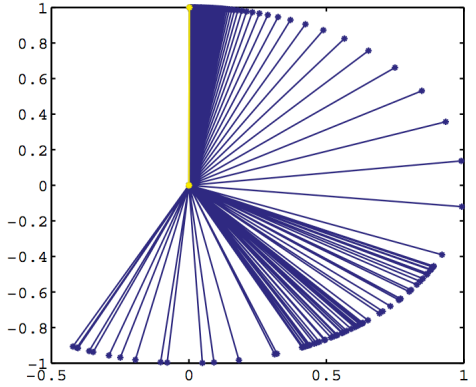
Results/settings

- 1000 basis functions
  - **Centers computed randomly**
  - **Widths set by abs-mean-dist-68% method**
- Typical error 1e-3Rad, t<5sec
- 100x100 grid
- 6sec vs. 5809sec









A higher dimensional example: a 1-dof pendulum with an uncertain wandering mapping

- Solving partially observable problems to which the separation principle does not apply
- Same problem but now consider an uncertain wandering mapping between the base angle and observation of the base angle
  - **In this case the gain of the sensor may be fluctuating**

$$dy = m(t)\theta(t)dt + d\omega_y$$

- This poses an additional problem which was addressed in our previous paper *(Simpkins and Todorov, 2008)*

Making the problem fully observable

- Take expectation of unobservable cost function

$$\ell(x,u) \approx E(k_\theta||m\theta - \pi/2||^2 + k_{\dot\theta}||\dot\theta||^2 + \frac{1}{2}||u||^2)$$

- Now we can estimate the unobservable parameter, and we have an exploration term

$$= k_\theta||\hat m\theta - \pi/2||^2 + k_\theta\theta^2\Sigma + k_{\dot\theta}||\dot\theta||^2 + \frac{1}{2}||u||^2$$

- We augment the state with the mean and covariance of the estimated quantity and the result is a higher dimensional but fully observable problem we can solve with the FAS

Kalman-Bucy filter

- Assume the prior over the initial state of the mapping is Gaussian, with mean $\widehat{m}(0)$ and covariance $\Sigma(0)$
  - **Then the posterior remains Gaussian for all t>0**
- Given the additive noise model, the optimal map estimate is propagated by the Kalman-Bucy filter

$$
\begin{aligned}
d\widehat{m} &= K\left(dy - \widehat{m}(t)\theta(t)dt\right), \\
K &= \Sigma(t)\theta(t)^{\mathsf{T}}\Omega_y^{-1}, \\
d\Sigma &= \Omega_m dt - K(t)\theta(t)\Sigma(t)dt.
\end{aligned}
$$

Augmenting the state

- We augment the state with the filter dynamics

  $x(t) = [\theta(t); \dot{\theta}(t); \widehat{m}(t); \Sigma(t)]$

- This yields a nonlinear stochastic optimal control problem which is fully observable
- No separation principle here
- We can approach this with our FAS algorithm

$$a(x) = \begin{bmatrix} \dot{\theta} \\ J^{-1}(-H\dot{\theta} - G(\theta, \dot{\theta})) \\ 0 \\ \Omega_m - \Sigma^2\theta^2\Omega_y^{-1} \end{bmatrix}$$

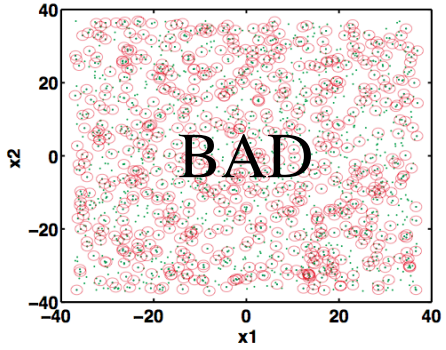
$$Bu = \begin{bmatrix} 0 & J^{-1}\tau & 0 & 0 \end{bmatrix}$$

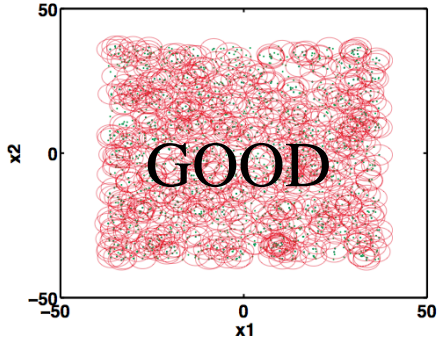
$$C(x) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \Sigma\theta\Omega_y^{-1} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

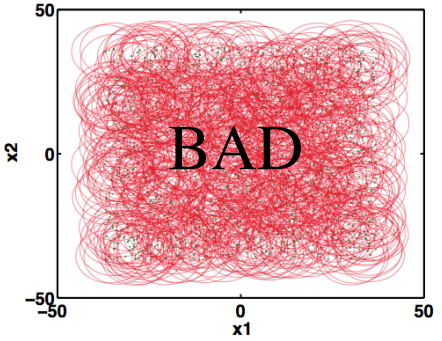
Choosing parameters

$$n_g = \prod_i \frac{p_i}{\sigma_i}$$

- How to choose number of parameters?
- Numerical stability
  - **Be sure to check the condition number of the M matrix before computing the least squares solution**
  - **Poorly conditioned matrices will lead to divergent results**
    - (at the very least poorly performing results)
    - Visualize Gaussian coverage two dimensions at a time or volumetrically (2sigma radius circles) - look for holes and reshuffle, add, or use a randomization scheme which takes into account the empty space remaining as each Gaussian center is generated
      - **Too much overlap causes the weights to 'fight' - alternating between very large positive and very large negative values**
      - **Tends to cause numerical problems and oscillations in the solution ('wavy')**

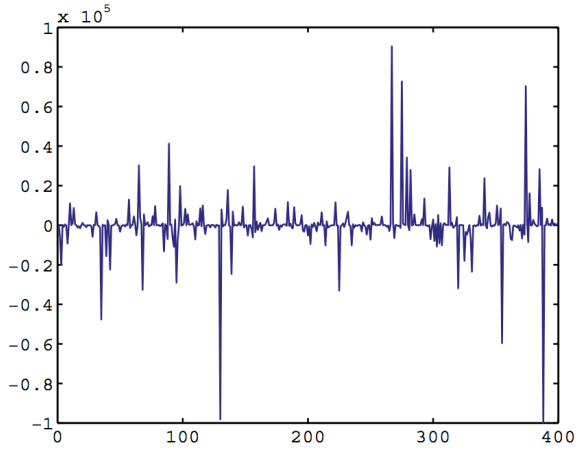


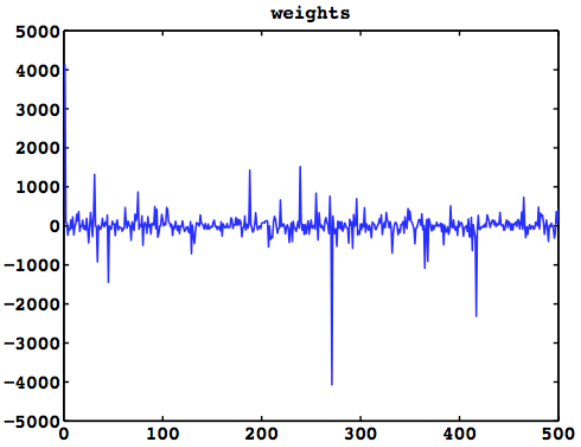




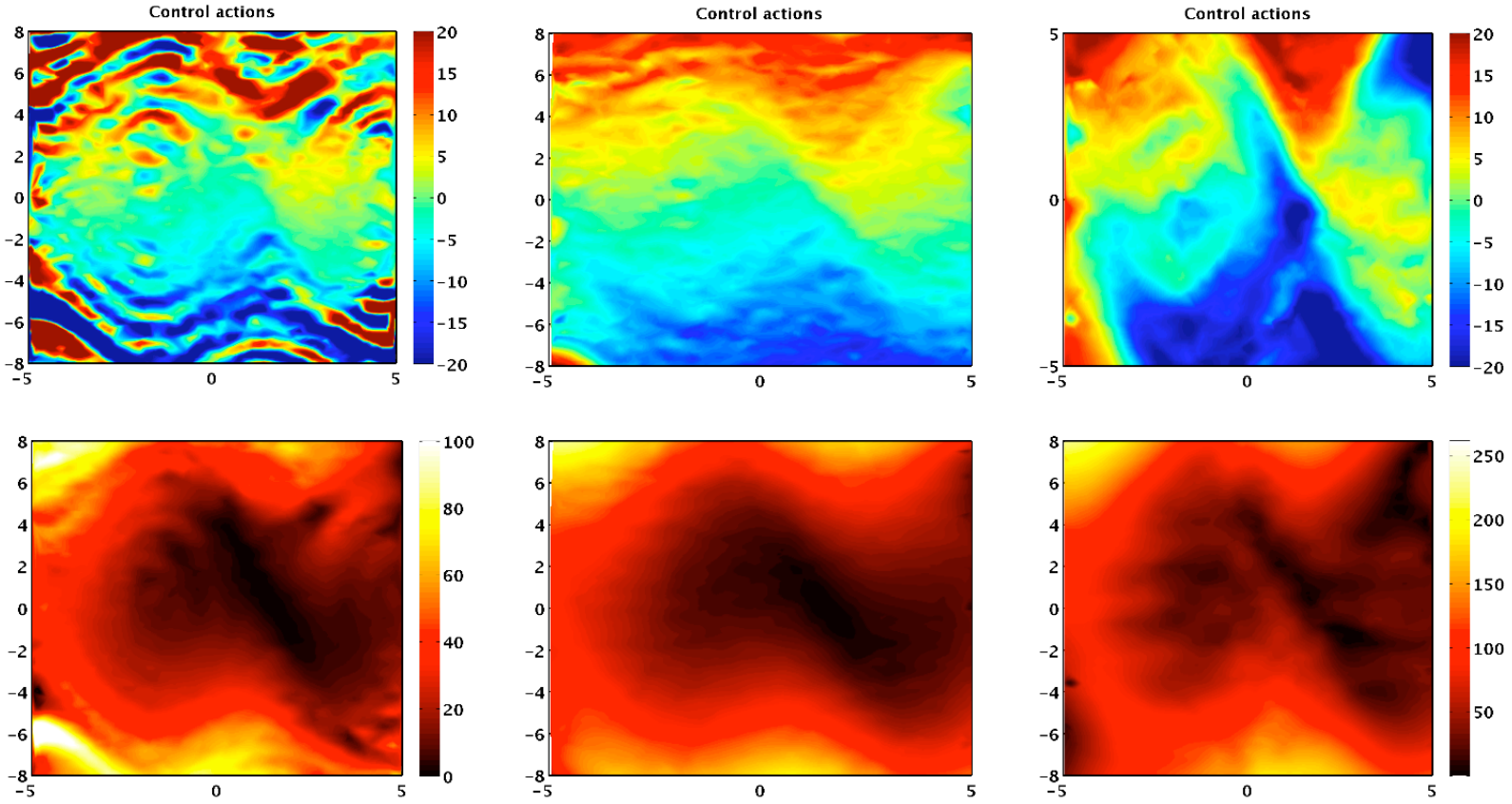
Choosing parameters (II)

- Choosing the widths of gaussians
  1. **Begin with a maximal number of gaussians, and use previous equation to compute the standard deviation**
  2. **Choose a number of Gaussians which reasonably divides the space which is most likely to have sharp or complex shapes**
     1. Compute mean absolute distance between centers, set the covariance to 68% of the max value



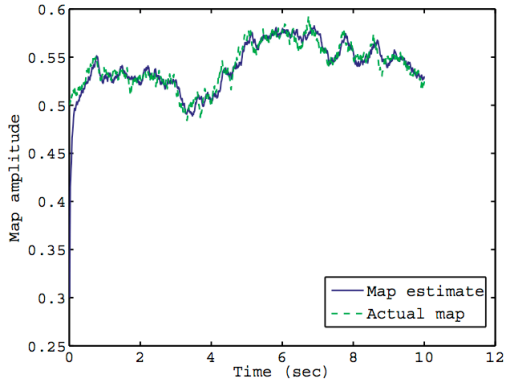


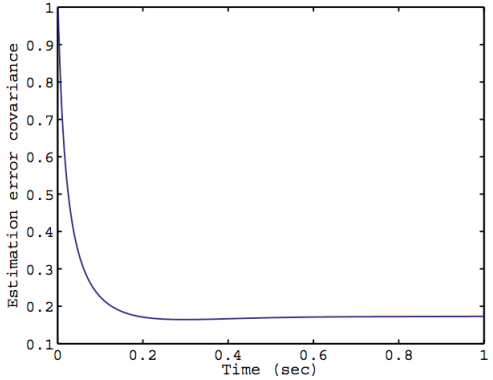
Visualize the value and policy functions

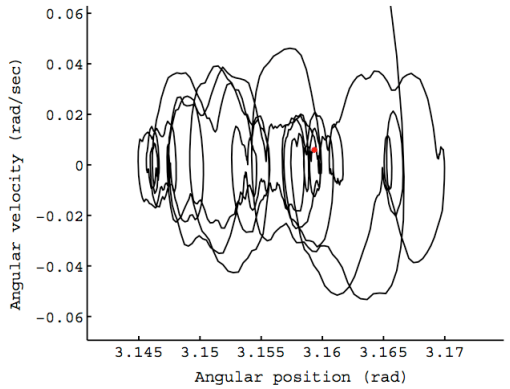


Convergence

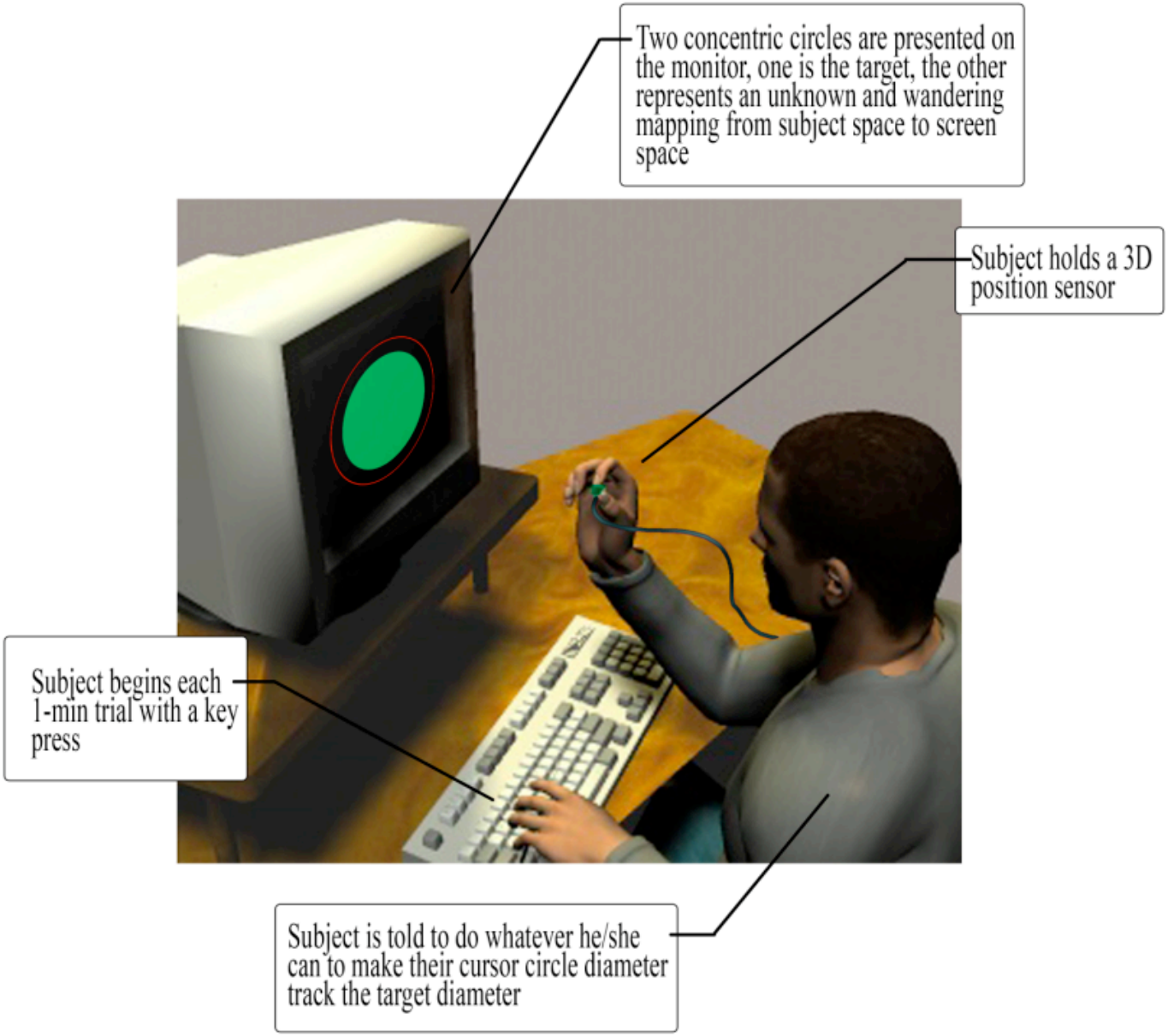
- Almost always <=2 iterations
  - **compare to nonlinear root finding such as newton's method which had to run overnight to converge and did not perform better**
  - **It is possible to make diverge**
    - Sometime the appearance of divergence
    - Renormalize weights if any issues with divergence
      - **Use different initial guess**


Map amplitude
Time (sec)
Map estimate
Actual map


Estimation error covariance
Time (sec)


Angular velocity (rad/sec)
Angular position (rad)

The task


Two concentric circles are presented on the monitor, one is the target, the other represents an unknown and wandering mapping from subject space to screen space
Subject holds a 3D position sensor
Subject begins each 1-min trial with a key press
Subject is told to do whatever he/she can to make their cursor circle diameter track the target diameter

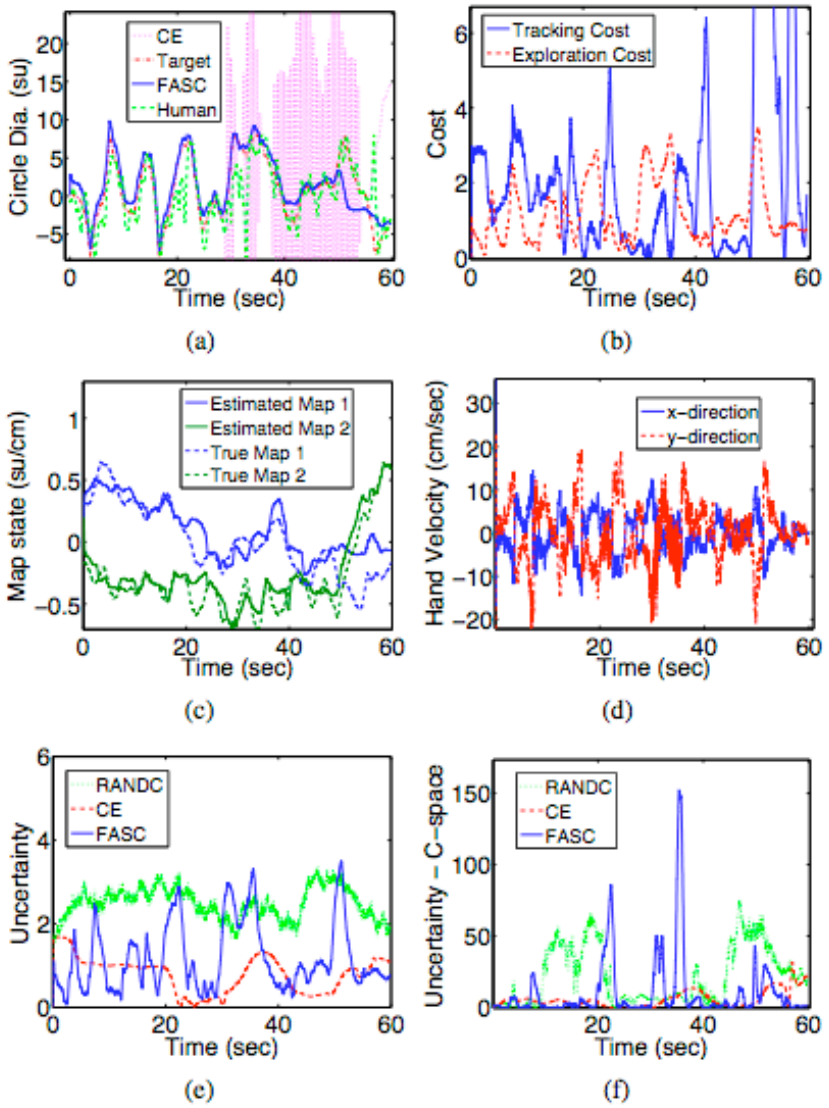


Fig. 2. (a) A section of a 60 second trial displaying human subject data, FAS, and proportional feedback controller tracking. (b) represents the two portions of the cost - the tracking and exploration costs. Plot (c) shows the true and estimated map. (d) shows the FAS control actions. (e) and (f) show two measures of uncertainty - (e) is in h-space, and is given by the trace term in 12, while (f) is the same, but in $h^{\perp}$-space.

TABLE I

$\sqrt{Norm}$ OF UNCERTAINTY QUANTITIES PLOTTED IN FIG. 2(E) AND 2(F), WHICH IS A STANDARD DEVIATION QUANTITY.

	h-space	$h^{\perp}$-space
FASC	8.8	36.7
CE	10.2	81.2
RAND	13.7	14.7