Learning for Control from Multiple Demonstrations

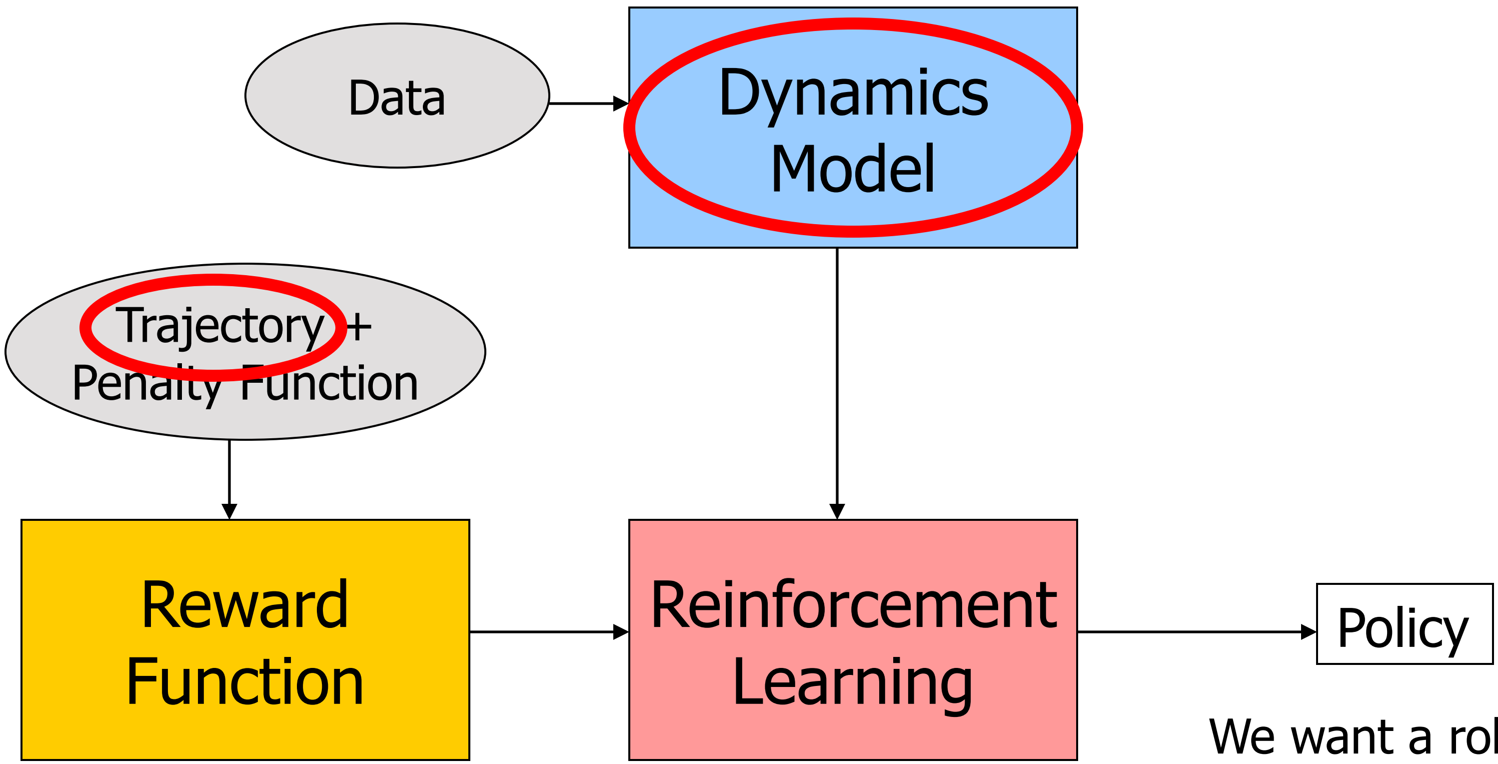
Adam Coates, Pieter Abbeel, and Andrew Y. Ng
Stanford University

ICML 2008

Motivating example

- How do we specify a task like this???

Introduction



We want a robot to follow a desired trajectory.

Key difficulties

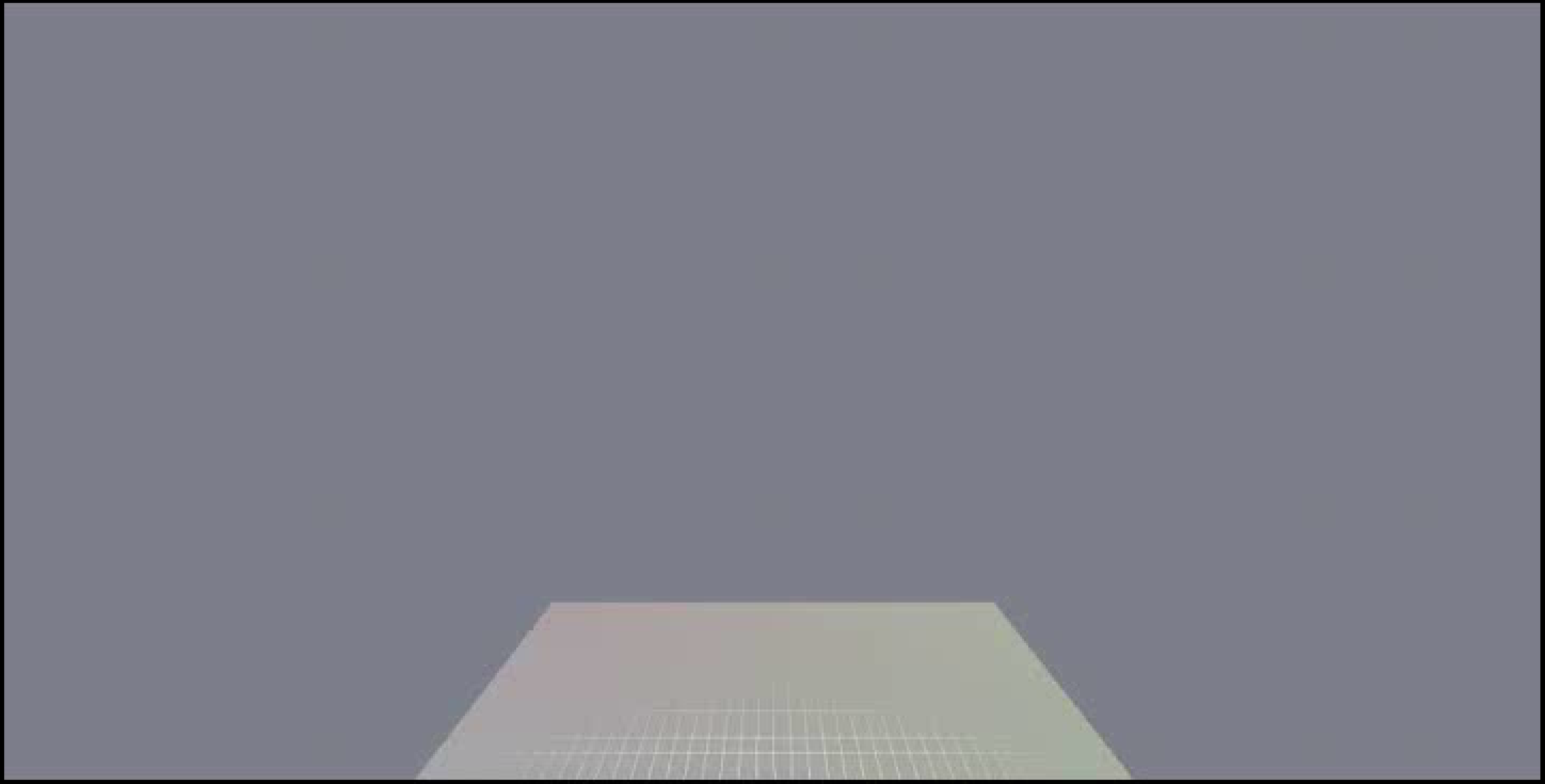
- Often very difficult to specify trajectory by hand.
  - Difficult to articulate exactly how a task is performed.
  - The trajectory should obey the system dynamics.
- ~~Use an expert *demonstration* as trajectory.~~
  - ~~But, getting perfect demonstrations is hard.~~
- Use multiple suboptimal demonstrations.

Outline

- Generative model for multiple suboptimal demonstrations.
- Learning algorithm that extracts:
  - Intended trajectory
  - High-accuracy dynamics model
- Experimental results:
  - Enabled us to fly autonomous helicopter aerobatics well beyond the capabilities of any other autonomous helicopter.

Adam Coates, Pieter Abbeel, Andrew Y. Ng

Expert demonstrations: Airshow



Graphical model

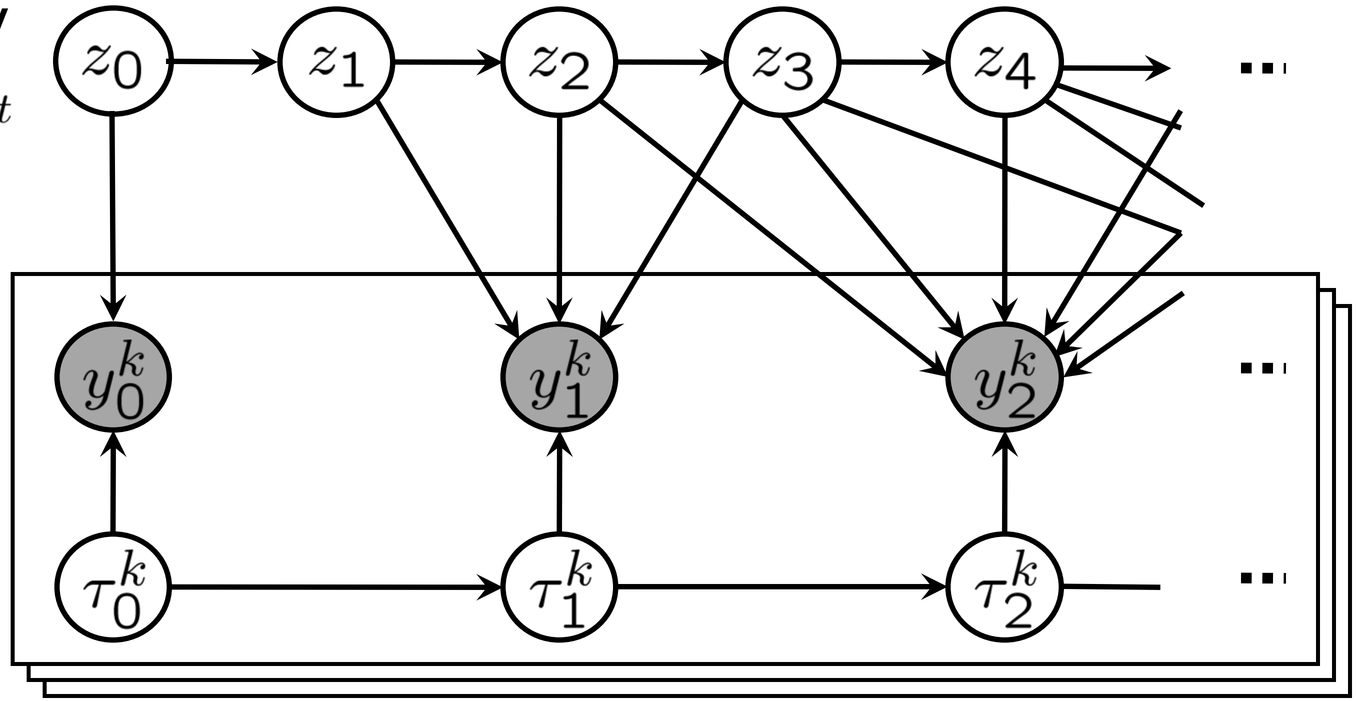
Intended trajectory

$$z_{t+1} = f(z_t) + \omega_t$$

Expert demonstrations

$$y_j = z_{\tau_j} + \nu_j$$

Time indices



- Intended trajectory satisfies dynamics.
- Expert trajectory is a noisy observation of one of the hidden states.
  - But we don't know exactly which one.

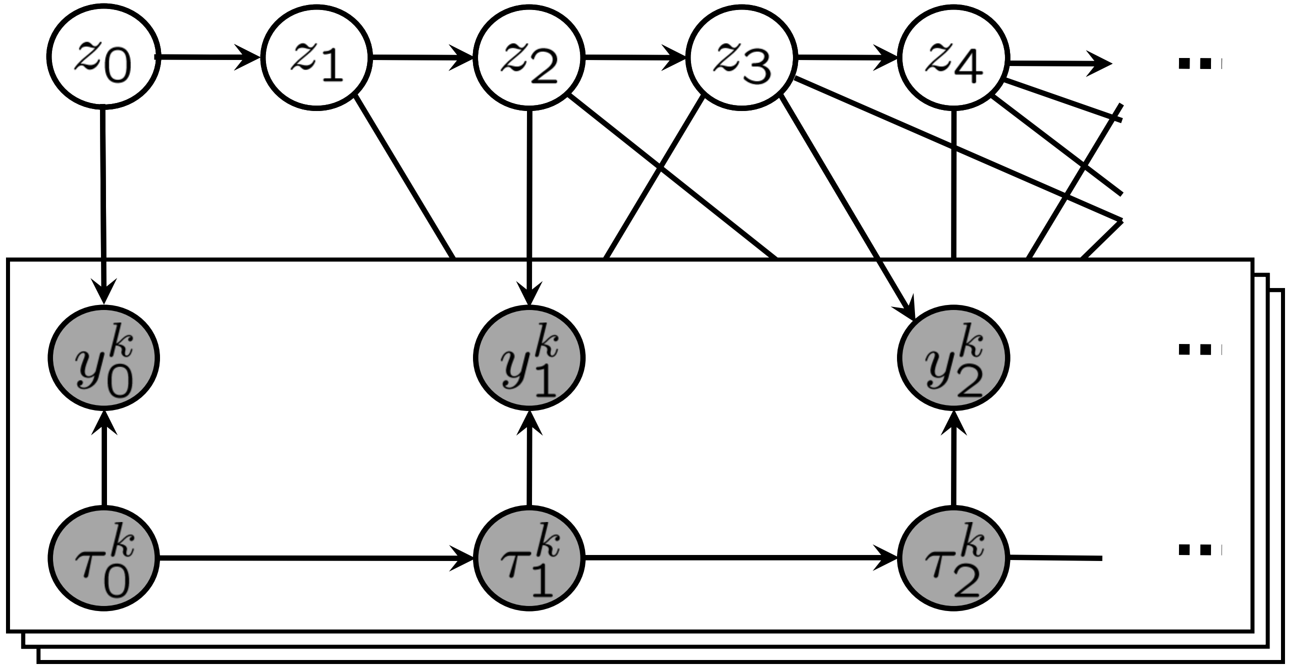
Learning algorithm

- Similar models appear in speech processing, genetic sequence alignment.
  - See, e.g., Listgarten et. al., 2005
- Maximize likelihood of the demonstration data over:
  - Intended trajectory states
  - Time index values
  - Variance parameters for noise terms
  - Time index distribution parameters

Learning algorithm

If τ is unknown, inference is hard.

If τ is known, we have a standard HMM.



- Make an initial guess for τ.
- Alternate between:
  - Fix τ. Run EM on resulting HMM.
  - Choose new τ using dynamic programming.

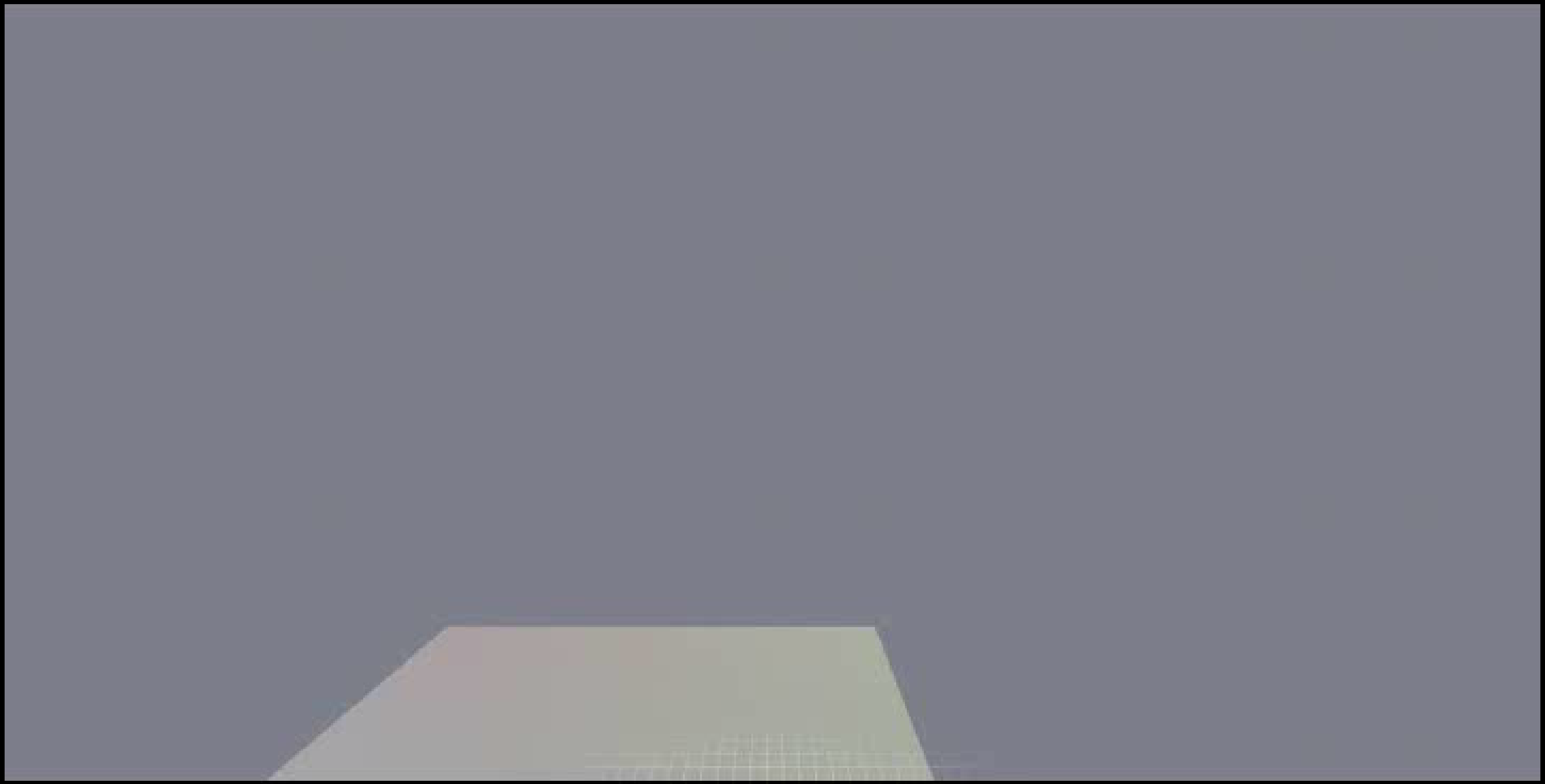
Details: Incorporating prior knowledge

- Might have some limited knowledge about how the trajectory should look.
  - Flips and rolls should stay in place.
  - Vertical loops should lie in a vertical plane.
  - Pilot tends to “drift” away from intended trajectory.

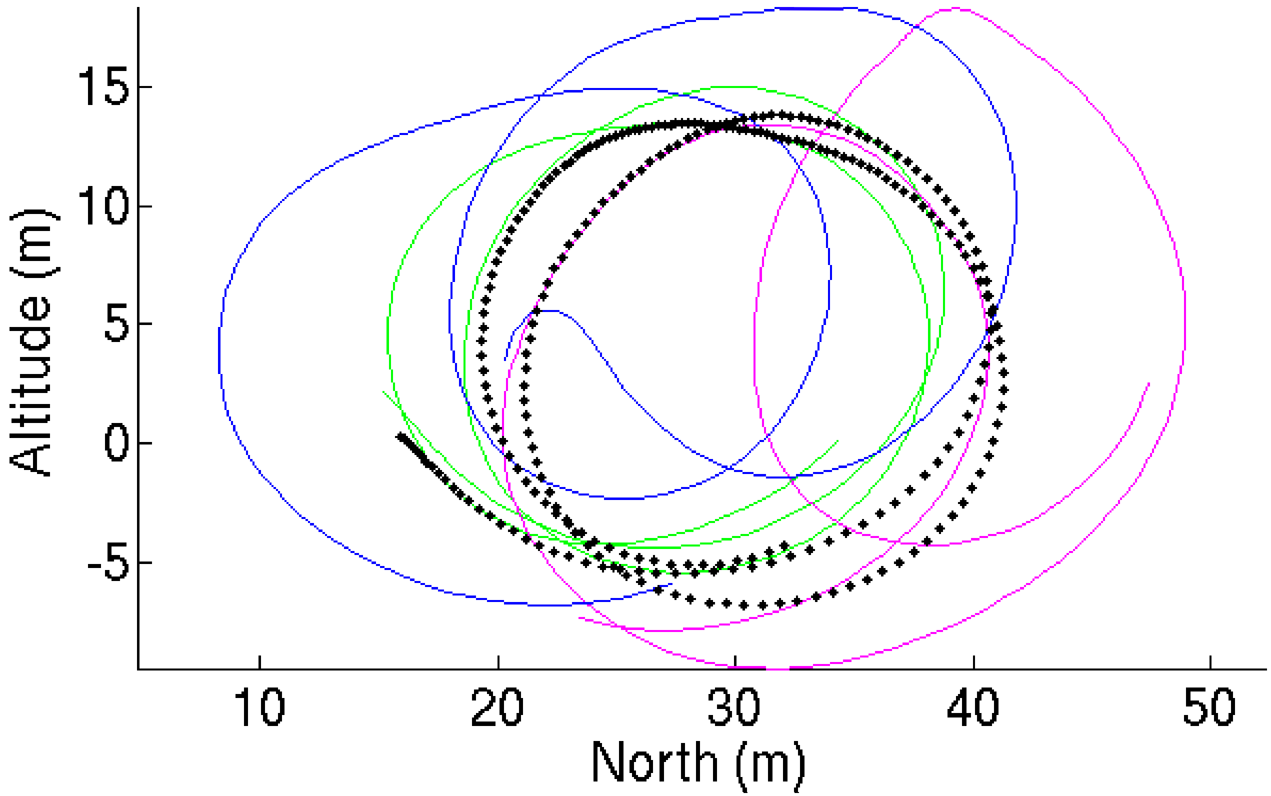
Adam Coates, Pieter Abbeel, Andrew Y. Ng

Results: Time-aligned demonstrations

- White helicopter is inferred “intended” trajectory.



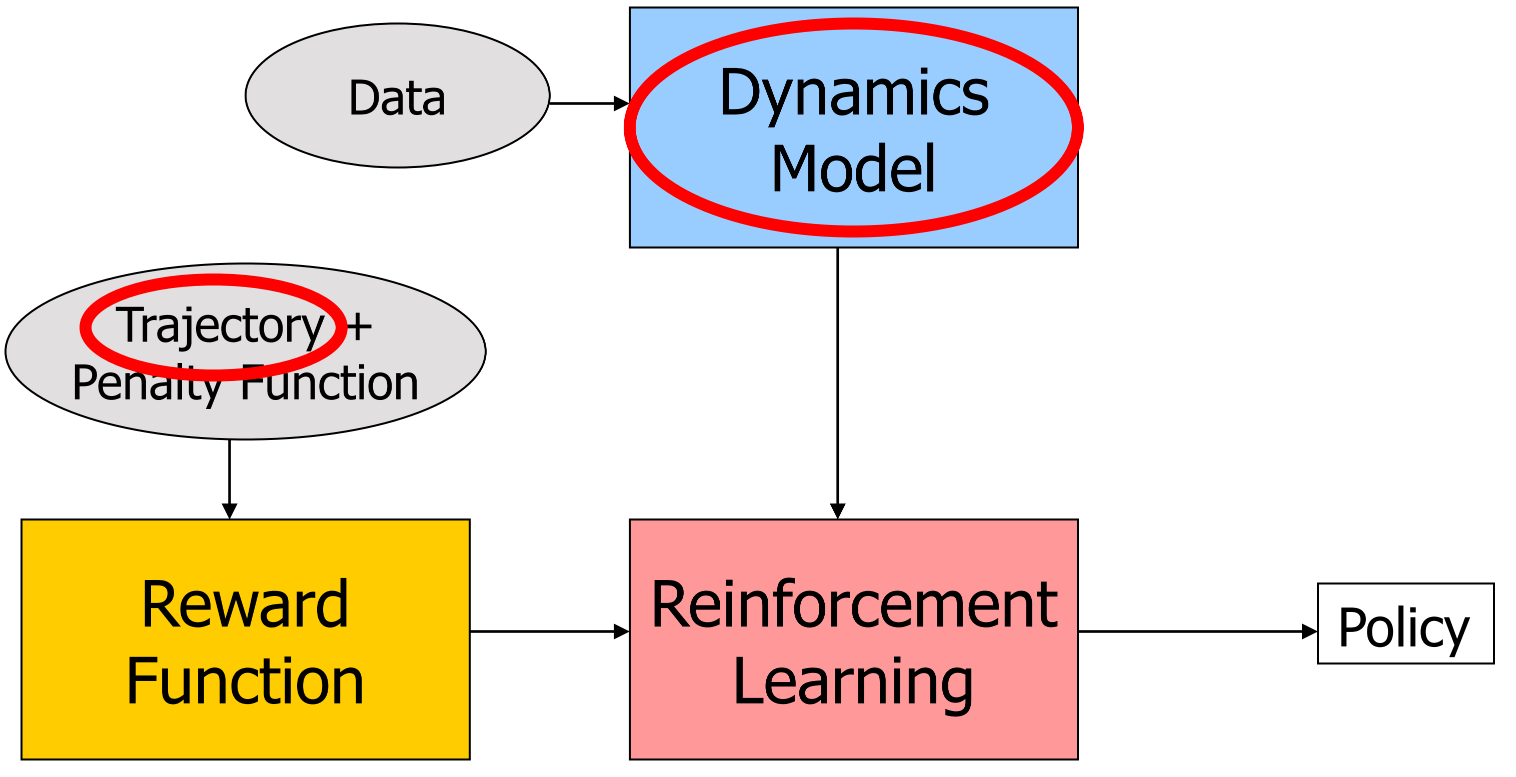
Results: Loops



- Even without prior knowledge, the inferred trajectory is much closer to an ideal loop.

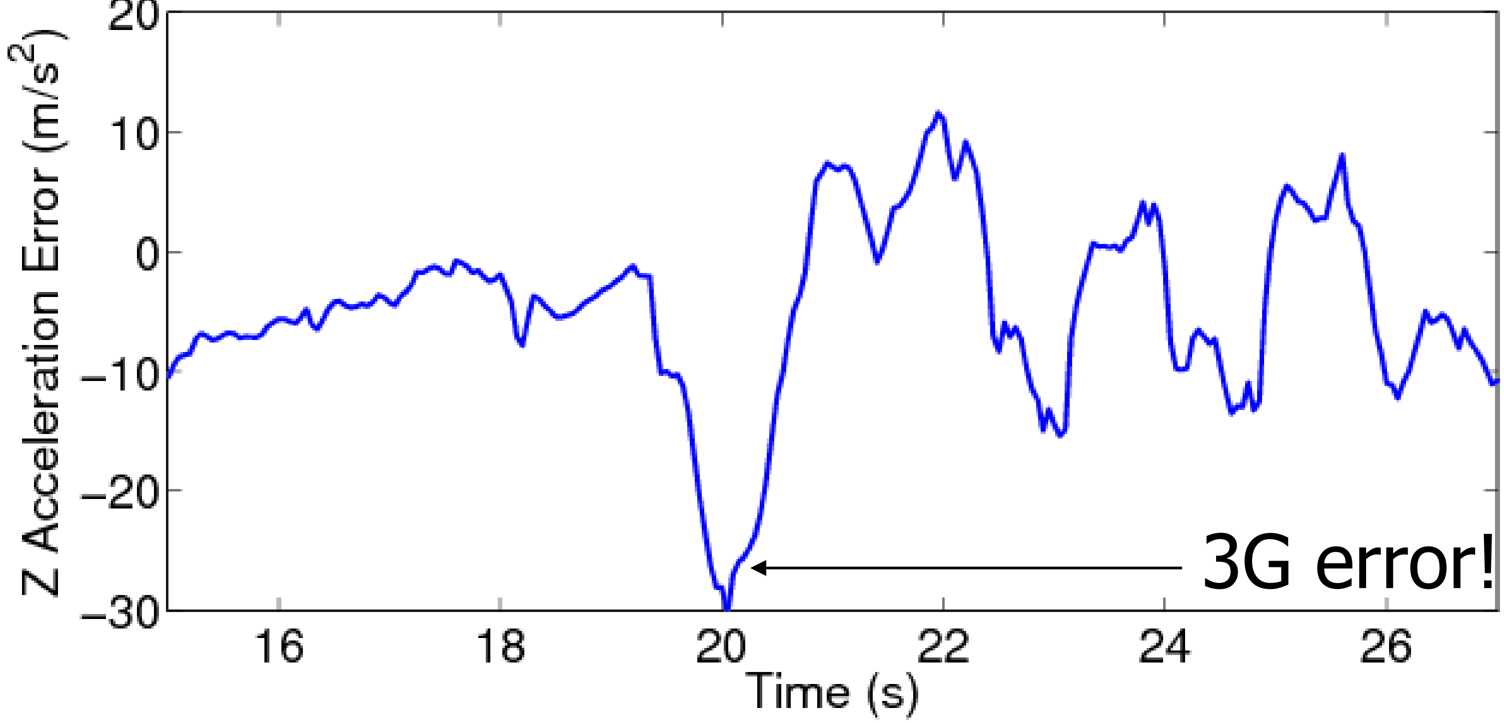
Adam Coates, Pieter Abbeel, Andrew Y. Ng

Recap


Data
Dynamics Model
Trajectory +
Penalty Function
Reward Function
Reinforcement Learning
Policy

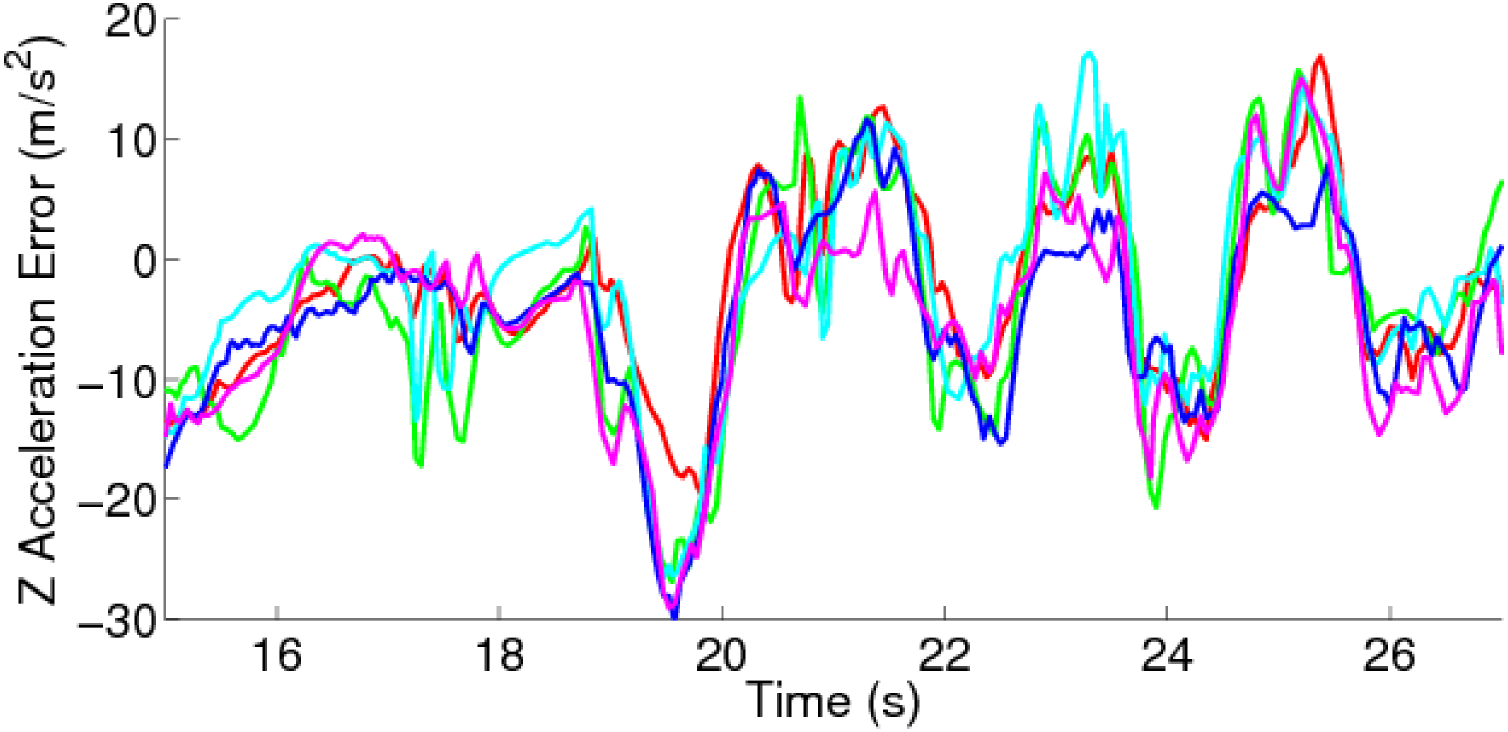
Standard modeling approach

- Collect data
  - Pilot attempts to cover all flight regimes.
- Build global model of dynamics



Adam Coates, Pieter Abbeel, Andrew Y. Ng

Errors aligned over time



- Errors observed in the "crude" model are clearly consistent after aligning demonstrations.

Adam Coates, Pieter Abbeel, Andrew Y. Ng

Model improvement

- Key observation:

  If we fly the same trajectory repeatedly, errors are consistent over time once we align the data.

  - There are many hidden variables that we can't expect to model accurately.
    - Air (!), rotor speed, actuator delays, etc.
  - If we fly the same trajectory repeatedly, the hidden variables tend to be the same each time.

Adam Coates, Pieter Abbeel, Andrew Y. Ng

Trajectory-specific local models

- Learn locally-weighted model from aligned demonstration data.
  - Since data is aligned in time, we can weight by *time* to exploit repeatability of hidden variables.
  - For model at time t: $W(t') = \exp(-(t - t')^2 / \sigma^2)$
- Suggests an algorithm alternating between:
  - Learn trajectory from demonstration.
  - Build new models from aligned data.
- Can actually infer an improved model *jointly* during trajectory learning.

Experiment setup

- Expert demonstrates an aerobatic sequence several times.
  - Inference algorithm extracts the intended trajectory, and local models used for control.

- We use a receding-horizon DDP controller.
  - Generates a sequence of closed-loop feedback controllers given a trajectory + quadratic penalty.

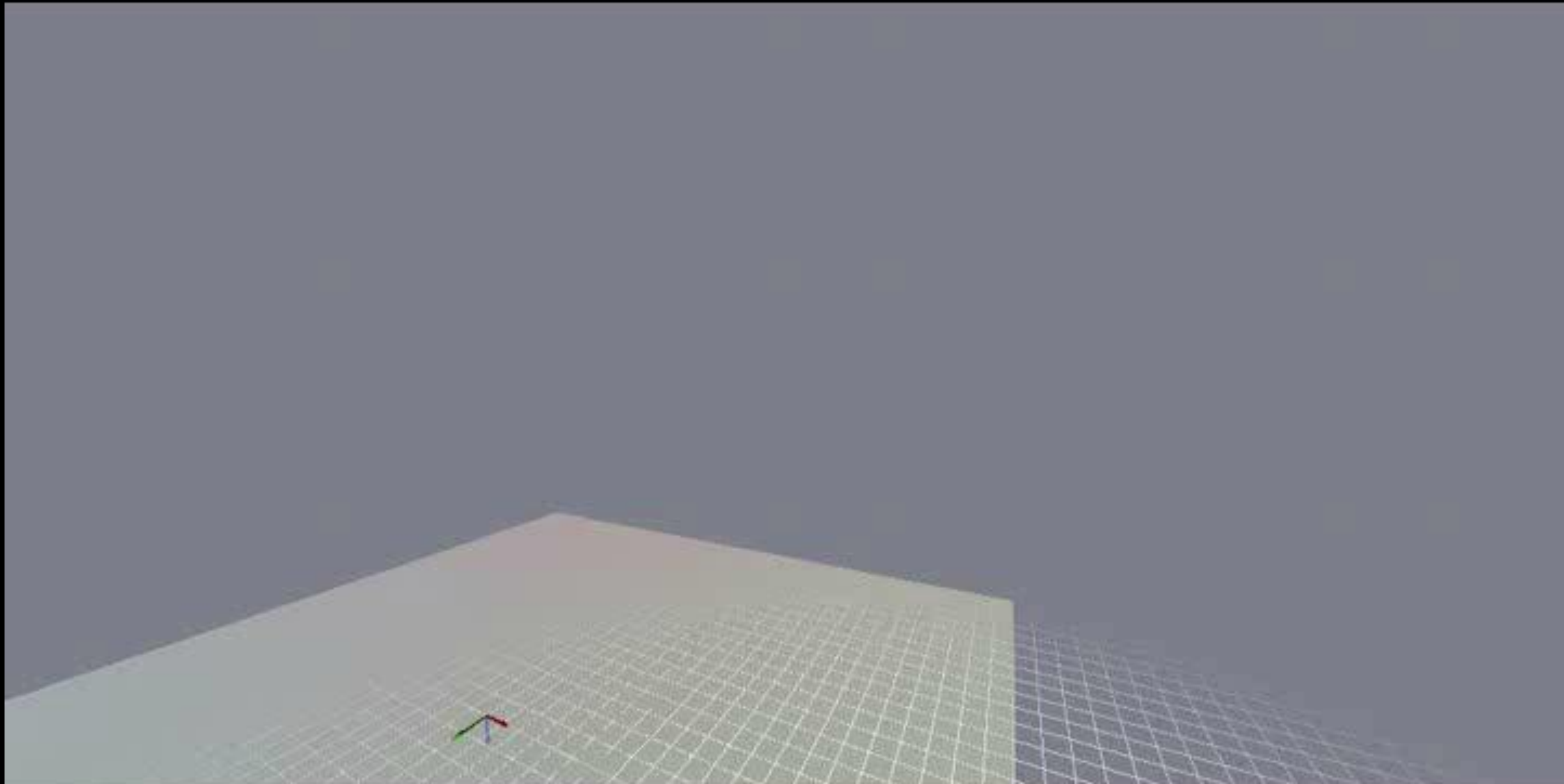
Related work

- Bagnell & Schneider, 2001; LaCivita, Papageorgiou, Messner & Kanade, 2002; Ng, Kim, Jordan & Sastry 2004a (2001);
- Roberts, Corke & Buskey, 2003; Saripalli, Montgomery & Sukhatme, 2003; Shim, Chung, Kim & Sastry, 2003; Doherty et al., 2004.
- Gavrilets, Martinos, Mettler and Feron, 2002; Ng et al., 2004b.
- Abbeel, Coates, Quigley and Ng, 2007.

- *Maneuvers presented here are significantly **more challenging** and **more diverse** than those performed by any other autonomous helicopter.*

Results: Autonomous airshow

Results: Flight accuracy



Conclusion

- Algorithm leverages multiple expert demonstrations to:
  - Infer intended trajectory
  - Learn better models along the trajectory for control.

- First autonomous helicopter to perform extreme aerobatics at the level of an expert human pilot.

Adam Coates, Pieter Abbeel, Andrew Y. Ng