Reinforcement Learning with Multiple Demonstrations
Adam Coates, Pieter Abbeel, Andrew Y. Ng

Overview

• Task: have a robot follow a desired trajectory.
• Key challenge: How to specify the desired trajectory.
  • By hand is extremely difficult for systems with complex dynamics, such as helicopters.
  • Even experts are often incapable of performing the true desired trajectory.
• We present a generative model and learning algorithm for learning the underlying, “ideal” trajectory from multiple suboptimal demonstrations.
• Our experimental results significantly extend the state-of-the-art in aerobatic helicopter flight; our resulting controllers perform the first autonomous tic-tocs and chaos---two of the most challenging aerobatic maneuvers.

Basic Generative Model

The latent, “ideal” trajectory \( x_0, x_1, \ldots \) has to satisfy the (stochastic) system dynamics:
\[
x_{t+1} = f(x_t) + \omega_t, \quad \omega_t \sim \mathcal{N}(0, \Sigma)
\]
• The /th state of the k th demonstration \( y_t \) is a noisy observation of a state in the hidden ideal trajectory at some time index \( v_k \): \( y_t = \theta_v(x_{tv}) + v_t, \quad v_t \sim \mathcal{N}(0, \Phi) \)
• “Time-warping” indices, which map demonstration states to hidden states, are distributed according to \( v_t \sim \mathcal{N}(\mu_v, \Sigma_v) \)

Incorporating Prior Knowledge

1. Position and heading drift in demonstrations Augment hidden state vector with position and heading drift variables. The dynamics model for these state variables enforces slow variation.
2. Prior knowledge about the trajectory Add an observation model, which directly observes the hidden trajectory. E.g., during loops we add observations of the hidden state in the loop’s plane. (Similarly for rolls, flips, tic-tocs, hurricanes, etc.)
3. Inferring an improved model: State vector is augmented with model error variables which can vary slowly over time. These variables will capture model inaccuracies that are consistently observed in demonstrations.

Challenges in Typical Demonstrated Trajectories

1. Expert demonstrations don’t always follow the same path.
2. Expert demonstrations are not consistent over time.
3. Demonstrations are suboptimal. E.g., a perfect loop need only move forward and up, never sideways.
4. We only have an approximate model for complex systems.

Experimental Results

Trajectory Learning

<table>
<thead>
<tr>
<th>Time-aligned demonstrations</th>
<th>Inferred ideal trajectory</th>
</tr>
</thead>
</table>
1. Inferred trajectory captures expert’s ideal path. | 2. Time warping removes variations from expert timing. |
3. Incorporating priors allows algorithm to enforce known constraints. | 4. Time-varying corrections improve the original model along the trajectory. |

Helicopter Control

Results include autonomous tic-tocs and chaos, arguably two of the most challenging aerobatic maneuvers.

Our results greatly extend the state-of-the-art in aerobatic helicopter flight.

Videos Available

Conclusion

• We presented a generative model and a learning algorithm to infer the underlying ideal trajectory from multiple suboptimal trajectory demonstrations.
• Our experimental results (i) successfully extract the ideal trajectory from wildly different demonstrations, and (ii) vastly outperform the state-of-the-art in aerobatic helicopter flight.

Related Work

• V. Gavrilets, I. Martinos, B. Mettler, and E. Feron, Control logic for augmented aerobatic flight of a miniature helicopter, AIAA GNCC 2002.