Reinforcement Learning with Multiple Demonstrations

Adam Coates, Pieter Abbeel, Andrew Y. Ng
Department of Computer Science
Stanford University

Many tasks in robotics can be described as a trajectory that the robot should follow. Unfortunately, specifying the desired trajectory is often a non-trivial task. For example, when asked to describe the trajectory that a helicopter should follow to perform an aerobatic flip, one would have to not only (a) specify a complete trajectory in state space that intuitively corresponds to the aerobatic flip task, but also (b) ensure that the state space trajectory is consistent with the helicopter's dynamics. This is a non-trivial task for systems with complicated dynamics.

In the apprenticeship learning setting, where an expert is available, one can instead have the expert demonstrate the desired trajectory. Unfortunately, this means that we must have an essentially optimal expert available—since any learned controller, at best, will only be able to repeat the demonstrated trajectory. Such a perfect demonstration may be hard, if not impossible, to acquire. However, even suboptimal expert demonstrations often embody many of the desired qualities. Even stronger, repeated expert demonstrations are often suboptimal in different ways, suggesting that a large number of suboptimal expert demonstrations could implicitly encode the optimal demonstration. In this piece of work we propose an algorithm that approximately extracts this implicitly encoded optimal demonstration from multiple suboptimal expert demonstrations. In doing so, the algorithm learns a target trajectory that not only mimics the behavior of the expert, but can even be significantly better.

The problem of extracting the underlying ideal trajectory from a set of suboptimal trajectories is not a matter of merely averaging the states observed at each time-step. A simple arithmetic average of the states would result in a trajectory that does not obey the dynamic constraints of the model. Also, in practice, each of the demonstrations will occur at different rates so that attempting to combine states from the same time-step in each trajectory will not work properly.

Our algorithm uses a generative model that describes the expert demonstrations as noisy observations of the hidden, optimal target trajectory, with each demonstration possibly occurring at a different rate. An EM algorithm is developed to infer the hidden target trajectory and the necessary model parameters using a Kalman smoother and an efficient dynamic programming algorithm to perform the E-step.

We also show how prior knowledge can be easily incorporated to further improve the quality of the resulting "averaged" trajectory. For example, often only an approximate dynamics model is known, and our algorithm can estimate an improvement to the general dynamics model specific to the trajectory being performed by incorporating data from multiple demonstrations. Our formulation also allows us to take out known expert flaws. For example, for a helicopter performing in-place flips, it is known that the helicopter can be centered around the same position over the entire sequence of flips. Our model incorporates this prior knowledge, and factors out the position drift in the expert demonstrations.

Our experimental results show that the resulting trajectories are not only good, feasible trajectories that can be used in reality, but also that the resulting performance meets or exceeds that of the expert (as evaluated by our expert helicopter pilot). The presented algorithm significantly extends the state of the art in aerobatic helicopter flight ([1], [4]). Specifically, the learned trajectories resulted in significantly better in-place flips and rolls than previously possible. The presented algorithm also resulted in the first autonomous tic-tocs, a maneuver considered even more challenging than flips and

rolls. Movies of the flight results can be found at the Stanford Autonomous Helicopter homepage:
http://www.cs.stanford.edu/groups/helicopter

Related work. In recent work on apprenticeship learning and inverse reinforcement learning ([2], [5], [7], [6]), the reward function is assumed be a linear combination of a known set of features (rather than being defined by a trajectory), and the weighting of the features is then estimated from expert demonstrations. Most similar to our work, Atkeson and Schaal [3] also estimate the desired trajectory (for a pendulum swing-up task) from a demonstration. However, they learn from a single demonstration only, which can significantly limit the performance obtained (or, equivalently, increase the requirements on the expert) as discussed in previous paragraphs.

References

[1] P. Abbeel, A. Coates, M. Quigley, and A. Y. Ng. An application of reinforcement learning to aerobatic helicopter flight. In *NIPS 19*, 2007.

[2] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proc. ICML*, 2004.

[3] Christopher G. Atkeson and Stefan Schaal. Robot learning from demonstration. In *Proc. ICML*. Morgan Kaufmann, 1997.

[4] V. Gavrilets, I. Martinos, B. Mettler, and E. Feron. Control logic for automated aerobatic flight of miniature helicopter. In *AIAA Guidance, Navigation and Control Conference*, 2002.

[5] Gergely Neu and Csaba Szepesvari. Apprenticeship learning using inverse reinforcement learning and gradient methods. In *Proc. UAI*, 2007.

[6] N. Ratliff, J. Bagnell, and M. Zinkevich. Maximum margin planning. In *Proc. ICML*, 2006.

[7] N. Ratliff, D. Bradley, J. Bagnell, and J. Chestnutt. Boosting structured prediction for imitation learning. In *Neural Information Processing Systems 19*, 2007.