(Inverse) Optimal Control for Matching Higher-Order Moments

Oleg Arenz, Hany Abdulsamad and Gerhard Neumann

***Abstract*— Defining the cost function for optimal control manually is often cumbersome. Instead it is often easier to demonstrate the desired behavior and learn the cost function using inverse optimal control. When aiming to match distributions over features, state of the art methods for inverse optimal control often suffer dramatically as they are not designed for matching higher order moments. We therefore present a new approach for inverse optimal control that is tailored for matching distributions by minimizing the relative entropy to a distribution over features. This distribution could be either estimated based on expert demonstrations for inverse optimal control, or defined manually for optimal control.**

I. Motivation and Problem Definition

Optimal control computes optimal behavior based on a given cost function and the dynamics of the system. However, manually defining a cost function such that the optimal controller solves the desired task is often cumbersome and has to be done by experts.

It is often easier to specify a task by providing desired distributions over features. Such distributions can be either estimated from human demonstration which results in imitation learning, or they can be specified manually. In both cases, better generalizations can be achieved by employing a two-staged approach that first infers a cost function via inverse optimal control and then computes the desired behavior via optimal control.

Many approaches for inverse optimal control are based on matching feature counts in expectation which directly translates to matching the mean of the observed feature distribution. Such methods can also be applied for matching higher-order moments by additionally matching the corresponding products of features in expectation. However, the feature dimension increases significantly in that progress, drastically impairing the performance of the algorithm.

Observing that adding artificial features in that manner looses all information about the relation between the original and the additional features suggests that this class of problems can be solved more efficiently by directly formulating the inverse optimal control problem of matching feature distributions.

II. Related Work

Yin et al. [1] demonstrate robotic handwriting by first solving the inverse optimal control problem via stochastic optimization and afterwards using optimal control to compute the controller. However, their inverse optimal control formulation does not take into account the system dynamics during demonstration and hence the resulting controller does not achieve the desired distribution over trajectories.

Englert et al. 2013 [2] directly learn a controller for matching trajectory distributions by framing imitation learning as a policy search problem. They use the Kullback-Leibler divergence between the induced trajectory distribution of the controller and the target distribution as long-term cost function and optimize it using the model-based policy search algorithm PILCO [3]. However, in contrast to our approach, the optimization is non-convex and computationally heavy.

III. Own Approach and Contribution

Our approach strongly relates to Maximum Entropy Inverse Reinforcement Learning (MaxEnt-IRL) [4], which produces a stochastic controller with maximum entropy while matching the demonstrated feature counts in expectation. However, instead of matching feature counts, we propose to minimize the relative entropy to the demonstrated feature distribution. The corresponding objective of the resulting optimization problem is given by

$$\underset{\pi_t(\mathbf{a}|\mathbf{s})}{\text{maximize}} \sum_{t=1}^{T-1} H(\pi_t(\mathbf{a}|\mathbf{s})) - \sum_{t=2}^{T} \beta_t D_{\text{KL}}\left(p_t(\boldsymbol{\phi})||q_t(\boldsymbol{\phi})\right),$$

where $H(\pi_t(\mathbf{a}|\mathbf{s}))$ denotes the entropy of the controller, β_t is a regularization coefficient trading off the oposing objectives, and $D_{\text{KL}}\left(p_t(\boldsymbol{\phi})||q_t(\boldsymbol{\phi})\right)$ denotes the relative entropy between the feature distribution produced by π, $p_t(\boldsymbol{\phi})$, and the empirical distribution $q_t(\boldsymbol{\phi})$.

Interestingly, solving this optimization problem yields the same gradient as MaxEnt-IRL for $\beta_t \to \infty$, but also yields a closed form estimate of the reward function

$$\tilde{r}_t(\boldsymbol{\phi}) = \beta_t \left(\log q_t(\boldsymbol{\phi} - \log \tilde{p}_t(\boldsymbol{\phi})\right) + \textit{const}\,,$$

based on the current estimate of the feature distribution, $\tilde{p}_t(\boldsymbol{\phi})$. Optimization based on this estimate is several thousand times faster than L-BFGS based on the gradient in our preliminary experiments.

References

[1] H. Yin, A. Paiva, and A. Billard, "Learning Cost Function and Trajectory for Robotic Writing Motion," in *14th IEEE-RAS International Conference on Humanoid Robots (Humanoids), 2014*. IEEE, 2014.

[2] P. Englert, A. Paraschos, J. Peters, and M. P. Deisenroth, "Model-based Imitation Learning by Probabilistic Trajectory Matching," in *IEEE International Conference on Robotics and Automation*, 2013.

[3] M. Deisenroth and C. Rasmussen, "PILCO: A Model-Based and Data-Efficient Approach to Policy Search," in *28th International Conference on Machine Learning (ICML)*, 2011.

[4] B. D. Ziebart, J. A. Bagnell, and A. K. Dey, "Modeling Interaction via the Principle of Maximum Causal Entropy," in *Proceedings of the Twenty-seventh International Conference on Machine Learning*, 2010.