Policy Decomposition: Approximate Optimal Control with Suboptimality Measure

Ashwin Khadke (akhadke@andrew.cmu.edu) and Hartmut Geyer (hgeyer@cmu.edu)
The Robotics Institute, Carnegie Mellon University

I. Introduction

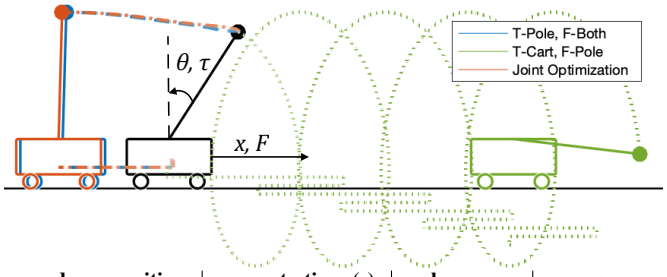
Dynamic programming (DP) is often applied to solve control problems in robotics. For complex robotic systems with many degrees of freedom, only approximate DP methods are computationally tractable. Several such approximations have been proposed, relying on either local search or state-space reductions [1–3]. However, these strategies do not yield an associated measure to gauge the suboptimality of the resulting control. We instead propose and explore policy decomposition, an alternative approximation strategy that includes such a measure. The measure indicates the similarity of the resulting closed-loop behavior to the one of the optimal control solution.

II. Policy Decomposition

We propose to generate control policies for complex systems based on cascaded, lower dimensional problems whose order is identified algorithmically and maximizes the similarity of the resulting control policy to the one that would be obtained if the complex system was computationally tractable.

Consider, for example, designing a control policy to swing-up a pole on a cart while moving the cart to a goal (Fig. 1). This system has two degrees of freedom (pole angle θ and cart position x) and two inputs (pole torque τ and cart force F). Although the example is simple and its control optimization tractable, imagine it were not. To approximate the optimization problem, we could first design an inner policy $\tau_\pi(\theta, \dot{\theta})$ for τ to swing up the pole assuming the cart is arrested and then design an outer policy $F_\pi(\theta, \dot{\theta}, x, \dot{x}, \tau_\pi)$ for F to move the cart with the torque control of the pole frozen to τ_π. As an alternative to this cascade, we could treat the cart and pole as separate subsystems and design control policies $\tau_\pi(\theta, \dot{\theta})$ and $F_\pi(x, \dot{x})$ independently. Both policy decompositions (and the four other possible ones, Fig. 1) reduce the dimensionality of the problem and make it computationally much more tractable. But the quality of the resulting control for the entire cart-pole differs considerably among them (examples shown in Fig. 1).

To measure and *predict* how closely a possible policy decomposition will approximate the optimal control of a complex nonlinear system $\dot{\boldsymbol{x}} = f(\boldsymbol{x}, \boldsymbol{u})$, we consider its corresponding linear system. The linear system approximates the dynamics of the original one around the goal state $\boldsymbol{x}_0$, $\dot{\boldsymbol{x}} = \frac{\partial f}{\partial \boldsymbol{x}}(\boldsymbol{x} - \boldsymbol{x}_0) + \frac{\partial f}{\partial \boldsymbol{u}}(\boldsymbol{u} - \boldsymbol{u}_0)$. Assuming quadratic costs, we design analogous policy decompositions for this linear system and readily compute their optimal control policies and value functions. We then use the difference between these analogous value functions and the optimal value function of the entire linear system as a measure to gauge the suboptimality of the controls obtainable from the corresponding policy decompositions of the original, nonlinear system.



	decomposition	compute time (s)	value error	measure
1:	F-cart, τ-both	581	0.5	1.26×10^{-2}
2:	τ-pole, F-both	565	0.15	1.1×10^{-2}
3:	τ-cart, F-both	543	470.1	0.53
4:	F-pole, τ-both	589	1.6	0.14
5:	F-cart, τ-pole	9	1.1	2.8×10^{-2}
6:	τ-cart, F-pole	16	994.4	∞
	entire cart-pole	8487	0	0

Fig. 1. Suboptimality of cart-pole control from different policy decompositions. **(Top)** Example trajectories for best (2, blue traces) and worst decomposition (6, green), and optimal control (red). **(Bottom)** Compute times, true value function error, and measure of error from analogous linear system.

III. Initial Results and Future Work

Our approach works well for the cart-pole example (Fig. 1). All policy decompositions largely reduce the compute time, but the resulting closed-loop behavior differs significantly. The example trajectories visualize the difference (top panel) and the mean deviations from the true optimal value function quantify it (bottom, column 3). The predictive measure correlates well with this error, although the two do not map proportionally (column 4). Thus, the measure could help to not only decide on suitable policy decompositions but also gauge the best ones similarity to the true optimal control. We currently study how well these findings generalize from the cart-pole to more complex systems, including legged systems.

References

[1] E. Todorov and W. Li, "A generalized iterative lqg method for locally-optimal feedback control of constrained nonlinear stochastic systems," in *Proceedings of the 2005, American Control Conference, 2005.* IEEE, 2005, pp. 300–306.

[2] R. Tedrake, I. R. Manchester, M. Tobenkin, and J. W. Roberts, "Lqr-trees: Feedback motion planning via sums-of-squares verification," *Int. J. Rob. Res.*, vol. 29, no. 8, p. 1038–1052, Jul. 2010. [Online]. Available: https://doi.org/10.1177/0278364910369189

[3] A. A. Gorodetsky, S. Karaman, and Y. M. Marzouk, "Efficient high-dimensional stochastic optimal motion control using tensor-train decomposition." in *Robotics: Science and Systems*, 2015.