Disentangling Sources of Uncertainty for Active Exploration

David Lines

Department of Engineering, University of Cambridge

UNIVERSITY OF CAMBRIDGE

Introduction

PILCO, a model-based reinforcement learning algorithm [2], offers state-of-the-art data efficiency for controlling mechanical dynamical systems (see Figure 1) despite the use of greedy policy selection. This project takes a Bayesian approach to the exploration-exploitation trade-off by quantifying the epistemic and aleatoric uncertainty in the transition and loss functions. These values are then used to identify areas of high value for active exploration.

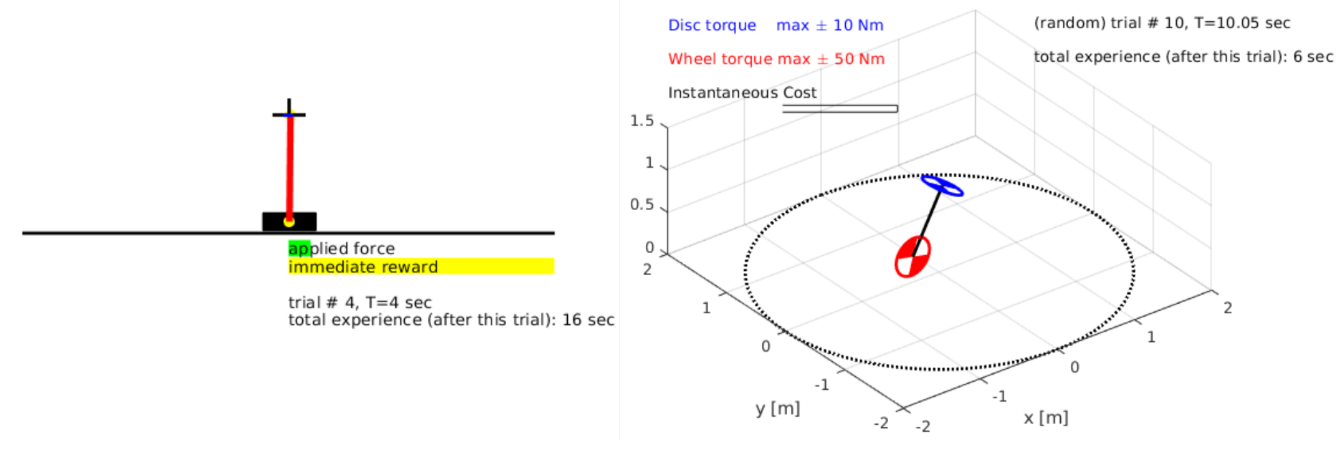

Fig. 1: Mechanical dynamical systems: cartpole (left) and unicycle (right) [2]

Model

Conditionally independent transition functions modelled for each target dimension.

- Gaussian prior on the weights: $\mathbf{w} \sim \mathcal{N}\left(\mathbf{0}, \sigma_{\mathbf{w}}^2 \mathrm{I}\right)$
- finite weight transition function: $f_{\mathrm{w}}(\mathbf{x}) = \boldsymbol{K}(\mathbf{x})^\top \mathbf{w}$
- random features provide an efficient and scalable kernel approximation [4]:

$$k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle \approx \mathbf{z}(\mathbf{x})^\top \mathbf{z}(\mathbf{x}'),$$

$$\mathbf{z}(\mathbf{x}) \equiv \sqrt{\frac{2}{D}} \left[\cos\left(\omega_1' \mathbf{x} + b_1\right) \cdots \cos\left(\omega_D' \mathbf{x} + b_D\right)\right]^\top$$

  for D iid offsets $b_1, \ldots, b_D \in \mathcal{R}$ from a uniform distribution on $[0, 2\pi]$

- trained using reparameterisation trick: $\boldsymbol{\omega} = \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and minimising the variational upper bound:

$$\mathbb{E}_{q(\mathbf{w})}[-\log p(\mathbf{y} | f_{\mathbf{w}}(\mathbf{x}))] + \mathbb{KL}(q(\mathbf{w}) \| p(\mathbf{w}))$$

Dealing with Uncertainty

Total loss uncertainty for policy π decomposed into the (i) aleatoric and (ii) epistemic components using the law of total variance [3]:

$$\mathbb{V}_{q(\mathbf{w})}\left(\mathcal{L}^\pi(\mathbf{x})\right) = \underbrace{\mathbb{E}_{q(\mathbf{w})}[\mathbb{V}(\mathcal{L}^\pi(\mathbf{x}))]}_{(i)} + \underbrace{\mathbb{V}_{q(\mathbf{w})}(\mathbb{E}[\mathcal{L}^\pi(\mathbf{x})])}_{(ii)}$$

Where $q(\mathbf{w})$ is the posterior distribution over the model weights and

$$\mathcal{L}^\pi(\mathbf{x}) = 1 - \exp\left(-\left\|\mathbf{x} - \mathbf{x}_{\text{target}}\right\|^2 / \sigma_c^2\right) \in [0, 1]$$

Preliminary Results

Figure 2 (right) shows the transition function for a single target dimension with 95% confidence interval (grey) and 3 functions sample (dotted) from the posterior distribution over the weights w.

Monte Carlo approximation to the distribution over trajectories (Figure 2 left) under policy π is generated by sampling $\mathbf{w} \sim q(\mathbf{w})$ a total of M times and then performing N roll-outs for each M with fixed $\mathbf{w}$ and start state s_0 sampled uniformly on the input space.

Trajectory samples used to calculate uncertainty decomposition and expected return.

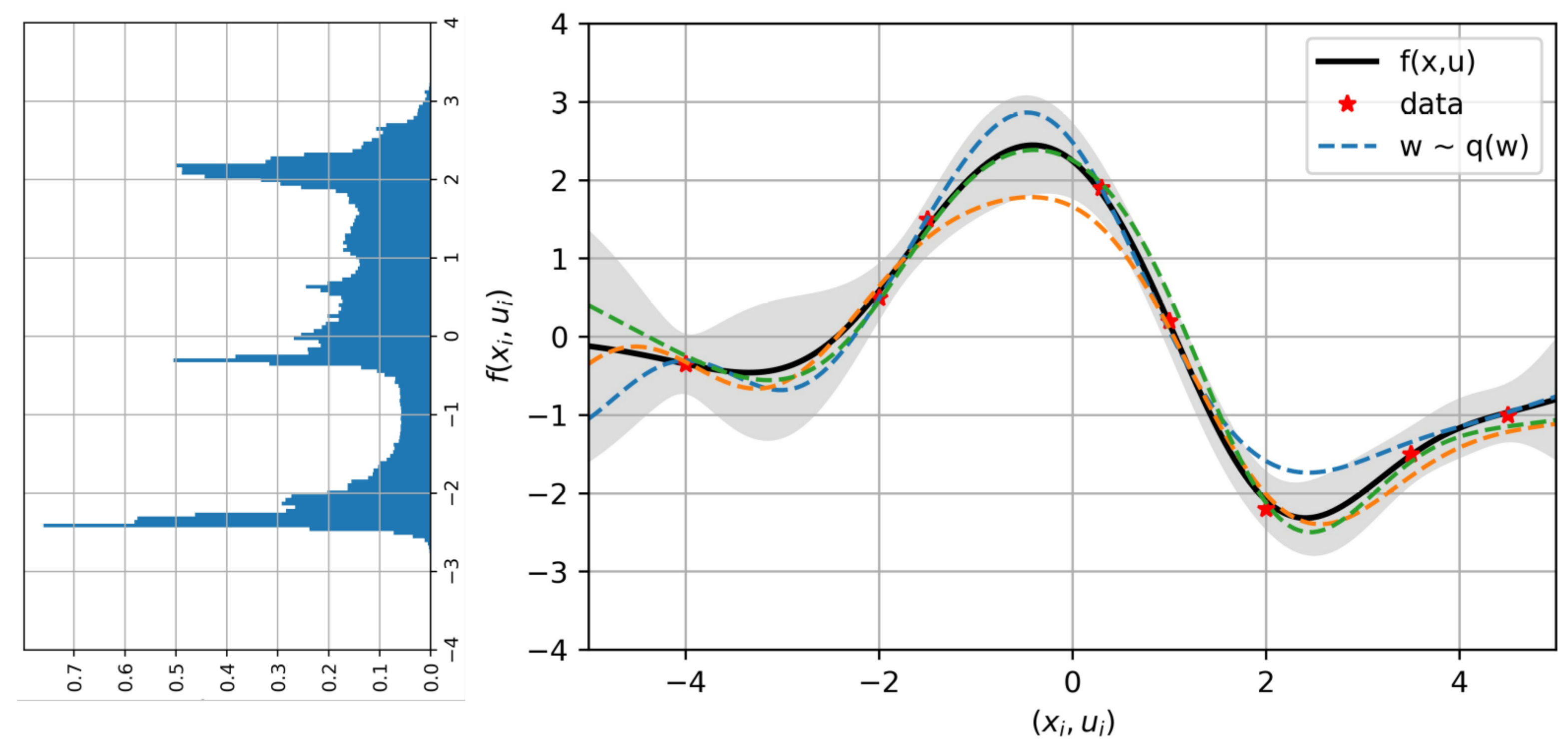

Fig. 2: Transition function (right) and MC approximation to the distribution over trajectories (left)

Ongoing Work

Work in progress includes:

- completion of PyLCO implementation
- uncertainty sensitive objective function to minimise the loss [1]:

$$\pi^* = \underset{\pi}{\operatorname{argmin}}\, \mathbb{E}_{q(\mathbf{w})}\left[\mathcal{L}^\pi\right] - \beta \sqrt{\mathbb{V}_{q(\mathbf{w})}\left(\mathcal{L}^\pi\right) - \mathbb{E}_{q(\mathbf{w})}\left[\mathbb{V}\left(\mathcal{L}^\pi\right)\right]}$$

- experiments and comparisons to current algorithm efficiency (see Figure 3)

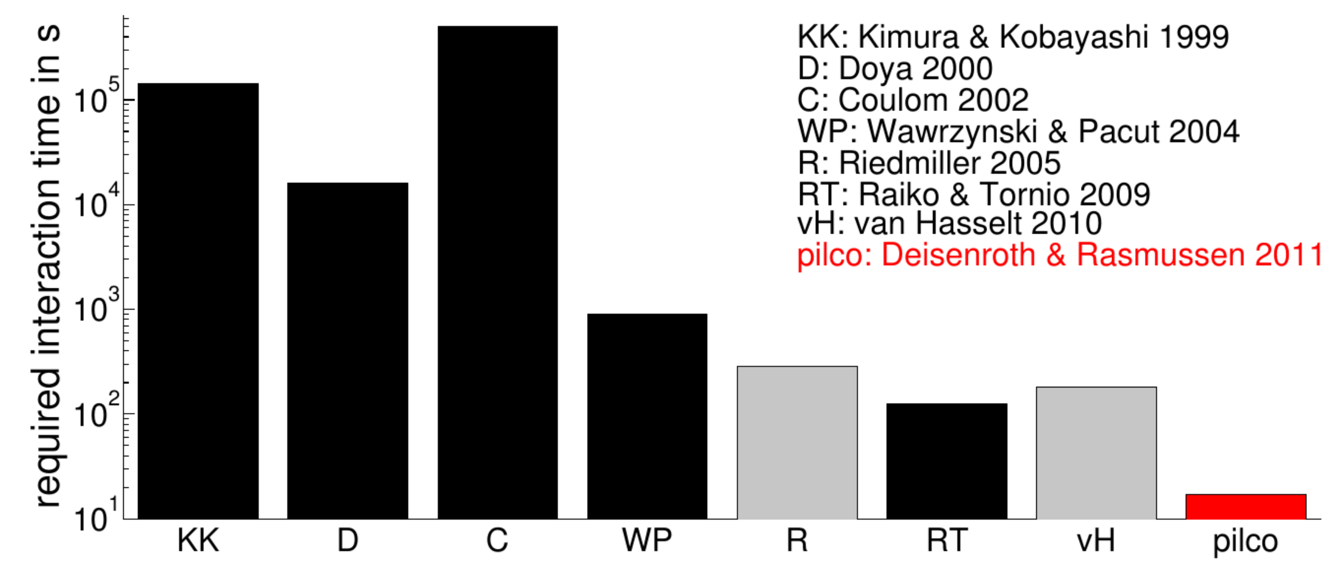

Fig. 3: Target algorithm efficiency [2].

References

[1] Peter Auer. "Using confidence bounds for exploitation-exploration trade-offs". In: *Journal of Machine Learning Research* 3.Nov (2002), pp. 397–422.

[2] Marc Deisenroth and Carl E Rasmussen. "PILCO: A model-based and data-efficient approach to policy search". In: *Proceedings of the 28th International Conference on machine learning (ICML-11)*. 2011, pp. 465–472.

[3] Stefan Depeweg et al. "Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning". In: *arXiv preprint arXiv:1710.07283* (2017).

[4] Ali Rahimi and Benjamin Recht. "Random features for large-scale kernel machines". In: *Advances in neural information processing systems*. 2008, pp. 1177–1184.