Learning and Decision Making in Dual Control Scenarios

Eric Schulz (eric.schulz.13@ucl.ac.uk), **Paula Parpart** (paula.parpart.10@ucl.ac.uk),
Neil R. Bramley (neil.bramley.10@ucl.ac.uk) & **Maarten Speekenbrink** (m.speekenbrink@ucl.ac.uk)

Abstract

In a dual control scenario participants have to control a non-episodic, finite-horizon dynamical system. This provides an informative paradigm to assess active learning and choice making behaviour in an exploration-exploitation environment, in which participants' actions influence outcomes, states, and where the reward function can change over time. We let participants control an environment called "Spaceship", in which they have to navigate a spaceship on a trajectory to avoid damage over time with changes of underlying reward functions. Results show that participants sensibly trade-off between exploration and exploitation behaviour in dependency of the current reward function and are overall best described by an approximate Bayesian dual control algorithm.

Keywords: Dual Control, Reinforcement Learning, Gaussian Process, Active Learning, Heuristics

Introduction

Many learning scenarios require us to trade-off *exploration* (learning about an underlying relationship) and *exploitation* (maximizing it). Within simple episodic settings in which the dynamics do not change, simple notions of trading off exploration and exploitation such as assigning an uncertainty based exploration bonus (Srinivas et al., 2009) or probability matching of expected outcomes (Agrawal & Goyal, 2011), also known as Thompson sampling, are efficient strategies that capture participants' behaviour well (Schulz et al., 2015b,a).

However, many real world scenarios are non-episodic, which means that our actions influence the system we are in and it is our task to control a single, ongoing trial. Within such scenarios, we might not return to known states and therefore have to treat exploration with great caution to avoid big disasters (Klenske & Hennig, 2015). These problems are normally approached via Bayesian reinforcement learning (Poupart, 2010), which assigns probabilistic beliefs over the dynamics of the environment and the occurring costs to reason about changes to beliefs from future observations, and their influence on future decisions (Duff, 2002). The attempt to combine the physical state with the parameters of the probabilistic model into an augmented dynamical model with the aim to control the resulting system is, however, computationally expensive (Hennig, 2011) and therefore its solution can only be approximated (Vlassis et al., 2012). The idea of augmenting the physical state with model parameters was coined early as *dual control* (Feldbaum, 1960) and has recently been adapted to approach inference in Bayesian reinforcement problems (Klenske & Hennig, 2015).

In what follows, we will build on Klenske & Hennig (2015) work and introduce different models for dual control and apply those to how humans learn in a simple control task. We find that participants are best described by a Gaussian Process approximate dual control algorithm. Given the assumption that participants active learning behaviour is indicative of their underlying cognitive processes (Parpart et al., 2015), this means tha

Problem set-up

Throughout all experiments we assume the following state dynamics

$$x_{k+1} = f_k(x_k, u_k) + \xi_k$$

paired with the following observation model

$$y_k = Cx_k + \gamma_k.$$

Furthermore, we assume a system with unknown dynamics that can, up to Gaussian noise, be described by a general linear model with nonlinear feature projections.

$$x_{k+1} = A_k\phi(x_k) + B_k u_k + \xi_k$$

Given a finite time horizon with terminal Time T and the following quadratic cost function.

$$\mathcal{L}(\mathbf{x}, \mathbf{u}) = \left[\sum_{k=0}^{T}(x_k - r_k)^\top W_k(x_k - r_k) + \sum_{k=0}^{T-1} u_k^\top U_k u_k\right]$$

where $\mathbf{r} = [r_0, \ldots, r_T]$ is the target trajectory and W_k and U_k are the potentially time-varying state and control costs. The goal is now at each time point k to find the action sequence $\mathbf{u}$ that minimizes the expected cost to the horizon T.

Models

Certainty equivalence

If target $r_k = 0$ and noise free observations, when a and b are known, optimal u_k to drive x_k to 0:

$$u_k^\star = \frac{abx_k}{U + b^2}$$

Certainty equivalence control simply replaces the parameter above with a mean estimate.

$$u_k^\star = \frac{a\mu_k x_k}{U + \mu_k^2}$$

This, however, can lead to bad estimations if the variance is high and overshooting reactions after regime changes.

Cautious control

Cautious control calculates expected cost $\mathbb{E}[x_{k+1}^2 + Uu_k^2]$ and then optimizes with respect to u_k.

$$u_k^\star = \frac{a\mu_k x_k}{U + \sigma_k^2 + \mu_k^2}$$

This control law scales down control actions in cases of high parameter uncertainty. However, this can also lead to one of the major drawbacks of the "cautious controller", which is that it decreases control with rising uncertainty, an effect commonly called "turn-off phenomenon".

Bayesian exploration bonus

Both of the aforementioned control algorithms are passive learners, that is they do not explore the parameter space actively and thereby might miss out on some important information that might become more important later on. A simple adaptation to introduce exploration into the certainty equivalent controller is to introduce a Bayesian exploration bonus, measure by the standard deviation at each observation point.

$$l_{BEB} = \tau \left[\sqrt{\mathrm{diag}(\Sigma^{\theta\theta})}\right]^\top \left[\sqrt{\mathrm{diag}(\Sigma^{\theta\theta})}\right]$$

Even though this model manages to trade-off between exploration and exploitation directly, it is still myopic in that it only calculates that step with one look ahead.

Approximate Dual Control

Approximate control is based on a simulation of the future that also tries to learn the underlying functions.

Experiment

Design

We are considering the example of a cart on a rail taken from (Klenske & Hennig, 2015). We will firstly describe the mathematical set up of that problem before then explaining how we transformed it into an actual experimental paradigm. The dynamics of the to be controlled environment are described by the following Equation:

$$x_{k+1} = \begin{bmatrix} 1 & 0.4 \\ 0 & 1 \end{bmatrix} x_k \begin{bmatrix} 0 & 0 \\ \theta^1 & \theta^2 \end{bmatrix} + \begin{bmatrix} \phi^1(x_k) \\ \phi^2(x_k) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u_k$$

The non-linear functions ϕ^1 and ϕ^2 are defined as shown below.

$$\phi^1(x) = -\frac{1}{1+\exp(x+5)}, \quad \phi^2(x) = \frac{1}{1+\exp(x+5)}$$

and the true underlying parameters

$$\theta_{true} = \begin{bmatrix} 0.8 \\ 0.4 \end{bmatrix}$$

This means that the true underlying functions and how they interact has to be learned while at the same time tracking the current reference point. The reference point is the current value that should be put out by the system at the next step. The to be tracked references are:

$$\mathbf{r}_{0:11} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \; \mathbf{r}_{12:14} = \begin{bmatrix} 10 \\ 0 \end{bmatrix} \; \mathbf{r}_{15} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\mathbf{r}_{16:18} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \; \mathbf{r}_{19:20} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

This means that the reference point changes over a total of 20 trials and the goal is always to create an output as close as possible to the current (and future) reference points.

The state weighting also changes over time as shown below.

$$\mathbf{W}_{0:5} = \begin{bmatrix} 10 & 0 \\ 0 & 0 \end{bmatrix} \; \mathbf{W}_{5:10} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \; \mathbf{W}_{11:20} = \begin{bmatrix} 100 & 0 \\ 0 & 0 \end{bmatrix}$$

This means that an initial stage of low state costs is followed by a stage of no costs at all which then leads to a stage of very high cost. Therefore, a well-adapted agent would first explore the space a little bit under low cost, then massively explore when there are no costs involved, just to control the outcome well during the last high cost stage.

Materials

Participants were told that they had to navigate a space ship through outer space.

Participants

N participants were recruited via Prolific Academic and received £1 as a basic reward as well as a performance-dependent reward of up to £1.

Results

Raw results of overall costs per participants are shown below.

It can be seen that participants learn over time and perform better than both the certainty equivalence or the cautious control model would predict. The denisty over sample points by round is shown below.
It seems to be clearly the case that participants use the phase of zero punishment to explore the function as much as they can, which speaks in favour of the Approximate Dual Control model.
All models were fitted to participant's choices by using a softmax function as shown below.

$$\sigma(z_{jt}) = \left\{ \frac{\exp(\theta z_{jt})}{\sum_{k=1}^{K} \exp(\theta z_{kt})} \right\}$$

The resulting optimise log-loss then was used to calculate Akaike's "An Information Criterion" for each model per participants. Results are shown in Table 1 below.

Table 1: AIC of different models.

Method	AIC_{mean}	AIC_{SD}	#best
Random			
Certainty equivalent			
Cautious control			
Exploration bonus			
Approximate Control			

Discussion and Conclusion

Scenarios in which we have to explore and exploit underlying functions are ubiquitous to every day life. In many cases, our current choices will influence where we will end up in the future and once visited states might not always be possible to return to. Dual control provides a good paradigm to assess participants' behaviour in such tasks. We showed that participants do efficient ahead planning in a simple control task (controlling a space ship) and were best described by a Gaussian Process approximate dual control algorithm. This can only be seen as a first step towards assessing human behaviour in such tasks and other, more psychologically plausible models, might explain participants' behaviour even better (Bramley et al.).

Acknowledgements

Work supported by the UK Centre of Financial Computing.

References

Agrawal, S., & Goyal, N. (2011). Analysis of thompson sampling for the multi-armed bandit problem. *arXiv preprint arXiv:1111.1797*.

Bramley, N. R., Dayan, P., & Lagnado, D. A. (????). Staying afloat on neuraths boat – heuristics for sequential causal learning.

Duff, M. O. (2002). *Optimal Learning: Computational procedures for Bayes-adaptive Markov decision processes*. Ph.D. thesis, University of Massachusetts Amherst.

Feldbaum, A. (1960). Dual control theory. i. *Avtomatika i Telemekhanika*, *21*(9), 1240–1249.

Hennig, P. (2011). Optimal reinforcement learning for gaussian systems. In *Advances in Neural Information Processing Systems*, (pp. 325–333).

Klenske, E. D., & Hennig, P. (2015). Dual control for approximate bayesian reinforcement learning. *arXiv preprint arXiv:1510.03591*.

Parpart, P., Schulz, E., Speekenbrink, M., & Love, B. C. (2015). Active learning as a means to distinguish among prominent decision strategies. In *Proceedings of the Thirty-Seventh Annual Conference of the Cognitive Science Society*.

Poupart, P. (2010). Bayesian reinforcement learning. In *Encyclopedia of Machine Learning*, (pp. 90–93). Springer.

Schulz, E., Konstantinidis, E., & Speekenbrink, M. (2015a). Exploration-exploitation in a contextual multi-armed bandit task.

Schulz, E., Konstantinidis, E., & Speekenbrink, M. (2015b). Learning and decisions in contextual multi-armed bandit tasks.

Srinivas, N., Krause, A., Kakade, S. M., & Seeger, M. (2009). Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*.

Vlassis, N., Ghavamzadeh, M., Mannor, S., & Poupart, P. (2012). Bayesian reinforcement learning. In *Reinforcement Learning*, (pp. 359–386). Springer.