Model-based Analysis of the Tower of London Task

Constantinos Mitsopoulos
Centre for Brain and Cognitive Development
Centre for Cognition, Computation and Modelling
Department of Psychological Sciences
Birkbeck, University of London, WC1E 7HX
c.mitsopoulos@bbk.ac.uk

Denis Mareschal
Centre for Brain and Cognitive Development
Centre for Cognition, Computation and Modelling
Department of Psychological Sciences
Birkbeck, University of London, WC1E 7HX
d.mareschal@bbk.ac.uk

Richard Cooper
Centre for Cognition, Computation and Modelling
Department of Psychological Sciences
Birkbeck, University of London, WC1E 7HX
r.cooper@bbk.ac.uk

Abstract

The planning process is central to goal-directed behaviour in any task that requires the organization of a series of actions aimed at achieving a goal. Although the planning process has been investigated thoroughly, relatively little is known about how this process emerges and evolves during childhood. In this paper we describe three reinforcement learning models of planning, in the Tower of London (ToL) task, and use Bayesian analysis to fit each model to pre-existing data from 3-4 year-old and 5-6 year-old children performing the task. The models all capture the increased organisation seen in the older children's performance. It is also shown that, at least for this dataset, the most complex model – that with discounting of future rewards and pruning of highly aversive states – provides no additional explanatory power beyond a simpler discounting-only model. Insights into developmental aspects of the planning process are discussed.

Keywords: Reinforcement Learning, Shaping Rewards, Planning, Tower of London

Acknowledgements

We are deeply indebted to Peter Dayan for all the discussions that inspired this work. This research was supported by the European Commission grant MC-ITN-289404-ACT.

1 Introduction

Environmental stimuli combined with external rewards or punishments may elicit certain responses, which ultimately lead to learned behaviours. In this context, extrinsic motivation, which means to be moved to do something because of a specific reward outcome, may be distinguished from intrinsic motivation, which means to be moved to do something because it is inherently enjoyable [1]. Intrinsic motivation is evident in animal behaviour, where it has been found that organisms engage in exploratory, playful and curiosity-driven behaviour even in the absence of an environmental reinforcement [4]. Similarly, researchers in many areas in cognitive science emphasize the importance of intrinsically motivated behaviour in human development and learning.

Our concern in this paper is whether instrinsic motivation might play a role in the cognitive processes underlying planning. Human planning has been studied extensively using look-ahead puzzles, in which subjects have to preplan mentally a sequence of moves in order to transform a starting configuration of the puzzle to a goal configuration, subject to a set of rules. In the Tower of London (TOL; [7]) task, for example, subjects are required to rearrange three balls on three pegs so that the configuration of balls matches a goal state (see figure 1), but in doing so they must adhere to a set of rules or constraints. Thus they must move only one ball at a time, and place it back on a peg before moving another ball. The task can be viewed as a sequential decision making puzzle, with reward obtained if / when the player achieves the goal state. Here, however, we use computational methods to explore the effects of more frequent feedback – reflecting intrinsic motivation – on appropriate moves that may guide the subject towards solving the puzzle. Specifically, we model an existing dataset from children's planning on the ToL by incorporating a *reward shaping function*, representing the intrinsic motivation of the child, within the framework of model-based reinforcement learning.

Figure 1: A typical Tower of London problem. The task consists of a board with three pegs, each one with different heights, and three different coloured balls. The right peg can contain up to three balls, the middle peg up to two balls and the left one only one ball. The balls are initialy arranged in one configuration on the pegs and the goal is to move balls – one at a time and from peg to peg – in order to achieve the given goal configuration. The problem shown requires 3 moves, but more difficult problems may require up to 7 moves.

2 Modelling the ToL task

In problems such as the ToL, the goal is achieved by decomposing it into subgoals and evaluating the order of simple moves towards the goal [3]. It is this evaluation procedure that guides our approach to planning in such a task. We model children's behaviour on the ToL as a Markov Decision Process (MDP) and follow model-based approaches.

2.1 The Extended State Space of ToL

Within the empirical study on which this work is based (described in more detail below), some children (especially the younger ones) failed to adhere to the task rules. That is, although it was explained to each child that he/she should only move one ball at a time using only one hand, and although the child in each case claimed to understand this restriction and demonstrated this knowledge in a series of practice trials, sometimes he/she would pick up one ball in one hand in order to reorder the position of the other two balls that would otherwise require a series of moves. This typically occurred when the state of the apparatus almost matched the goal state, with two balls on one peg being in the wrong order (e.g., red immediately above blue when blue should have been immediately above red). One possibility in this case is that the child's look-ahead process suggests to him/her that there is great similarity between the current and goal states, yet any (legal) move would result in a decrease in similarity. From the perspective of search through a decision tree, pruning of the tree might take place when facing such situations, leaving the child with only one viable option – to move both balls at the same time and hence break the task rules.

In order to accommodate breaking the task rules by subjects, we expand both the state space (from 36 states to 114 states) by adding two more locations representing the hands of the child (effectively two additional pegs, each of which can hold at most one ball), and the set of available actions (adding actions corresponding to moving balls to and from the hands). This yields an extended state transition matrix $T : S \times A$ with $[114 \times 25]$ entries.

2.2 The Reward Function

The design of the transition matrix is straightforward, as the task is deterministic, but for the reward function further assumptions are necessary. In the Tower of London task, the reward from the environment is given to the subject only at the goal state. In addition to this, however, we assume that subjects are driven step-by-step towards the goal state by an internal reward function, which is

related to the similarity of the current configuration of the task, state s_t at time t, to the desired configuration (i.e., the goal state). By "similarity" we mean the degree of overlap, in terms of positions of the balls at the pegs, between two states (as defined in the following paragraph). In other words, in the planning process we assume that subjects evaluate their future actions in terms of not just whether they achieve the goal state, but (for other states) how close they bring them to the goal state. Previous work has shown that such a modification to the reward structure often suffices to render straightforward otherwise intractable learning problems. Additionally, proper modification can leave the optimal strategy invariant (see [6]).

To calculate state similarity, as required by the internal reward function, we represent each state within the ToL by a set of 24 binary features (bits). For each ball we assign three bits to represent its vertical position on the peg and five bits for the peg that the ball is placed on (three for the real pegs and two representing the hands). According to this scheme if the red ball is at the lowest position on the first peg then it will be represented as $R_{pos} = (1,0,0)$ and $R_{peg} = (1,0,0,0,0)$. The state vector is the concatenation of the vectors for each ball: $\mathbf{s}_t = (R_{pos}, R_{peg}, G_{pos}, G_{peg}, B_{pos}, B_{peg})$. We then define the *similarity between two states* as the inner product between those states. The reward shaping function therefore has the form $F(s,s') = \phi(\mathbf{s}') - \phi(\mathbf{s})$ where $\phi(\mathbf{x}) = \mathbf{x}^T \mathbf{x}_{goal}$ is the inner product between the state representations under comparison (with bold letters denoting the state vector of features).

It seems that some children perceive some configurations ("towers" where all three balls are on the longest peg, or "flats" where all three balls are on different pegs) as being the same, independent of the arrangement of colour. The above approach helps us capture similarity in the structure of a particular configuration (i.e., the number of balls on each of the pegs is the same for both configurations independent of colour).

3 Methods

3.1 Model-based analysis

In this section we describe three model-based RL models used to describe the planning process. The models, the model fitting and model comparison procedures are described in detail in [5] but we repeat the description here for completeness. All three models assume that subjects choose actions stochastically, with the probability of choosing action (or choice) c from state s at time t given by:

$$p(c_t|s_t) = \frac{e^{\beta Q(s_t,c_t)}}{\sum_{c'} e^{\beta Q(s_t,c')}} \tag{1}$$

The parameter β is an inverse temperature that represents the subject's sensitivity to rewards. The three models differ in the calculation of the function $Q(s_t,c_t)$.

The first model is the *Lookahead* model. This model is simply a tree search model in the sense that searches all available options until the end of the tree:

$$Q_{lo}(s,c) = R(s,c) + \max_{c'} Q_{lo}(s',c') \tag{2}$$

where s' is the successor state from state s by selecting choice c. In all the models we set $R(s,c) = (1-w)R_{ext} + wR_{int}$. This means that a low w will indicate goal directed behaviour whereas high w indicates planning driven by state similarity.

For the extended ToL with 25 available actions at each state[1], and a decision tree of depth $D = 3$, the total number of action choices considered by the lookahead model is 16275. This number is large and children are unlikely to evaluate this number of actions during planning. One possibility is that they prune the decision tree and evaluate action trajectories according to their expected outcome. This leads to the second model, the *Discount* model.

We assume that at each depth of the decision tree, a biased coin is flipped in order to determine whether the tree search should be terminated and return zero reward. Let the probability of stopping be γ , and using a mean-field approximation (in order not to use the immense number of possible choices at the branches of the decision tree), the Q values are estimated by:

$$Q_d(s,c) = R(s,c) + (1-\gamma)\max_{c'} Q_d(s',c') \tag{3}$$

Then the future outcome, k steps ahead, is weighted by the probability $(1-\gamma)^{k-1}$ that it is encountered.

The third model we used is a modification of the Discount model, and we refer to this as the *Pruning* model. We test the hypothesis that subjects avoid states with great dissimilarity with the goal state. Thus we modify the calculation of Q from the Discount model to the following:

$$Q_{pr}(s,c) = R(s,c) + (1-x)\max_{c'} Q_{pr}(s',c') \tag{4}$$

where

$$x = \begin{cases} \gamma_S & \text{if } R_{int}(s,c) \text{ is a large dissimilarity} \\ \gamma_G & \text{else} \end{cases} \tag{5}$$

[1]The actual available choices, at each state are given by counting all possible transitions of the balls at the pegs, including the two extra pegs which represent the hands of the child.

γ_S (Specific pruning parameter) is the probability that the subject stops evaluating the decision tree at a state where the immediate reward leads to a state with great dissimilarity with the goal state. γ_G (General pruning parameter) is γ as in the Discount model.

3.2 Model fitting procedure

We assume an hierarchical model on how data are generated for each age group. Each model is characterized by a set of parameters $\mathbf{k}_i$, for each subject i, that are generated by a Gaussian distribution $\mathbf{k}_i \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{v}^2)$ with parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}, \mathbf{v}^2)$. We will refer to these as *hyperparameters*. The whole analysis is applied separately for each of the two age groups. We fit the model parameters and the hyperparameters in a joint scheme, using the EM algorithm [2], maximising the marginal likelihood given all data by all N subjects:

$$\hat{\boldsymbol{\theta}}^{ML} = \arg\max_{\boldsymbol{\theta}} P(\mathcal{C}|\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \left(\prod_i^N \int P(\mathbf{c}_i|\mathbf{k}_i) P(\mathbf{k}_i|\boldsymbol{\theta}) d^N \mathbf{k}_i \right) \tag{6}$$

where $\mathcal{C} = \{\mathbf{c}_i\}_{i=1}^N$ is the set of all actions performed by each subject i. Actions are assumed to be independent. Thus they factorize over trials. According to the above, for the E-step at the j^{th} iteration we use the Laplace approximation to approximate the integral of the marginal (eq. 6) and the parameters are estimated at the maximum of the posterior distribution (MAP). For the M-step we estimate the hyperparameters $\boldsymbol{\theta} = (\boldsymbol{\mu}, \mathbf{v}^2)$, by maximising the expectation computed at the E-step. For the Lookahead model we fitted 1 parameter, for the Discount model 2 parameters, and for the Pruning model 3 parameters. All parameters were transformed before inference to enforce constraints ($\beta \geq 0, 0 \leq \gamma_S, \gamma_G \leq 1$). The above procedures were verified by using surrogated data from a known decision process.

3.3 Model comparison

Given the three models, and given that the models have different number of parameters, it is important to compare the models to understand which best accounts for the children's behaviour. Having no prior knowledge about each model, we assume that models are equally likely a priori. Thus we examine only the log likelihood of each model $\log P(\mathcal{C}|\mathcal{M})$. This quantity can be approximated by the Bayesian Information Criterion (BIC) as:

$$\log P(\mathcal{C}|\mathcal{M}) = \int P(\mathcal{C}|\boldsymbol{\theta}) P(\boldsymbol{\theta}|\mathcal{M}) d\boldsymbol{\theta} \approx -\frac{1}{2}\mathrm{BIC}_{int} = \log P(\mathcal{C}|\hat{\boldsymbol{\theta}}^{ML}) - \frac{1}{2}|M| \log |\mathcal{C}| \tag{7}$$

where $|\mathcal{C}|$ is the total number of choices made by all subjects of the group being examined, and $|\mathcal{M}|$ is the number of prior parameters (mean and variance for each parameter) that we estimated empirically above. The first term on the right hand side of equation 7 was estimated by standard Monte Carlo approximation. The Bayesian Information Criterion (BIC_{int}), apart from penalizing the model for extra parameters, is not the sum of individual likelihoods, but the sum of *integrals* over individual parameters thus the *int* (integral) subscript. With this approach we compare not only how well a model fits the data when its parameters are optimized, but also how well a model fits the data when we use information about where the group level parameters lie on average [5].

Although the above gives a good comparative measure of model fit, an absolute measure is needed in order to ensure that the best model does indeed describe the data generation procedure efficiently. Given the MAP estimation of each subject's parameters, we compute the mean total "predictive probability" for subjects N, in a number of trials T, which is the geometric mean of all the $P(c_t|s_t, \mathbf{k}_i)$ where c_t is the action selected at trial t by the i^{th} subject, at the state s_t from a decision process parametrized by $\mathbf{k}_i$.

4 Experimental results and discussion

To evaluate the three planning models we consider existing data from seventeen 3-4 year olds (mean age 47 months) and seventeen 5-6 (mean age 68 months) years old on six ToL problems of graded difficulty [8]. The younger children in this study struggled to complete many of the problem, and in both groups some children failed to complete all problems. Therefore in the analysis below we restrict attention to data from 10 of the younger children and 13 of the older children. Given the population and the number of problems we obtained 60 and 78 action sequences for younger children and older children respectively. Among younger children, 7 out of 10 performed illegal moves (37 total), whereas 5 out of 13 of the older children used illegal moves (22 total). As illegal moves we count the transition of a ball from a peg to the hand and not the other way around. The original dataset and experimental conditions are described in detail in [8] and summarised in Table 1.

The phenomenon of the direction of the behaviour towards a perceptual match with the goal state (i.e., reaching a state with the same configuration of the balls as the goal state, except the colour of the balls is different in the goal state, and declaring that the goal state reached) is much evident in younger children. The inferred parameter w was 0.28 and 0.52 (Pruning Model estimation) for older and younger children respectively, revealing a significant difference between the planning mechanisms between the two groups. This suggests that younger children pursue a similarity match between goal state and their current state whereas the older children follow more goal directed behaviour. However, by comparing BIC_{int} scores (e.g., 5-6yrs old group: Lookahead (1455), Discount (1105), Pruning (1115)) and mean predictive probabilities (e.g., 5-6yrs old group: Lookahead (0.85), Discount (0.89), Pruning (0.89)), we found that the Discount and Pruning models describe children's behaviour better than the Lookahead model,

Table 1: The percentage of children showing different behaviours in each age group

Sequences	3-4yrs	5-6yrs
Reached goal correctly	38.3%	64.1%
Reached goal with illegal moves	40.0%	24.3%
Perception matching	20.0%	5.1%
Interrupted/Stopped	1.7%	6.4%

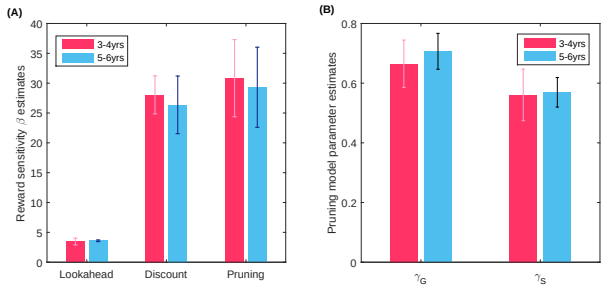

Figure 2: Model parameters estimates: (A) Reward sensitivity β estimates from the three models. (B) Estimations of General pruning γ_G and Specific pruning γ_S parameters for the two age groups given the pruning model.

though the extra parameter of the Pruning model does not improve the model predictions beyond that of the Discount model (at least in the specific ToL problems tested). This may reflect a lack of sophistication in planning ability at these ages. Further investigation of behaviour on specific ToL problems could reveal the importance of various features that affects their planning process.

An analysis of choice behaviour according to our models shows that older children prune, in general, more than the younger children. However the difference is very small. This early termination of the decision tree for the younger ones, appears to be mainly because are driven by the (perceived) similarity of the current state to the goal state, leading to them "cheating" by holding two balls at the same time In addition, younger children tend to overprune their decision tree and confuse the objective similarity between states. On the other hand older children demonstrated a better level of planning (i.e., reaching the goal state by following the rules consistently). They tend to prune but in a way that leads them to the goal state without shortcuts (i.e., without picking up more than one ball at a time). Furthermore, the older children tended not to show confusion in distinguishing very similar states. Finally looking at the reward sensitivity parameter β (Fig 2A), younger children are much more greedy to rewards than older children pursuing a perceptual match between current state and goal state.

We have demonstrated a method for analysing human behaviour in puzzle tasks where the main reward factor is the internal reward, represented by a shaping reward function. By testing it in a real world example as the above, useful insights can be gained concerning differences in mental planning between age groups, though further work needs to be conducted to formally explore the relationship between internal reward representations and planning in child development.

References

[1] E L Deci and R M Ryan. *Intrinsic Motivation and Self-Determination in Human Behavior*. Springer Science & Business Media, 1985.

[2] A P Dempster, N M Laird, and D B Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[3] K J Gilhooly, L H Phillips, V Wynn, R H Logie, and S Della Sala. Planning processes and age in the five-disc tower of london task. *Thinking & reasoning*, 5(4):339–361, 1999.

[4] H F Harlow. Learning and satiation of response in intrinsically motivated complex puzzle performance by monkeys. *Journal of comparative and physiological psychology*, 43(4):289–294, 1950.

[5] Q J M Huys, N Eshel, E O'Nions, L Sheridan, P Dayan, and J P Roiser. Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS computational biology*, 8(3):e1002410, 2012.

[6] A Y Ng, D Harada, and S J Russell. Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping. *International Conference on Machine Learning*, 16:278–287, 1999.

[7] T Shallice. Specific Impairments of Planning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 298(1089):199–209, 1982.

[8] R Waldau. A developmental study of problem solving in children aged 3–6: Development of planning strategies, 1999. Undergraduate disertation.