Helping people choose subgoals with sparse pseudo rewards

Frederick Callaway
Dept. of Psychology, UC Berkeley
fredcallaway@berkeley.edu

Falk Lieder
Dept. of Psychology, UC Berkeley
falk.lieder@berkeley.edu

Thomas L. Griffiths
Dept. of Psychology, UC Berkeley
tom_griffiths@berkeley.edu

Abstract

Many decisions require planning multiple steps into the future, but optimal planning is computationally intractable. One way people cope with this problem is by setting subgoals, suggesting that we can help people make better decisions by helping them identify good subgoals. Here, we evaluate the benefits and perils of highlighting potential subgoals with pseudo-rewards. We first show that sparse pseudo-rewards based on the value function of a Markov decision process (MDP) lead a limited-depth planner to follow the optimal policy in that MDP. We then demonstrate the effectiveness of these pseudo-rewards in an online experiment. Each of 88 participants solved 40 sequential decision-making problems. In control trials, participants only saw the state-transition diagram and the reward structure. In experimental trials, participants additionally saw pseudo-rewards equal to the value (sum of future rewards) for the states 1-, 2-, or 3-steps ahead of the current state. When participants reached one of those states, the display would again reveal the values of the states located 1-, 2-, or 3-steps ahead of the current state. We found that showing participants the value of proximal states induced goal-directed planning and improved their average score per second. This benefit was largest when the incentives were 1 or 2 steps away and decreased as they were moved farther into the future. Although these pseudo-rewards were beneficial overall, they also caused systematic errors: Participants sometimes neglected the costs and rewards along the paths to potential subgoals, leading them to make "unwarranted sacrifices" in the pursuit of the most valuable highlighted states. Overall, our results suggest that highlighting valuable future states with pseudo-rewards can help people make better decisions. More research is needed to understand what constitutes optimal subgoals and how to better assist people in selecting them.

Keywords: pseudo-rewards; shaping; planning; goals; gamification

Acknowledgements

This work was supported by grant number ONR MURI N00014-13-1-0341 and a grant from the Templeton World Charity Foundation.

1 Introduction

Many important decisions require sacrificing immediate reward in the pursuit of greater payoffs in the future. Incorporating potential future rewards requires solving a Markov decision process. The computational challenge of solving this problem may partly explain why people tend to underappreciate future rewards relative to immediate rewards (Myerson & Green, 1995).

Inspired by the use of *shaping rewards* in reinforcement learning (Ng, Harada, & Russell, 1999), Lieder and Griffiths (2016) proposed that this problem can be ameliorated by providing people with *pseudo-rewards* that align short term reward with long term value. They showed that optimal pseudo-rewards (i.e., the value of the next state minus the value of the current state) enable myopic agents (i.e., agents maximizing immediate rewards) to follow the optimal policy and improve human performance in sequential decision problems.

In the real world, however, it is generally infeasible to provide pseudo-rewards after every action a person takes. Furthermore, computing optimal pseudo-rewards exactly is intractable for the decision problems that people face in real-life. Instead, a decision support system may only have access to the approximate values of a limited number of states. Fortunately, although people may not be able to look far into the future, research on human planning has shown that they are not entirely myopic either (Morris & Ward, 2004). One strategy people use to approximate optimal planning is to set subgoals (Newell, Simon, et al., 1972), an approach that has also been applied in artificial intelligence (Shivashankar, Kuter, Nau, & Alford, 2012).

However, to use subgoals effectively one must choose good subgoals, and humans sometimes rely on simple heuristics and habits when setting subgoals (Cushman & Morris, 2015). Perhaps then, humans and decision support systems have complementary stregnths and weaknesses. Humans can often achieve subgoals, but they may not always select the best subgoals to pursue. Decision support systems, on the other hand, cannot feasibly advise on every decision, but they may have information about the values of certain states. This suggests that a decision support system can improve human decision making by helping people choose better subgoals using its knowledge of valuable states.

We propose that a decision support system can inform peoples' choice of subgoals by providing *sparse pseudo-rewards* to convey the value of potential subgoal states that are within reach. People could then select one of these states as their target by combining their knowledge of proximal rewards with the state values conveyed by pseudo-rewards. Here, we investigate the potential benefits and perils of this approach through mathematical analysis and behavioral experiments.

2 Optimal sparse pseudo-rewards for non-myopic agents

Lieder and Griffiths (2016) proved that pseudo-rewards generated by a potential function of the optimal value function lead a *myopic* agent (one that greedily maximizes immediate reward) to follow the globally optimal policy, taking $r'(s_t, a_t, s_{t+1}) = r(s_t, a_t, s_{t+1}) + f(s_t, a_t, s_{t+1})$ where $f(s_t, a_t, s_{t+1}) = V^\star(s_{t+1})\gamma - V^\star(s_t)$. Here, we extend this result, showing that sparse pseudo-rewards can provide the same benefit for an agent with some ability to plan ahead. Let an N-step planner be the policy

$$\pi^N(s_t) = \arg\max_{a_t} \max_{a_{t+1}\dots a_{t+n-1}} \sum_{i=0}^{N-1} r'(s_{t+i}, a_{t+i}, s_{t+i+1})\gamma^i \tag{1}$$

Lieder and Griffiths (2016) note that an optimal planning agent will still perform optimally when given their pseudorewards because of the shaping theorem (Ng et al., 1999). However, because sparse pseudo-rewards are necessarily not potential-based, the shaping theorem cannot apply. Thus, we begin by proving directly the optimality of dense pseudo-rewards in order to build intuition for the sparse case.

Lemma 2.1. *The N-step planner π^N is equivalent to the optimal policy π^* when the pseudo-reward function is that of Lieder and Griffiths (2016).*

Proof. We begin by breaking up the modified reward into its pieces.

$$\sum_{i=0}^{N-1} r'(s_{t+i}, a_{t+i}, s_{t+i+1})\gamma^i = \sum_{i=0}^{N-1} r(s_{t+i}, a_{t+i}, s_{t+i+1})\gamma^i + \sum_{i=0}^{N-1} f(s_{t+i}, a_{t+i}, s_{t+i+1})\gamma^i \tag{2}$$

Expanding the second term, we find that the fs telescope.

$$\sum_{i=0}^{N-1} f(s_{t+i}, a_{t+i}, s_{t+i+1})\gamma^i = -V^\star(s_t) + V^\star(s_{t+1})\gamma^1 \cdots - V^\star(s_{t+N-1})\gamma^{N-1} + V^\star(s_{t+N})\gamma^N = V^\star(s_{t+N})\gamma^N - V^\star(s_t) \tag{3}$$

Thus the N-step planner chooses the first action in a series of actions that maximizes the sum of rewards plus the difference in value between the final state and the current state. The value of the current state is constant under the maximization, so we can remove it.

$$\pi^N(s_t) = \arg\max_{a_t} \max_{a_{t+1}\dots a_{t+n-1}} \sum_{i=0}^{N-1} r(s_{t+i}, a_{t+i}, s_{t+i+1})\gamma^i + V^\star(s_{t+N})\gamma^N \tag{4}$$

Starting from the optimal policy and working backwards, we expand the recursive definition of $V^\star$ $N-1$ times to attain

$$\pi^\star(s_t) = \arg\max_{a_t} r(s_t, a_t, s_{t+1}) + \dots + \max_{a_{t+N-1}} r(s_{t+N-1}, a_{t+N-1}, s_{t+N})\gamma^{N-1} + V^\star(s_{t+N})\gamma^N = \pi^N(s_t) \tag{5}$$

■

Note that this proof can be easily extended to stochastic MDPs by replacing the functions f, r and $V^\star$ with the expected values of the corresponding distributions.

Given that it may not be feasible to provide pseudo-rewards after every action, we would like would like to construct an f that has this optimality property but is also sparse (i.e. $f(s_t, a_t, s_{t+1}) = 0$ for most t). Unfortunately, it is likely impossible to construct such an f without making assumptions about the agent and environment. As shown in Equation 4, the key contribution of the pseudo-rewards is in the term $V^\star(s_{t+N})\gamma^N$, which comes from the pseudo-reward $f(s_{t+N-1}, a_{t+N-1}, s_{t+N})$. That is, the proof relies critically on the pseudo-reward at the final transition of each path considered while planning. In general, this could be any transition in the MDP, thus we cannot safely set f to 0 for any transition, precluding sparsity in the general case.

Designing optimal sparse-pseudo rewards may be impossible without knowing what paths the agent could consider. However, in some cases, this information may be available. Assume we know the current state of the agent s_t as well as its planning depth N. This information defines a set S_{t+N} of all possible states the agent could be in after N steps. We can then define pseudo-rewards relative to the current time step as

$$f_t(s, a, s') = \begin{cases} V^\star(s') & \text{if } s' \in S_{t+N} \\ 0 & \text{otherwise} \end{cases}$$

Lemma 2.2. *The N-step planner π^N is equivalent to the optimal policy π^* when the pseudo-reward function is f_t and the environment in non-cyclic.*

Proof. Substituting f_t into Equation 3, we find $\sum_{i=0}^{N-1} f_t(s_{t+i}, a_{t+i}, s_{t+i+1})\gamma^i = V^\star(s_{t+N})\gamma^N$ which differs from the original case only in the constant that is removed in Equation 4. Thus, we can follow the same steps starting at Equation 4 to show that π^N with shaping function f_t is equivalent to π^*. Note that the first step of this proof requires that f_t be 0 for all states the agent visits except s_{t+N}. This assumption holds in non-cyclic environments. ■

This formulation is less than ideal because it only applies to non-cyclic environments. Furthermore, the agent will never actually attain a pseudoreward because f_t moves a little further every step. Both problems can be solved by assuming that the N-step planner never plans past the nearest pseudo-reward (an assertion that we make without proof for lack of space). In this case we can reset f_t only when the agent attains the pseudo-reward. We follow this assumption in designing our experimental stimuli.

3 Testing the effects of sparse pseudo-rewards on planning

To investigate the effects of pseudo-rewards on planning, we employ a modified version of the paradigm of Lieder and Griffiths (2016) which presents a series of sequential decision problems with versus without pseudo-rewards. We hypothesized that spacing out the pseudo-rewards such that they occurred every N time steps would lead participants to form subgoals to reach states with high pseudo-rewards. This simplifies the planning problem, allowing participants to plan only as far as the next pseudo-reward, and thus making their planning process to be less error-prone and less time-consuming. We expected this benefit to be proportional to the reduction in planning distance. Thus, we predicted that participants would achieve higher scores and spend less time on trials with more frequent pseudo-rewards.

3.1 Methods

We recruited 88 participants on Amazon's Mechanical Turk using the psiTurk experimental framework (Gureckis et al., 2016). Participants were paid \$0.75 plus a performance-dependent bonus of up to \$1.85 for completing 40 trials

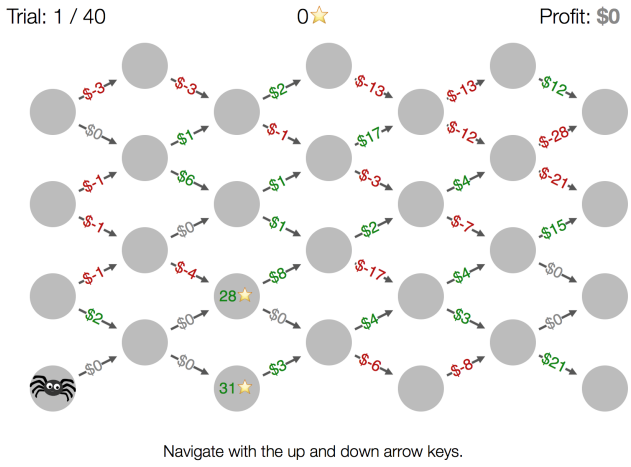

Figure 1: Screenshot of the experiment: Participants control the spider with the arrow keys to maximize profit. The number of stars conveys the value of a state.

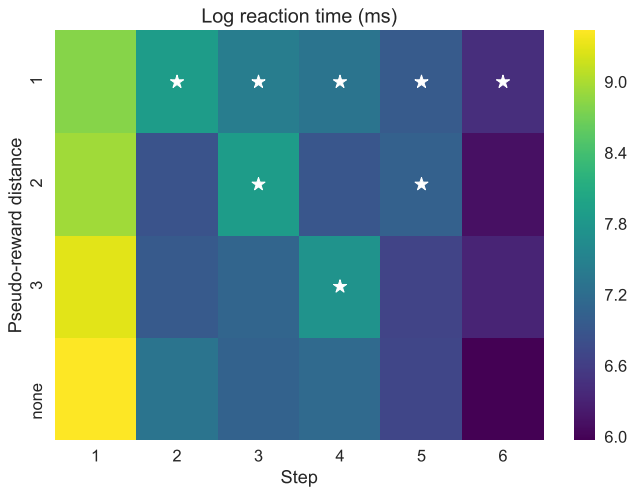

Figure 2: Planning time is higher on states with pseudo-rewards, indicated by stars.

(roughly 14 minutes of work). The experiment was programmed with the JsPsych library (De Leeuw, 2015) using the Mouselab-MDP plugin (Callaway, Lieder, Krueger, & Griffiths, in press). Trials were counterbalanced such that each reward structure occurred with each pseudo-reward distance an equal number of times across all participants.

A screenshot of the task is shown in Figure 1. Participants navigated a spider through the *Web of Cash*, aiming to attain the highest possible profit. (Their bonus was 5% of their profit on a random round). The graph was structured such that participants could only move rightwards, and only one row up or down on each step. This induces a long-term planning problem in which one may need to make a short-term suboptimal decision to reach a highly valuable state. To increase the importance of planning, rewards were drawn from a zero-centered normal distribution with variance increasing with depth, Normal$(0, 2 + 2d)$, and clipped to be integers. We selected stimuli that discriminated 1-step, 2-step, and 6-step planners by rejection sampling.

On some trials, pseudo-rewards, displayed as stars, were presented every 1, 2, or 3 steps. Only the pseudo-rewards at the nearest such column were visible; when one of those pseudo-rewards was attained, the next round appeared. Participants were explicitly informed that the number of stars on a state was the maximum total profit one could gain starting in that state. In MDP jargon, the number of stars on a state was equal to the value of that state. Thus a participant could take the optimal path by considering all paths up to the next column of pseudo-rewards and taking the path with the largest sum of rewards (dollars) and pseudo-rewards (stars).

3.2 Results

Planning time Before conducting our primary analyses, we sought evidence that the pseudo-rewards induced goal-targeted planning. We defined an intuitive behavioral signature of planning based on reaction times: If a participant chooses to plan ahead at a given state, her reaction time will be higher than average; however, once a plan has been formed, she can quickly execute the full sequence. Thus, if the pseudo-rewards induced subgoal-targeted planning, we would expect to see longer reaction times on states with pseudo-rewards because participants would need to re-plan after reaching the pseudo-reward induced subgoal.

Log reaction times broken down by step and pseudo-reward frequency are shown in Figure 2. We excluded the first action (the first column in Figure 2) from our analysis because one would expect unusually high reaction time at the first state regardless of the presence of pseudo-rewards. We constructed a mixed effects linear regression model over log reaction time at each step with one fixed effect for the presence of a pseudo-reward on that step and random effects for participant, stimulus, trial index, and step number. An Anova comparing the full model to the model with only random effects revealed a significant effect of pseudo-rewards on planning time ($\beta = 0.708$, $\chi^2(1) = 619.37$, $p < 0.001$), indicating that pseudo-rewards induced goal-targeted planning.

Score rate We predicted that pseudo-rewards would either increase the quality of actions, reduce the time spent planning, or both. To capture both of these possibilities in one metric, we defined *score rate* to be the average score per second. Results are plotted in Figure 3. For each measure (score, time, score rate), we constructed a mixed effects linear regression model with one fixed effect for distance between pseudo-rewards (or 6, the number of steps, when pseudo-rewards were

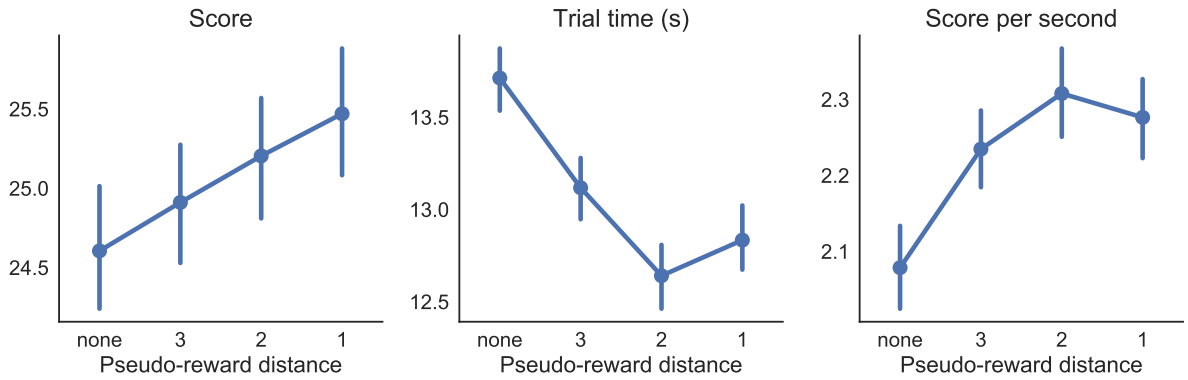

Figure 3: Average score, trial time, and score rate (score per second) for different frequencies of pseudo-rewards. Error bars show 68% confidence intervals by bootstrapping.

absent) and random effects for participant, stimulus, and trial index. The effect of distance between pseudo-rewards on score rate was statistically significant ($\beta = -0.045$, $\chi^2(1) = 15.14$, $p < 0.001$). The effect on score ($\beta = -0.166$, $\chi^2(1) = 4.90$, $p < 0.05$) and trial time were also significant ($\beta = 0.204$, $\chi^2(1) = 44.74$, $p < 0.001$).

Unwarranted sacrifices So far we have seen that sparse subgoal-directed planning induced by pseudo-rewards can have positive effects. However, there may be negative side effects. Increasing the salience of valuable states may lead participants to neglect the cost of reaching the state, or the many small rewards along a path to a less valuable state. We term a suboptimal plan made in tenacious pursuit of a goal an *unwarranted sacrifice*. This effect can be quantified in our paradigm as traveling to the state with maximal value at a given time step—perhaps highlighted by the largest pseudo-reward—when that state does not lie on the optimal path. Let $s_t^V = \arg\max_{s \in S_t} V^*(s)$ be the state with maximal value at step t. Similarly, let s_t^* be the state at step t of the optimal path. We can define a state s_t as the result (and perhaps also the cause) of an unwarranted sacrifice if $s_t = s_t^V \wedge s_t \neq s_t^*$. A χ^2 test revealed a significant effect of pseudo-rewards on unwarranted sacrifices ($\chi^2(1) = 12.52$, $p < 0.001$;), supporting the hypothesis that pseudo-rewards increase the likelihood that people will pursue a valuable state when doing so is suboptimal.

4 Conclusion

We found that pseudo-reward induced subgoals enabled people to solve sequential decision problems more efficiently, suggesting that pseudo-rewards could be used in real-world settings to help people make better decisions. However, our results also suggest that this approach is not without perils: When choosing subgoals, people may neglect the cost of achieving them. In future research, we will examine the extent to which pseudo-rewards exacerbate this effect, and what can be done to ameliorate it.

References

Callaway, F., Lieder, F., Krueger, P. M., & Griffiths, T. L. (in press). Mouselab-MDP: A new paradigm for tracing how people plan. In *The 3rd multidisciplinary conference on reinforcement learning and decision making*.

Cushman, F., & Morris, A. (2015). Habitual control of goal selection in humans. *Proceedings of the National Academy of Sciences*, *112*(45), 13817-13822. doi: 10.1073/pnas.1506367112

De Leeuw, J. R. (2015). jsPsych: A javascript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1–12.

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., … Chan, P. (2016). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior research methods*, *48*(3), 829–842.

Lieder, F., & Griffiths, T. L. (2016). Helping people make better decisions using optimal gamification. In *Proceedings of the 38th annual meeting of the cognitive science society* (pp. 2075–80). Austin, TX: Cognitive Science Society.

Morris, R., & Ward, G. (2004). *The cognitive psychology of planning*. Psychology Press.

Myerson, J., & Green, L. (1995). Discounting of delayed rewards: Models of individual choice. *Journal of the experimental analysis of behavior*, *64*(3), 263–276.

Newell, A., Simon, H. A., et al. (1972). *Human problem solving* (Vol. 104) (No. 9). Prentice-Hall Englewood Cliffs, NJ.

Ng, A. Y., Harada, D., & Russell, S. (1999). Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml* (Vol. 99, pp. 278–287).

Shivashankar, V., Kuter, U., Nau, D., & Alford, R. (2012). A hierarchical goal-based formalism and algorithm for single-agent planning. In *Proceedings of the 11th international conference on autonomous agents and multiagent systems-volume 2* (pp. 981–988).