Generalization in planning by metalearning to metareason

Frederick Callaway
Princeton University
Princeton, NJ, USA
fredcallaway@princeton.edu

Bas van Opheusden
Princeton University
Princeton, NJ, USA
svo@princeton.edu

Erin Grant
UC Berkeley
Berkeley, CA, USA
eringrant@berkeley.edu

Thomas L. Griffiths
Princeton University
Princeton, NJ, USA
tomg@princeton.edu

Introduction

Recent successes of machines in sequential decision-making problems such as Atari and Go have involved exponential increases in domain-specific experience and computational power. By contrast, humans are able to rapidly and accurately identify solutions to novel sequential decision-making problems—all with a biological computer that uses less power than a lightbulb. Two critical components of human intelligence that allow us to think and learn so efficiently are metareasoning and metalearning (Griffiths et al. 2019), that is, *thinking about how to think* and *learning how to learn*. These two ideas have primarily been considered separately, but we propose that the ability to efficiently form plans in variable and novel contexts can be achieved by applying metalearning to metareasoning problems.

Planning as metareasoning

Given an accurate model of the environment and infinite computational resources, one can always identify an optimal policy by planning—in principle. In practice, however, constraints on computational resources make exact planning impossible. Planning algorithms should thus be evaluated not only by their theoretical ability to find an optimal solution, but also by their empirical performance given some fixed computational budget. This suggests that planning is fundamentally a problem of computational resource allocation. The theory of rational metareasoning frames this as a sequential decision problem in which an agent attempts to maximize the sum of the rewards received from executing external actions minus the cost of the computations used to select those actions (Horvitz, Cooper, and Heckerman 1989; Russell and Wefald 1991). This idea is formalized in a *metalevel* Markov decision process (Hay et al. 2012).

A metalevel Markov decision process (MDP) $M_{\text{meta}} = (\mathcal{B}, \mathcal{C}, T_{\text{meta}}, r_{\text{meta}})$ is an MDP where the states, $\mathcal{B}$, encode an agent's beliefs over task-relevant variables, the actions, $\mathcal{C}$, are computations, the transition function, T_{meta}, describes how computations update beliefs, and the reward function, r_{meta}, describes the costs and benefits of computation. In particular, r_{meta} is strictly negative for all computations except a special operation, $\perp$, which terminates the decision making process and executes the external action (or sequence of actions) that is optimal given the final belief state, b_T. The final metalevel reward, $r_{\text{meta}}(b_T, \perp)$, is the expected utility of that action. In this framework, a planning algorithm is formalized as a metalevel Markov policy, π_{meta}, which selects the next computation to perform based on the current belief state.

To make things concrete, we consider a process-tracing task that has been used to study human planning (Callaway et al. 2018b). In the Mouselab-MDP task (illustrated in Figure 1A), participants navigate a spider through a tree-structured MDP in order to maximize total reward. The reward gained at each state is initially occluded, but may be revealed by clicking. This task is naturally modeled as a metalevel MDP. A belief state specifies a joint distribution over the reward function at each state, the distribution at each state either being the prior or a delta distribution on the observed value. A computation reveals the reward at a state, except for $\perp$ which executes a sequence of actions with maximal expected return given the current belief state. The transition function specifies the probability that each possible reward is revealed after a click. Finally, the metalevel reward function penalizes each click with a constant negative reward and rewards the $\perp$ action with the maximal expected return of any path.

A critical challenge for metareasoning is that the benefit of selecting computations intelligently must not be outweighed by the cost of deciding which computations to execute. Research in psychology suggests that humans amortize metalevel decision cost through model-free RL (Jain et al. 2019). However, despite early proposals to apply such a technique in machines (Harada and Russell 1999), this idea is yet to be fully explored. In a recent instantiation of this approach, Callaway et al. (2018a) present a direct policy search method to learning a metalevel planning policy. However, because the metalevel policy is optimized separately for each environment, the upfront cost of learning may still outweigh the benefits of future computational thrift.

Metalearning to metareason

For a metalevel learning algorithm to be useful, it must be able to quickly and flexibly adapt to new domains. This is exactly the problem considered by metalearning (Schmidhuber, Zhao, and Wiering 1996; Thrun and Pratt 2012). In

general, the target of metalearning is a learning algorithm that can achieve high performance with minimal experience on a new task drawn from some family of related tasks. Here, we focus on metalearning applied to RL, in which a "task" corresponds to an MDP and "minimal experience" corresponds to a small number of episodes interacting with that MDP (Wang et al. 2016; Duan et al. 2017). While meta-RL increases the sample efficiency of learning to solve a new MDP (Rakelly et al. 2019; Zintgraf et al. 2019; Humplik et al. 2019), it does not explicitly reason about the computational cost of online planning, and therefore does not by design maximize computational efficiency. Metalearning to metareason addresses this deficiency by allowing a metalevel policy to transfer experience across metalevel MDPs when deliberating about the sequence of computations to perform.

As a proof of concept, we applied metalearning to metareason to the Mouselab-MDP planning task. We defined a distribution of metalevel MDPs, p_M, in which several dimensions were allowed to vary: the height and width of the decision tree, the cost of revealing a reward, and the probabilities of each reward in the set $\{-1, 0, 1\}$. As a learning algorithm, we used Bayesian Q-learning with linear function approximation and Monte Carlo estimates of return. That is, we assume $Q(b, c; \boldsymbol{w}) = \boldsymbol{w}^\top \phi(b, c)$ and $G_t \sim \text{Normal}(Q(b_t, c_t; \boldsymbol{w}), \sigma^2)$, where G_t denotes the cumulative reward from timestep t to the end of the episode and the features, ϕ, are taken from Callaway et al. (2018a). We specified priors $p_0(\boldsymbol{w}) \sim \text{Normal}(\boldsymbol{\mu}, \text{diag}(\boldsymbol{\lambda}^{-1}))$ and $p_0(\sigma) \sim \text{Gamma}(a, b)$, updating these distributions at the end of each episode based on all previous experience. The policy uses Thompson sampling over weights for exploration, sampling $\tilde{\boldsymbol{w}}_t \sim p_t(\boldsymbol{w})$ and then executing the computation $c_t = \arg\max_c Q(b_t, c; \tilde{\boldsymbol{w}}_t)$.

The prior over the Bayesian regression parameters is a natural target for metalearning because it describes not only the starting point in weight space, but also the exploration policy and the rate at which each weight is updated from experience. As a metalearning objective, we chose the expected total reward achieved over the course of 20 episodes in an MDP sampled from p_M. We approximate this objective by Monte Carlo (sampling 100 environments) and optimize it with Bayesian optimization.

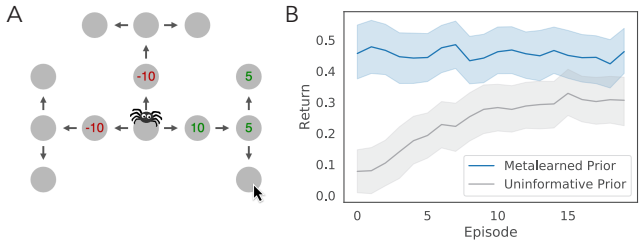

Figure 1: Proof of concept. (A) The Mouselab-MDP paradigm. Rewards are revealed by clicking, prior to selecting a path. (B) Average performance of the metalearning vs simple learning agent over a distribution of Mouselab-MDP planning problems. Bands show 95% CI.

We found that the metalearned prior yielded dramatically improved performance when compared with an uninformative baseline prior. However, as shown in Figure 1B, this improvement is expressed as strong immediate generalization to new domains, apparently making fine-tuning unnecessary. This result may be explained by our use of hand-engineered features which naturally generalize within this relatively constrained domain. Thus, in future work we will parameterize components of the learner with neural networks, allowing metalearning of metareasoning policies across environments with more complex state and action spaces.

References

Callaway, F.; Gul, S.; Krueger, P.; Griffiths, T. L.; and Lieder, F. 2018a. Learning to select computations. In *Uncertainty in Artificial Intelligence: Proceedings of the Thirty-Fourth Conference*.

Callaway, F.; Lieder, F.; Das, P.; Gul, S.; Krueger, P. M.; and Griffiths, T. L. 2018b. A resource-rational analysis of human planning. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.

Duan, Y.; Schulman, J.; Chen, X.; Bartlett, P. L.; Sutskever, I.; Abbeel, P.; and Science, C. 2017. RL2: Fast reinforcement learning via slow reinforcement learning. 1–14.

Griffiths, T. L.; Callaway, F.; Chang, M. B.; Grant, E.; Krueger, P. M.; and Lieder, F. 2019. Doing more with less: meta-reasoning and meta-learning in humans and machines. *Current Opinion in Behavioral Sciences* 29:24–30.

Harada, D., and Russell, S. 1999. Learning search strategies. In *In Proc. AAAI Spring Symposium on Search Techniques for Problem Solving under Uncertainty and Incomplete Information*. Citeseer.

Hay, N.; Russell, S. J.; Tolpin, D.; and Shimony, S. 2012. Selecting computations: Theory and applications. In *Proceedings of the 28th Conference of Uncertainty in Artificial Intelligence*.

Horvitz, E. J.; Cooper, G. F.; and Heckerman, D. E. 1989. Reflection and action under scarce resources: Theoretical principles and empirical study. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, 1121–1127. San Mateo, CA: Morgan Kaufmann.

Humplik, J.; Galashov, A.; Hasenclever, L.; Ortega, P. A.; Teh, Y. W.; and Heess, N. 2019. Meta reinforcement learning as task inference. *arXiv preprint arXiv:1905.06424*.

Jain, Y. R.; Gupta, S.; Rakesh, V.; Dayan, P.; Callaway, F.; and Lieder, F. 2019. How do people learn how to plan?

Rakelly, K.; Zhou, A.; Quillen, D.; Finn, C.; and Levine, S. 2019. Efficient off-policy meta-reinforcement learning via probabilistic context variables. *arXiv preprint arXiv:1903.08254*.

Russell, S. J., and Wefald, E. 1991. Principles of metareasoning. *Artificial Intelligence* 49(1-3):361–395.

Schmidhuber, J. H.; Zhao, J.; and Wiering, M. 1996. Simple principles of metalearning. *Technical report IDSIA-69-96* 1–23.

Thrun, S., and Pratt, L. 2012. *Learning to learn*. Springer Science & Business Media.

Wang, J. X.; Kurth-Nelson, Z.; Tirumala, D.; Soyer, H.; Leibo, J. Z.; Munos, R.; Blundell, C.; Kumaran, D.; and Botvinick, M. 2016. Learning to reinforcement learn. 1–17.

Zintgraf, L.; Shiarlis, K.; Igl, M.; Schulze, S.; Gal, Y.; Hofmann, K.; and Whiteson, S. 2019. Varibad: A very good method for bayes-adaptive deep rl via meta-learning. *arXiv preprint arXiv:1910.08348*.