Goal-Specific Strategies for the Explore-Exploit Dilemma

Brian Montambault
(brian.montambault@tufts.edu)
Department of Computer Science, Tufts University
Christopher Lucas
School of Informatics, University of Edinburgh

Abstract

When an agent is presented with a set of options, each with an uncertain reward, their behavior depends on how they balance the conflicting goals of exploiting known rewards and exploring unknown rewards that might uncover information that will be valuable in the future. While heuristic approaches tend to do well across a range of tasks, it is unclear whether human learners are using these same approaches or employing a rational, goal-specific strategy. We compare the performance of human learners on both a reinforcement learning and optimization task to that of rational, goal-based and heuristic, general approaches.

Introduction

explore-exploit dilemma

Given a set of possible actions, each with an uncertain outcome, a learning agent's performance over time depends on their ability to gather information about each outcome. However, information gathering is often in direct conflict with seeking short-term reward. This is known as the explore-exploit dilemma, and how human learners resolve it remains an open question. A number of heuristics have been proposed that directly address this trade-off, but fail to take into account the goal of the agent. Our main contribution is to draw the distinction between different goals, their associated loss functions, and how they might be solved by a Bayesian rational agent. In particular, we contrast the goal of maximizing cumulative reward typically considered in reinforcement learning contexts (e.g. Sutton & Barto, 1998) with that of finding the action that maximizes reward considered in optimization (e.g. Snoek et al., 2012).

contextual multi-armed bandits

We will consider these goals in the contextual multi-armed bandit (CMAB) paradigm, where k possible actions, $a_1,...,a_k$, where a_i yield the random rewards $r_1,...,r_k$ drawn from an unknown distribution. Actions are associated with a set of shared features and the function $R: x \rightarrow r$ mapping features to rewards. The policy $\tilde{\pi}$ maps the observed history of actions and rewards $h_{1:t-1} = [(a_1,x_1,r_1),...,(a_{t-1},x_{t-1},r_{t-1})]$ to the next action a_t for all possible histories $h_{1:t-1} \in \mathcal{H}$. When there is uncertainty about R, the explore-exploit dilemma exists as a balance between choosing actions with high expected reward and those that will yield relevant information about the function.

probabilistic model of function learning

Each distinct goal that might be supported in this paradigm first requires the mapping from features to rewards to be learned. As this step happens prior to determining the best action at the next time step, we assume that the underlying mechanism is identical across all possible goals. While a number of models of human function learning exist, we adopt Gaussian process regression (GPR) as proposed by Griffiths et al. (2009). GPR solves the problem of learning to map inputs to outputs by assuming that the outputs $\mathbf{y}_N$ are drawn from the N dimensional distribution $\mathcal{N}(m(\mathbf{x}_N), k(\mathbf{x}_N,\mathbf{x}_N))$, where m defines a mean function that broadly captures the function in the absence of data, and k defines how the inputs relate to each other. For the unobserved input x', the output y' is drawn from a normal distribution with the posterior mean and variance functions

$$m(x'|\mathbf{x}_N,\mathbf{y}_N) = m(x') + \mathbf{k}(x')^T(\mathbf{K}+\sigma^2\mathbf{I})(\mathbf{y}_N - m(\mathbf{x}_N))$$
$$v(x'|\mathbf{x}_N,\mathbf{y}_N) = k(x',x') - \mathbf{k}(x')^T(\mathbf{K}+\sigma^2\mathbf{I})^{-1}\mathbf{k}(x')$$

where $\mathbf{k}(x') = [k(x_1,x'),...,k(x_N,x')]$ and $\mathbf{K} = k(\mathbf{x}_N,\mathbf{x}_N)$. The main advantage of this approach in the context of CMAB problems is that the uncertainty about the reward associated with each action is captured by the posterior variance of R, making it clear which actions should be explored to gain the most information given a particular goal. Bayesian optimization (Snoek et al., 2012) provides a framework in which actions are evaluated in terms of *acquisition functions* depending on the posterior means and variances of an underlying Gaussian process.

explore-exploit heuristics

A number of acquisition functions have been proposed that attempt to optimally balance exploration and exploitation. Rather than explicitly working to maximize utility given a particular goal, these functions assign utility based strictly on local measures about each action. In general, these functions assign high utility to actions where the marginal mean and variance (equation 1) of its associated reward are both high. the *upper confidence bound* (Krause & Ong, 2011) algorithm defines this trade-off explicitly:

$$\text{acq}_{UCB}(a) = m(a|h_{1:t-1}) + \beta\sqrt{v(a|h_{1:t-1})}$$

with β as a free parameter determining the agent's preference for exploration. *probability of improvement* (Kushner, 1963) defines utility as the probability that each action will result in a reward that is higher than the current highest expected value:

$$\text{acq}_{PoI}(a) = \Phi\left(\frac{m(a|h_{1:t-1}) - m(a^*|h_{1:t-1})}{\sqrt{v(a|h_{1:t-1})}}\right)$$

where a^* is the action with the highest expected reward and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal. A similar heuristic is the *expected improvement*:

$$\text{acq}_{EI}(a) = (m(a|h_{1:t-1}) - m(a^*|h_{1:t-1}))\Phi(z) + \sqrt{v(a|h_{1:t-1})}\phi(z)$$

where $z = \frac{m(a|h_{1:t-1}) - m(a^*|h_{1:t-1})}{\sqrt{v(a|h_{1:t-1})}}$ and $\phi(\cdot)$ is the probability density function of the standard normal. While all three of these functions have the advantage of a closed analytical form, the use of local measures does not allow for one to capture uncertainty about global attributes of the reward function such as its maximum. As such, these heuristics are unable to take into account the goal of an agent, instead opting for an inexpensive but general solution to the balancing exploration and exploitation. In the next section, we will describe strategies that directly solve the strategy of the agent.

goals and optimal solutions to CMAB tasks

While the heuristics described above provide general strategies for explore-exploit problems, a rational agent should act in such a way that minimizes loss with respect to a specific goal (Anderson, 1991). Often encountered in reinforcement learning, a common goal is to maximize cumulative reward over the course of the task. Loss can thus be defined in terms of *cumulative regret*:

$$\mathcal{L}_{CR}(h_{1:T}) = \sum_{t=1}^{T} r^* - r_t$$

or the sum of the differences between the best possible reward r^* and the actual rewards observed over T trials. The optimal solution to a CMAB task given is given by the solution to the Bellman equation

$$\mathbf{E}_{CR}[r_{1:T}|\tilde{\pi}(h_{1:t-1})] = \mathbf{E}_{CR}[r_t|\tilde{\pi}(h_{1:t-1})] + \mathbf{E}_{CR}[r_{t+1:T}|\tilde{\pi}(h_{1:t})]$$

where the first term yields the expected reward on trial t and the second term recursively defines the expected reward on all subsequent trials up to the last trial T. This second term can be evaluated for trial $t+1$ by weighing the expected reward given all possible histories $\mathcal{H}_{1:t} = [(a_1, x_1, r_1), ..., (a_{t-1}, x_1, r_{t-1}), (a_t, x_1, r_t)]$ for $r_t \in \mathbb{R}$ by the probability of that history occurring:

$$\mathbf{E}_{CR}[r_{t+1:T}|\tilde{\pi}(h_{1:t})] = \int_{\mathbb{R}} p(r_t|a_t, h_{1:t-1})\mathbf{E}_{CR}[r_{1:T}|\tilde{\pi}(h_{1:t})]$$

This solution is intractable, so an appropriate approximation to long-term reward must be constructed. Short-term loss, that is, loss after any single trial t can be easily evaluated:

$$\mathcal{L}_{ST}(h_t) = r^* - r_t.$$

Since the action taken on trial $t+1$ depends on information gained about the maximum reward r^* on trial trial t, long-term loss can be described by:

$$\mathcal{L}_{LT}(h_{1:t}) = \int_{\mathbb{R}} [\min_{r_{t+1}} \mathcal{L}_{ST}(h_{t+1})] P_{r^*}(r_{t+1}|h_{1:t}).$$

and long-term loss for trials $t+1$ to T given a particular given the set of rewards observed up to trial t is thus $[T-t]\mathcal{L}(h_{1:t})$. Assuming the measures of short and long-term loss are proportional up to some constant, we can define total expected loss on trial t as

$$\mathbf{E}[\mathcal{L}_{CR}(h_{1:t})] = \mathcal{L}_{ST}(h_t) + \beta[T-t]\mathcal{L}_{LT}(h_{1:t})$$

where β is a scaling constant. Following Wang & Jegelka (2017), we define utility given $\mathcal{L}(h_{1:t})$ in terms of reduction of entropy about r^*, yielding the maximum entropy search acquisition function:

$$\text{acq}_{MES}(a) = H(P_{r^*}(r|h_{1:t-1}, x)) - \mathbf{E}[H(P_{r^*}(r|h_{1:t-1}, x, r^*))]$$

and the associated acquisition function:

$$\text{acq}_{RL}(a) = m(a|h_{1:t-1}) + \beta[T-t]\text{acq}_{MES}(a).$$

We note that while this appears similar to the UCB acquisition function described earlier, it uses a global measure of uncertainty rather than a local one, and is thus better suited for describing an agent acting rationally with respect to its goals.

We also consider the goal of optimization, where loss is defined in terms of *simple regret*:

$$\mathcal{L}_{SR}(h_{1:T}) = r^* - \max(r_{1:T})$$

The optimal solution given the goal of optimization is similar to that of reinforcement learning

$$\mathbf{E}_{SR}[r_{1:T}|\tilde{\pi}(h_{1:t-1})] = \max(\mathbf{E}_{SR}[r_t|\tilde{\pi}(h_{1:t-1})] + \mathbf{E}_{SR}[r_{t+1:T}|\tilde{\pi}(h_{1:t})])$$

with the exception being that only the maximum reward obtained is counted. Since only the best reward that is obtained over all trials is counted towards simple regret only long-term loss needs to be considered for the first $T-1$ trials, while on the last trial only short-term loss should be considered:

$$\mathbf{E}[\mathcal{L}_{SR}(h_{1:t})] = \begin{cases} \mathcal{L}_{ST}(h_t), & \text{if } t = T \\ \mathcal{L}_{LT}(h_{1:t}), & \text{otherwise.} \end{cases}$$

This also gives us the corresponding acquisition function

$$\text{acq}_{opt}(a) = \begin{cases} m(a|h_{1:t-1}), & \text{if } t = T \\ \text{acq}_{MES}(a), & \text{otherwise.} \end{cases}$$

It is unclear whether human learners act rationally with respect to their goal, or employ cheaper, more general heuristics across all goals. We approach this question by testing the performance of human learners against the performance of the heuristic and rational acquisition functions in reinforcement learning and optimization CMAB tasks.

Experiments

In each of the following experiments, we consider a set of tasks where each relies on a similar mapping from features to rewards. In each task participants are shown a set of uniformly spaced vertical bars on their computer screen. At the beginning of a task, each bar appears 500 pixels tall and gray in color. For each task there is an unknown function mapping each bar's order from left to right $(1,...,N)$ to a reward (between 0 and 500). Participants are invited to click on any of the bars over a number of trials. When a gray bar is clicked its color changes to black and its height is adjusted to match it's corresponding reward (between 0 and 500 pixels). After each trial the reward associated with the chosen bar is used to update the participant's goal-specific reward, which is displayed on the screen alongside the bars. On each trial, any bars that were clicked on previous trials remain black and the height in pixels of their associated rewards.

In the RL task goal-specific reward is determined by cumulative reward of all previous actions. In the optimization task goal-specific reward is determined by the height of the maximum reward that has been gained on previous trials. In the AL task, no goal-specific reward is displayed during the initial task. Instead, a secondary task is completed to determine. On each trial of the secondary task one of the remaining gray bars from the previous tasks is highlighted in red. Participants are then asked to indicate the correct height of the highlighted bar according to the underlying reward function. Participants indicated the correct height by clicking anywhere inside the highlighted bar. Goal-specific reward was increased by the maximum height (500) minus the mean absolute error of the participants judgment (0 to 500) on each trial.

For each of the following experiments the height of each bar was determined by one of three possible reward functions:

$$R_{linear}(x) = x$$
$$R_{quadratic}(x) = -(x-55)^2$$
$$R_{sinc}(x) = \frac{\sin(x/2-30.000001)}{x/2-30.000001}$$

Experiment 1A

On tasks like those described above, where each has a distinct goal but all can share similar reward functions, optimal behavior is determined by the goal-specific reward that is unique to each task. However, given the added cognitive burden of determining and following a unique strategy for each new task, it is possible that human learners will instead employ a general heuristic that performs reasonably well on all tasks. The goal of this experiment was to test whether human learners act rationally with respect to their goal in a set of similar tasks, or use a general strategy for all tasks.

Participants

0 participants were recruited using Amazon Mechanical Turk and received \$0.0 in addition to a bonus based on their final score.

Procedure

After giving their informed consent, participants received instructions outlining either the RL, AL, or optimization task. 80 vertical bars were displayed on the screen, with the hidden height of each bar determined by either the linear, quadratic, or sinc function. Participants were then invited to click one of the 80 on each of 25 trials. After the chosen bar was clicked the height of the chosen bar was revealed according to the reward function, and the next trial immediately began. For those in the AL condition a second task began immediately following the last trial of the first task, and consisted of 10 additional trials.

Results

Behavioral Results.

Model Comparisons. We considered a number of strategies for these tasks, each of which required that the reward function be modeled using a Gaussian process. Specifically, we considered a mixture of Gaussian processes (Tresp, 2001). This mixture consisted of four kernel functions that allow for a wide range of functions that participants might have expected to encounter in this experiment: short/long length-scale squared exponential (equation 3, $l = 1$ / $l = 5$), linear (equation 4, $d = 1$), and quadratic (equation 4, $d = 2$). The mixture weight on each kernel was set to .25.

Two classes of models were considered. The first class consisted of strategies that maximize goal-specific reward. For RL and optimization-specific strategies, we chose to approximate long-term reward as the reduction in entropy about the maximum of the reward function (equation 9). For the RL task, we consider the acquisition function

$$\text{acq}_{RL}(a) = m(a|h_{1:t-1}) + (T-(t-1))\beta \cdot \text{acq}_{MES}(a) \quad (1)$$

where β determines the relative preference for long-term over short-term reward and $T-(t-1)$ is the number of trials remaining as of trial t. On the last trial, the second term will be zero and the optimal action will be determined only by the expected short-term reward. For the optimization task, we consider the acquisition function

$$\text{acq}_{opt}(a) = \begin{cases} m(a|h_{1:t-1}), & \text{if } t = T \\ \text{acq}_{MES}(a), & \text{otherwise} \end{cases} \quad (2)$$

for which the optimal action on each trial up to the last is the one that most reduces uncertainty about the maximum possible reward, and the optimal action on the last trial is the one with with the greatest expected value. For the AL task, we consider the acquisition function

$$\text{acq}_{RL}(a) = v(a|h_{1:t-1}) \quad (3)$$

for which the optimal action is the one whose associated reward is most uncertain. The second class of strategies consist of those that act to maximize some local approximation of short and long-term reward, including the GP UCB, expected improvement, probability of improvement, and ε-greedy strategies.

Discussion

Experiment 1B

On each task, the optimal strategy for selecting the next action that will yield the maximum total goal-specific reward can be determined evaluating the expected reward of each action. However, this type of strategy requires that an explicit model of the underlying reward function be updated on each trial. An alternative, cheaper strategy might be to estimate the optimal next action using some model-free method. The goal of this experiment was to test whether human learners act rationally with respect to observed evidence of the underlying reward function. In particular, we investigate whether providing learners with additional information about the underlying reward function before the task begins improves performance.

Participants

0 participants were recruited using Amazon Mechanical Turk and received $0.0 in addition to a bonus based on their final score.

Procedure

After giving their informed consent, participants received instructions outlining either the RL, AL, or optimization task. As in experiment 1A, 80 vertical bars were displayed on the screen. However, instead of the height of all bars being hidden, 10 randomly selected bars were revealed before the first trial began. Aside from this the same procedure was followed as in experiment 1A.

Results

Behavioral Results.

Model Comparisons. We considered the GP strategies that were considered in experiment 1A in addition to a model-free strategy based on gradient descent.

Discussion

Experiment 2

For a sequential set of tasks with sharing a common reward function, optimal goal-specific reward can be earned by sacrificing performance on earlier tasks by spending more trials exploring the reward function in order to exploit this knowledge during later tasks. The goal of this experiment was to test whether human learners act rationally given shared information in a set of distinct tasks. In particular, we hoped to determine whether human learners use strategies that value improving shared knowledge or act only to maximize reward on the current task.

Procedure

Participants were asked to complete each of the three (RL, optimization, AL) tasks one after the other in random order. If the height of a bar was revealed during any particular task it would remain visible for all subsequent tasks. For each individual task the procedure was identical to that in experiment 1A.

Results

Behavioral Results.

Model Comparisons. For this task we considered each of the strategies discussed in the results of experiment 1A. We also considered this set of models augmented to put additional value on learning more about the reward function across each task. For these models we placed a Dirichlet ([1., 1., 1., 1.]) prior over kernel mixture weights. Long-term reward gained after from reducing uncertainty about the mixture parameters is approximated by the weighted variance $\hat{\sigma}_w^2 = \sum_{i=1}^{K} w_i (m_i(a|h_{1:t-1}) - \mu_w)^2$ for K kernels and u_w is the weighted mean of the mixture $\sum_{i=1}^{K} w_i m_i(a|h_{1:t-1})$.

Discussion

General Discussion

References

Anderson, J. (1991, 07). The adaptive nature of human categorization. , *98*, 409-429.

Griffiths, T. L., Lucas, C., Williams, J., & Kalish, M. L. (2009). Modeling human function learning with gaussian processes. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems 21* (pp. 553–560). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/3529-modeling-human-function-

Krause, A., & Ong, C. S. (2011). Contextual gaussian process bandit optimization. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 24* (pp. 2447–2455). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/4487-contextual-gaussian-proc

Kushner, H. (1963). A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. *Joint Automatic Control Conference*, *1*, 69 - 79. doi: 10.1109/JACC.1963.4168566

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. In *Proceedings of the 25th international conference on neural information processing systems - volume 2* (pp. 2951–2959). USA: Curran Associates Inc. Retrieved from http://dl.acm.org/citation.cfm?id=2999325.2999464

Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning* (1st ed.). Cambridge, MA, USA: MIT Press.

Tresp, V. (2001). Mixtures of gaussian processes. *Advances in Neural Information Processing Systems 13*.

Wang, Z., & Jegelka, S. (2017, 06–11 Aug). Max-value entropy search for efficient Bayesian optimization. In D. Precup & Y. W. Teh (Eds.), *Proceedings*

of the 34th international conference on machine learning (Vol. 70, pp. 3627–3635). International Convention Centre, Sydney, Australia: PMLR. Retrieved from http://proceedings.mlr.press/v70/wang17e.html