Goal-Specific and Multi-Task 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

How learners balance exploring new actions with unknown rewards and exploiting familiar actions with known rewards has typically been studied in the context of reinforcement learning, where the goal is to maximize cumulative reward over time. However, a number of other goals exist that have not been studied side-by-side in this paradigm, including optimization and active learning. We consider how human learners might adjust their explore-exploit strategy to perform optimally in each of these tasks. Additionally, we consider how human learners might take advantage of common structure in a multi-task setting.

Introduction

Many decisions require finding a balance between exploiting known information for short-term gain and exploring new sources of information that may lead to future reward. For example, a person might choose to go to a restaurant that they know well and have had multiple satisfying meals at, or choose to try a new restaurant that could result in a better or worse experience. This is known as the explore-exploit trade-off, and is often studied in the context or *reinforcement learning* (RL) (Sutton & Barto, 1998) where the learners actions with an environment elicit a reward, and the goal is to maximize cumulative reward over the course of the task. In particular, multi-armed bandit (MAB) problems consists of A possible actions, where the a-th action yields the random reward r_t drawn from an unknown distribution. Over T trials, the agent must choose $a_{1:T}$ such that $\sum r_{1:T}$ is maximized. The policy $\tilde{\pi}$ maps the observed history of actions and rewards $h_{1:t-1} = [(a_1, r_1), ..., (a_{t-1}, r_{t-1})]$ to the next action a_t for all possible histories $h_{1:t-1} \in \mathcal{H}$. The goal in a these problems is to find the optimal policy such that the expected reward over T trials, $\mathbf{E}[r_{1:T}|\tilde{\pi}(h_{1:t-1})]$, is maximized. This is the solution to the Bellman equation

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

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 T. This second term can be evaluated for trial $t+1$ by weighing the expected reward given all possible histories $\mathcal{H}_{1:t} = [(a_1, r_1), ..., (a_{t-1}, r_{t-1}), (a_t, r_t)]$ for $r_t \in \mathcal{R}$ by the probability of that history occurring:

$$\mathbf{E}[r_{t+1:T}|\tilde{\pi}(h_{1:t})] = \sum_{r_t \in \mathcal{R}} p(r_t|a_t, h_{1:t-1})\mathbf{E}[r_{1:T}|\tilde{\pi}(h_{1:t})] \quad (2)$$

If the reward r observed after choosing action a is drawn from the distribution $H(\theta_a)$, an agent must employ a strategy that balances choosing the action that maximizes the reward on trial t, $a_t^* = \text{argmax}_{a_t \in A}\mathbf{E}[H(\theta_{a_t})]$, and learning more about θ_A by choosing novel actions to maximize reward on future trials. As the problem of evaluating all possible histories is intractable, one of two general classes of approximations can be used. The first, simulation-based methods, involves using a subset of possible histories to evaluate the long-term reward of an action, and includes Thompson sampling (Thompson, 1933) and Monte Carlo tree search (Coulom, 2007). Rather than relying on simulations, myopic strategies define a value function for approximating long-term reward. Common myopic strategies include upper confidence bound (UCB) (Agrawal, 1995) and epsilon-greedy (Sutton & Barto, 1998) algorithms.

Contextual multi-armed bandit (CMAB) problems introduce additional information into the standard MAB problem by way of a set of features associated with the set of possible examples (Langford & Zhang, 2008). For example, a standard MAB formulation of the problem of choosing which restaurant to eat at assumes that the reward yielded by any two restaurants will be uncorrelated. However, it might be the case that these restaurants share a set of features (e.g. size, location, menu items) such that choosing similar restaurants can be assumed to yield similar rewards. Rather than having to execute an action to be able to evaluate its expected reward, considering shared features allows one to learn a function, $R : x_t \rightarrow r_t$ that maps features of the action a_t, x_{a_t} to that action's expected reward, r_t.

With the inclusion of context, contextual multi-armed bandits require the additional step of learning the function R between updating the history, $h_{1:t-1} = [(a_1, x_1, r_1), ..., (a_{t-1}, x_{t-1}, r_{t-1})]$, and choosing $a_t = \tilde{\pi}(h_{t-1})$. Formally, function learning describes how people predict a continuous-valued output given an input, and can be thought of as a continuous extension of category learning. Theories of function learning typically follow either a rule-based or similarity-based approach. Rule-based approaches posit that people learn this mapping by assuming that the unknown function belongs to a particular parametric family, then inferring the most likely parameterization after observing input/output pairs. For example Carroll (1963) considers polynomials up to degree 6, and Koh & Meyer (1991) consider power-law functions. While this approach attributes rich representations to learners, it not clear how these representations are acquired. Similarity-based theories suggest instead that learning is the result of forming associations between input/output pairs and generalizing these

associations to similar inputs. Busemeyer et al. (2005) implement a connectionist network where inputs activate a set of input nodes according to a Gaussian similarity function and each output node is activated according to learned weights between the input and output nodes. This approach does not require any assumptions about functional form and allows for flexible interpolation, but does not support generalization to inputs that are distant from past examples. While neither approach seems to fully capture human function learning, hybrid models have been introduced that take advantages from both rule and similarity-based theories. McDaniel & Busemeyer (2005) extend their connectionist model by including a layer of hidden nodes, each corresponding to a parameterization of a particular functional family.

More recently, Lucas et al. (2015) proposed Gaussian process regression (GPR) as a unified approach to function learning. 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. The squared exponential kernel:

$$k(x_i, x_j) = \sigma^2 \exp(-\frac{(x_i - x_j)^2}{2l^2}) \tag{3}$$

follows the assumption made by similarity-based models that similar inputs will map to similar outputs, with the parameter l determining how quickly the correlation between two inputs decreases as the distance between them increases. Similarly, many rule-based models can be expressed in terms of polynomial kernels

$$k(x_i, x_j) = \sigma^2 (x_i x_j + c)^d \tag{4}$$

of degree d. For the unobserved input x^*, the output y^* is drawn from a normal distribution with the posterior mean and variance functions

$$\begin{aligned} m(x^*|\mathbf{x}_N, \mathbf{y}_N) &= m(x^*) + \mathbf{k}(x^*)^T (\mathbf{K} + \sigma^2 \mathbf{I})(\mathbf{y}_N - m(\mathbf{x}_N)) \\ v(x^*|\mathbf{x}_N, \mathbf{y}_N) &= k(x^*, x^*) - \mathbf{k}(x^*)^T (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{k}(x^*) \end{aligned} \tag{5}$$

where $\mathbf{k}(x^*) = [k(x_1, x^*), ..., k(x_N, x^*)]$ and $\mathbf{K} = k(\mathbf{x}_N, \mathbf{x}_N)$.

Bayesian optimization (Snoek et al., 2012) provides a framework for solving CMAB problems when the reward function is assumed to be drawn from a Gaussian process. On each trial, the acquisition function $\text{acq}(a) = \text{acq}(m(a|h_{1:t-1}), v(a|h_{1:t-1}))$ is used to approximate the expected long-term reward of each action, and the next action is chosen according to the policy $\tilde{\pi}(h_{1:t-1}) = \text{argmax}_{a \in A} \text{acq}(a)$. Once the action/reward pair has been observed, the mean and variance functions at each action are updated. Of the commonly used acquisition functions, most follow the heuristic of favoring actions where both the mean and variance of the associated distribution over rewards are high; that is, where there is opportunity to either exploit a known reward or explore an unknown area of the reward function. The Gaussian process UCB acquisition function directly weighs exploration and exploitation by taking a weighted sum of the mean and variance of the reward function at a

$$\text{acq}_{UCB}(a) = m(a|h_{1:t-1}) + \beta \sqrt{v(a|h_{1:t-1})} \tag{6}$$

where β determines the preference for exploration. Another strategy is to maximize the probability of improving the current best reward:

$$\text{acq}_{PoI}(a) = \Phi\left(\frac{m(a|h_{1:t-1}) - m(a*|h_{1:t-1})}{\sqrt{v(a|h_{1:t-1})}}\right) \tag{7}$$

where a^* is the action that is believed to maximize reward and Φ is the cumulative distribution function of the standard normal. A similar strategy is to maximize the expected improvement:

$$\text{acq}_{EI}(a) = (m(a|h_{1:t-1}) - m(a^*|h_{1:t-1}))\Phi(z) + \sqrt{v(a|h_{1:t-1})} \tag{8}$$

where $z = \frac{m(a|h_{1:t-1}) - m(a^*|h_{1:t-1})}{\sqrt{v(a|h_{1:t-1})}}$. Rather than exploring based on the uncertainty of each action, the $\varepsilon -$ greedy strategy

$$\text{acq}_{\varepsilon - greedy}(a) = m(a|h_{1:t-1})p + \frac{1}{N}(1-p) \tag{9}$$

where ε determines the probability of choosing a random action with $p \sim \text{Bernoulli}(\varepsilon)$.

An alternative approach to maximizing this heuristic is to work directly with the distribution that describes the maximum of the function. Entropy search (Hennig & Schuler, 2012) achieves this by maximizing the expected reduction in entropy about $p(x^*)$, where $x^* = \text{argmax}_{x \in X} R(x)$, yielded after choosing the action a:

$$\text{acq}_{ES}(a) = H(p(x^*|h_{1:t-1})) - \mathbf{E}[H(p(x^*|h_{1:t}))] \tag{10}$$

and predictive entropy search (Henrández-Lobato et al., 2014) maximizes an equivalent value:

$$\text{acq}_{PES}(a) = H(p(y|h_{1:t-1}, x)) - \mathbf{E}[H(p(y|h_{1:t-1}, x, x^*))] \tag{11}$$

Max-value entropy search Wang & Jegelka (2017) takes a slightly different approach, instead working to minimize entropy about $p(y^*)$:

$$\text{acq}_{MES}(a) = H(p(y|h_{1:t-1}, x)) - \mathbf{E}[H(p(y|h_{1:t-1}, x, y^*))] \tag{12}$$

While each of these strategies has the advantage of working with a global distribution over features and their associated rewards rather than a local heuristic, none of the required entropies can be computed analytically. However, a number of approximations have been proposed.

Human learners' approach to the explore/exploit trade-off has typically been studied in the context of RL where the goal is to maximize cumulative reward (e.g. Bechara et al., 2005; Steyvers et al., 2009; Schulz et al., 2017). However, the need for this trade-off can be observed in a number of other tasks.

For instance, an agent might be instead interested in finding the maximum possible reward after T trials. This problem is known as *optimization* (e.g.). In these cases, earning a high reward on any particular task t is not important, as long as the actions lead to finding the global maximum. Another example is when an agent might instead be interested in learning the the reward function across all actions. These cases are described by *active learning* (AL) (e.g. L. Bramley & Speekenbrink, 2015; N. Bramley et al., 2016). In typical learning paradigms learners are asked to make judgments based on evidence that has been preselected for them. In contrast, AL describes tasks where the learner plays a role in selecting the evidence to observe.

While reinforcement learning, optimization, and active learning are often studied in isolation, there are cases where different tasks might rely on the same mapping between features and rewards. *Multi-task learning* (e.g.) involves exploiting common structure among distinct tasks in order to learn each task more effectively.

Modeling the Explore/Exploit Trade-Off

We consider models of the three types of task requiring an explore/exploit trade-off that were described above. While the mapping from features to rewards, $R : x_N \rightarrow \mathcal{R}$, can be described in similar terms for all three tasks, their goal-specific reward, and thus the function used to approximate long-term reward, are distinct.

Since the goal in RL tasks is to maximize cumulative reward over time, the goal-specific reward is the same as the reward, that is:

$$\mathbf{E}_{rl}[r'_t|a_t, h_{1:t-1}] = \mathbf{E}[r_t|a_t, h_{1:t-1}] \tag{13}$$

In contrast, the goal of the optimization problem is simply to find the maximum possible reward withing T steps. As such, the goal specific reward is defined as:

$$\mathbf{E}_{opt}[r'_t|a_t, h_{1:t-1}] = max(\mathbf{E}[r_t|a_t, h_{1:t-1}] - \sum_{i=1}^{t-1} r'_i, 0) \tag{14}$$

That is, on trial if trial t yields an increase in reward over the previous maximum reward, r'_t is the difference. If not, r'_t is 0. Since active learning is concerned with learning the reward function rather than the magnitude of the rewards themselves, its goal-specific reward can be described as the sum of the decrease in variance across all possible actions:

$$\mathbf{E}_{al}[r'_t|a_t, h_{1:t-1}] = \sum_{a \in A} \mathbf{Var}[r|a, h_{1:t-1}] - \mathbf{Var}[r|a, h_{1:t}] \tag{15}$$

We hope to address three questions through the comparison of a number of models and there abilities to capture human performance on these tasks. First, we ask whether human learners act rationally with respect to information about the underlying reward function; that is, are people explicitly modeling the underlying reward function as an intermediate step in their decision making process, or relying on a less costly heuristic? To answer this question, we compare the class of Gaussian process-based strategies (e.g. GP UCB, probability of improvement, expected improvement, entropy search) with a model based on gradient descent. Second, we ask whether human learners act rationally with respect to their task-specific goal. In other words, do people use different strategies depending on their current goal, or rely on a more general measure of utility (e.g. choose actions with high expected reward and high uncertainty). We compare three task specific strategies with the general strategies of GP UCB, probability of improvement, expected improvement, and epsilon greedy. Third, we ask whether people adopt strategies that take into account higher-order information about the underlying reward function across multiple distinct tasks. We compare both strategies that include information about R in their estimation of long-term reward across tasks and with those that consider only the long-term reward of the current task.

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:

$$\begin{aligned} R_{linear}(x) &= x \\ R_{quadratic}(x) &= -(x-55)^2 \\ R_{sinc}(x) &= \frac{\sin(x/2-30.000001)}{x/2-30.000001} \end{aligned} \tag{16}$$

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

$$\text{acq}_{RL}(a) = m(a|h_{1:t-1}) + (T-(t-1))\beta \cdot \text{acq}_{MES}(a) \tag{17}$$

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

$$\text{acq}_{opt}(a) = \begin{cases} m(a|h_{1:t-1}), & \text{if } t = T \\ \text{acq}_{MES}(a), & \text{otherwise} \end{cases} \tag{18}$$

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

$$\text{acq}_{RL}(a) = v(a|h_{1:t-1}) \tag{19}$$

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

Agrawal, R. (1995). Sample mean based index policies with o(log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, *27*(4), 1054-1078. Retrieved from http://www.jstor.org/stable/1427934

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (2005). The iowa gambling task and the somatic marker hypothesis: Some questions and answers. *Trends in Cognitive Sciences*, *9*(4), 159–162.

Bramley, L., & Speekenbrink. (2015, 5). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. , *41*, 708-31.

Bramley, N., Gerstenberg, T., & Tenenbaum, J. B. (2016). *Natural science: Active learning in dynamic physical microworlds.* 38th Annual Meeting of the Cognitive Science Society.

Busemeyer, J. R., Byun, E., & McDaniel, M. A. (2005). Learning functional relations based on experience with input-output pairs by humans and artificial neural networks..

Carroll, J. D. (1963). Functional learning: The learning of continuous functional mappings relating stimulus and response continua. *ETS Research Bulletin Series*, *1963*(2), i–144. Retrieved from http://dx.doi.org/10.1002/j.2333-8504.1963.tb00958.x doi: 10.1002/j.2333-8504.1963.tb00958.x

Coulom, R. (2007). Efficient selectivity and backup operators in monte-carlo tree search. In H. J. van den Herik, P. Ciancarini, & H. H. L. M. J. Donkers (Eds.), *Computers and games: 5th international conference, cg 2006, turin, italy, may 29-31, 2006. revised papers* (pp. 72–83). Berlin, Heidelberg: Springer Berlin Heidelberg.

Hennig, P., & Schuler, C. J. (2012, June). Entropy search for information-efficient global optimization. *J. Mach. Learn. Res.*, *13*(1), 1809–1837. Retrieved from http://dl.acm.org/citation.cfm?id=2503308.2343701

Henrández-Lobato, J. M., Hoffman, M. W., & Ghahramani, Z. (2014). Predictive entropy search for efficient global optimization of black-box functions. In *Proceedings of the 27th international conference on neural information processing systems - volume 1* (pp. 918–926). Cambridge, MA, USA: MIT Press. Retrieved from http://dl.acm.org/citation.cfm?id=2968826.2968929

Koh, K., & Meyer, D. (1991, 10). Function learning: Induction of continuous stimulus-response relations. , *17*, 811-36.

Langford, J., & Zhang, T. (2008). The epoch-greedy algorithm for multi-armed bandits with side information. In J. C. Platt, D. Koller, Y. Singer, & S. T. Roweis (Eds.), *Advances in neural information processing systems 20* (pp. 817–824). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/3178-the-epoch-greedy-algorit

Lucas, C., L Griffiths, T., Williams, J., & Kalish, M. (2015, 03). A rational model of function learning. , *22*.

McDaniel, M. A., & Busemeyer, J. R. (2005). The conceptual basis of function learning and extrapolation: comparison of rule-based and associative-based models. *Psychonomic bulletin & review*, *12*, 24-42.

Schulz, E., Konstantinidis, E., & Speekenbrink, M. (2017). Putting bandits into context: How function learning supports decision making. *bioRxiv*. Retrieved from http://www.biorxiv.org/content/early/2017/06/15/081091 doi: 10.1101/081091

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

Steyvers, M., Lee, M. D., & Wagenmakers, E. J. (2009). A bayesian analysis of human decision-making on bandit problems. *Cognition and Brain Theory*, *53*, 168–179.

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

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, *25*(3/4), 285-294. Retrieved from http://www.jstor.org/stable/2332286

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