Reward Function Complexity and Goals in Exploration-Exploitation Tasks

Brian Montambault (brian.montambault@tufts.edu)
Department of Computer Science, Tufts University

Christopher Lucas
School of Informatics, University of Edinburgh

Abstract

People are often faced with choices where there is a conflict between seeking reward and gathering information. In many of these cases there exists a functional relationship between the features associated with actions and their corresponding rewards. Accounts of how people make decisions in these circumstances have not considered how peoples' strategies depend on the complexity of this function, as well as the person's goal. In a sequential decision making task we found that people chose between a number of different exploration strategies, but that strategy selection did not necessarily align with goal or account for function complexity.

Keywords: Decision Making; Exploration-Exploitation; Contextual Multi-Armed Bandits

Introduction

In many of the decisions that people make in life there is a conflict between choices that are likely to have good results and choices where the result is more uncertain, but could possibly lead to a better outcome than the known option. For example, one might choose to eat at a familiar restaurant that is known to be good, or a new restaurant where the quality could be either better or worse. This trade-off is known as the explore-exploit dilemma. A structurally similar problem, with a slightly different goal is identifying the best candidate from a set of possible choices within a fixed time frame. For example, someone planning a party might wish to sample several possible caterers in order to find who will provide the best meal. Unlike the dilemma of choosing a restaurant for dinner, it is only important that the best option is found; the quality of any single meal is unimportant.

A common task for studying how people navigate explore-exploit dilemmas is the multi-armed bandit (MAB) task (Steyvers & Wagenmakers, 2009; Lee, Zhang, Munro, & Steyvers, 2011), where a decision-maker chooses between discrete actions, each with an unknown reward distribution, in order to maximize total reward over the course of several trials. While these tasks provide a simple environment for studying decision-making, real world tasks often contain additional contextual information about how rewarding an option might be. For example, we might have the option between two new restaurants, where the first has a menu with similar items to a past favorite, and the second has a menu that is full of new options. If we want to maximize the chance we will be satisfied, it would be prudent to pick the first. If we want to learn something new, we should choose the second. More formally, we can describe each option, a_i with the set of features s_i, with a_i yielding the reward $r_i = f(a_i, s_i)$, where f is a reward function mapping actions and features (or contexts) to rewards. We can call this a *contextual* multi-armed bandit (CMAB) (Li, Chu, Langford, & Schapire, 2010). In this setting, successful learners must make inferences about what this function might be – especially if there are many actions to choose from.

How people learn mappings between inputs and outputs, or *function learning*, has been widely studied (DeLosh, Busemeyer, & McDaniel, 1997). Recently, Gaussian process regression (GPR) has been presented as a model of function learning (Lucas, Griffiths, Williams, & Kalish, 2015). In addition to being a flexible non-parametric model capable of representing a wide range of functions, GPR is distinct from other accounts in that it directly allows for the representation of uncertainty in outputs. For CMAB tasks, this lays bare the trade-off between exploration and exploitation: An exploration-oriented agent can target options where uncertainty is greatest, an exploitation-oriented agent can target options with the highest expected reward, and it is possible to strike a balance between the two extremes. Bayesian optimization (Snoek, Larochelle, & Adams, 2012) is a flexible framework for transforming predictions from GPR models into actions. Several algorithms have been proposed for handling these tasks (Snoek et al., 2012) and have shown to both perform well (Srinivas, Krause, Kakade, & Seeger, 2010) and describe human behavior (Schulz, Konstantinidis, & Speekenbrink, 2018) in CMAB tasks. However, these accounts do not consider how one's strategy might be contingent on their ability to learn the reward function. This ignores a prominent result from the function learning literature: that some families of functions (e.g. linear) are easier to learn than others (e.g. periodic) (Kalish, Lewandowsky, & Kruschke, 2004).

While most work on MABs and CMABs study tasks where the goal is to maximize cumulative reward, there are circumstances where a decision-maker might instead be interested in finding the best action (Audibert & Bubeck, 2010). In the case of CMABs, this can be understood in terms of optimization, where the goal is to find some configuration of features (contexts) that maximize an objective function (reward). Bayesian optimization has shown to be of great practical use in these cases, in particular when the objective

function is expensive to evaluate as in the optimization of machine learning algorithm hyperparameters (Snoek et al., 2012). Bayesian optimization typically selects actions that have both a high expected reward and are highly uncertain, as in upper confidence bound (Auer, Cesa-Bianchi, & Fischer, 2002) and expected improvement (Mockus, 1974) algorithms. While these are reasonable strategies when the goal is to earn large rewards on each trial while still exploring new actions, they are ill-suited for optimization, where rewards on each trial are not important. Algorithms based instead on reducing uncertainty about the maximum of the reward function have been recently introduced (Hennig & Schuler, 2012; Wang & Jegelka, 2017) and appear better suited to this goal. Other recent work has examined the idea that people adapt their strategies to the tasks they face, accounting for both the expected performance of a strategy and the cost (e.g., in time) of executing it (Lieder, Helen, & Griffiths, 2017). If one hypothesis is that people adapt their strategies to the task at hand, and distinguish between optimization problems and ongoing trade-offs between exploration and exploitation, another is that people use a "one size fits all" strategy that supports multiple goal types reasonably well, as suggested by some past results, e.g., (Borji & Itti, 2013; Wu, Schulz, Speekenbrink, Nelson, & Meder, 2018).

For both reward maximization and optimization problems, good strategies must seek out information or reduce uncertainty. They can do this in an explicit or *directed* way, or achieve it implicitly by adopting a *stochastic* policy. In *directed* exploration one seeks actions that are most informative about the underlying reward distributions. One popular class of algorithms choose actions with high *upper confidence bounds* (UCB) (Auer et al., 2002), which typically include a free parameter β that controls the width of the confidence bound, directly controlling the preference for exploration over exploitation. In the case of UCB, exploration is directed by uncertainty about individual actions, where those with high uncertainty about their reward are more appealing than those with low uncertainty. In contrast, entropy-based strategies (Hennig & Schuler, 2012; Wang & Jegelka, 2017) are directed by uncertainty about global properties of the function – in particular uncertainty about the function maximum. In *stochastic* exploration, one seeks to explore the space of actions by applying some level of randomness to one's actions. While these methods are implicitly sensitive to reward uncertainty, they do not explicitly minimize it. Thompson sampling (Thompson, 1933) applies randomness to actions by first sampling a reward structure given previous observations, and then choosing the best action given the sampled rewards. Another method of random exploration is to choose actions with probabilities based on the softmax function

$$p(a_t = k) = \frac{\exp[m_t(k)/\tau]}{\sum_{k' \in A} \exp[m_t(k')/\tau]}$$

where $m_t(k)$ is the expected reward of arm k on trial t, and τ controls the level of randomness of actions, with all actions being equally likely as $\tau \to \infty$ and one deterministically choosing the action with the highest expected reward as $\tau \to 0$. While evidence for both directed and random exploration has been found in human behavior (Gershman, 2018), it has yet to be determined whether the criteria for directed exploration is dependent on the goal of the task or is exclusively based on uncertainty about individual actions, and under what conditions random exploration might be preferred over directed exploration.

Many of the real world explore-exploit dilemmas faced by people require learning a mapping between contexts and rewards, making CMABs an attractive environment for studying this phenomenon. While Bayesian optimization and other GPR-based approaches have been widely demonstrated to be a good model of human behavior in these tasks, there has been little research investigating how these frameworks capture different behaviors across distinct environments. While there has been work demonstrating that people are capable of learning functions and applying that representation to their decisions (Schulz et al., 2018), it is unclear how people's strategies might change when faced with functions of varying complexities, though some have varied function complexity by comparing smooth and rough non-parametric functions (Wu et al., 2018), and compared linear to quadratic reward functions (Stojic, 2016). While Bayesian optimization has been shown to describe human behavior well both when the goal is to maximize cumulative reward and when the goal is to find the best arm, it is unclear whether people choose a strategy to match their goal or use a more general strategy regardless of goal. Our contribution is to demonstrate how these factors influence people's strategies. We introduce a model based on Bayesian optimization that is capable of representing a rich set of behaviors revealed in prior work, and how different reward function complexities and goals might result in different parameterizations describing behavior.

Methods

Experiment. We designed a CMAB task in which participants were allowed to click one of several "actions" represented by a set of vertical bars situated along the x-axis of a plot. Upon clicking a bar, the reward of the associated action was revealed to the participant by displaying the height of the bar. Actions (bars) were related to rewards by their position on the x-axis: the i^{th} bar from left to right, a_i, was associated to the reward r_i by the function $r_i = f(a_i, i)$. We tested behavior on CMABs with three different reward functions of varying complexity:

$$f_{linear}(a_i, i) = i$$
$$f_{quadratic}(a_i, i) = -(i - 55)^2$$
$$f_{sinc}(a_i, i) = \frac{\sin(i/2 - 30.000001)}{i/2 - 30.000001}$$

Reward functions were scaled to fall within minimum and maximums drawn from uniform distributions, $\mathbf{U}(0,100)$ and $\mathbf{U}(400,500)$ respectively. Participants were shown 10 reward sample functions before they began the task. The quadratic and sinc samples were generated by uniformly sampling the location of the maximum, the function minimum, and function maximum ($\mathbf{U}(1,80)$, $\mathbf{U}(0,100)$, and $\mathbf{U}(400,500)$ respectively). Linear functions were generated by samples of the intercept and slope drawn from uniform distributions $\mathbf{U}(0,250)$ and $\mathbf{U}(0,6.25)$.

Participants were given one of two possible goals: In the maximum-finding condition, participants were asked to find the bar associated with the maximum possible reward. Participants final score in this condition was equal to the maximum reward uncovered across all trials. In the score-maximization condition, participants were asked to maximize their cumulative scores across all trials.

Procedure and participants. Participants (n=69, mean age=33.0 years) were recruited using Amazon's Mechanical Turk service. They were randomly assigned one of 6 (3 reward functions $\times$ 2 goals) conditions. They were first shown 10 different sets of 80 bars with their heights already revealed. Depending on a participant's function condition, the heights of the bars in each set was determined by either linear, quadratic, or sinc functions. Participants were then shown a new set of 80 bars, each 500 pixels tall and gray in color, and instructed to either find the bar with the largest height (find-max) or to maximize the cumulative heights of bars clicked across all trials (max-score) for a new set of bars. Participants were invited to click on any of the 80 bars over 25 trials. When a gray bar was clicked its color changed to black and its height was adjusted to match its corresponding reward (between 0 and 500 pixels). After each trial the reward associated with the chosen bar was used to update the participants goal-specic reward, 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. To incentivize performance participants were given a bonus up to \$0.75 proportional to the total number of points they earned.

Model

Our goal was to uncover strategies used in an CMAB task with different reward function complexities and goals. We take inspiration from Bayesian optimization, taking action probabilities to be a function of a GPR predictions of the reward function. Like previous accounts, we characterize exploration as a mixture of directed and random behavior. However, While previous accounts have assumed that directed exploration only uses uncertainty about each action, we extend this framework to include uncertainty about the function maximum.

In GPR a kernel function is used to encode prior beliefs about a function. We use the radial basis function (RBF) kernel:

$$k(x,x') = \sigma^2_{var}\exp(-\frac{(x-x')^2}{2l^2})$$

where l determines the smoothness of the function, or how quickly the similarity of two points falls off as they become more distant, and σ^2_{var} determines the average distance of the function from its mean. This kernel function is well suited to flexibly modelling function learning, as it is capable of learning any smooth function. For each reward function condition a set of 10 functions from the same family that were shown to participants prior to the CMAB task were used to fit the hyperparameters of the kernel function by maximizing the log marginal likelihood of the sample functions (Rasmussen & Williams, 2005). Fitting kernel hyperparameters in this way for each function allows us to model participants' expectations about the smoothness of the reward function, given the observed set of sample functions.

To estimate each participant's trial-by-trial predictions we compute the posterior mean and variance of the reward function at each action:

$$m_t(a) = \mathbf{k}_t(a)^\top(\mathbf{K}_t + \sigma^2_{noise}\mathbf{I})^{-1}r_t$$
$$v_t(a) = k(a,a) - \mathbf{k}_t(a)^\top(\mathbf{K}_t + \sigma^2_{noise}\mathbf{I})^{-1}\mathbf{k}_t(a)$$

where $k(a,a')$ is the covariance of two actions given the hyperparameters learned from a participant's training functions, $\mathbf{k}_t(a)$ is a vector of covariances for the action a and all previous observed actions, and $\mathbf{K}_t$ is the covariance matrix of all previously observed actions. σ^2_{noise} is the noise observed in the data. The reward functions in our task are deterministic, so we set this to a very small but non-zero number 10^{-4} to avoid numerical instability. We encode exploration directed by uncertainty about the function maximum by approximating the mutual information between the reward r revealed by action a and the highest possible reward r^*, $I(\{a,r\};r^*)$, the approximation used in max-value entropy search (Wang & Jegelka, 2017). We define the *utility* of each action on trial t to be

$$u(a,\beta,\lambda) = m_t(a) + \beta v_t(a) + \lambda I(\{a,r\};r^*)$$

and the probability of each action was defined using the softmax function

$$p(a|\beta,\lambda,\tau) = \frac{\exp[u(a,\beta,\lambda)/\tau]}{\sum_{a'\in A}\exp[u(a',\beta,\lambda)/\tau]}$$

We use an infinite groups model (Navarro, Griffiths, Steyvers, & Lee, 2006) to uncover common strategies across participants. Using the probability of actions, if the i^{th} participant belongs to group z,

$$p(a^i_T|g_i = z) = \prod_t^{T-1} p(a^i_{t+1}|a^i_t, r^i_t, \beta_z, \lambda_z, \tau_z)$$

where a^i_T is the set of all actions performed by the i^{th} participant. Groups were assigned priors according to a

stick-breaking procedure (Ishwaran & James, 2001). Under this prior we imagine a stick of length 1 that we break in two, keeping the length of the first stick to be the prior probability of our first group. We can then break the remaining piece in two again, with one of its pieces representing the prior probability of our second group. This process can be extended to represent a countably infinite number of groups, with the sum of their prior probabilities guaranteed to sum to 1. The stick-breaking prior has one parameter, α, that determines the dispersion among groups, with a higher α resulting in likelihoods being spread across a greater number of groups. We place a Gamma(a,b) prior over α, setting $a = b = 10^{-10}$ to represent our ignorance of the true number of groups in the data. We set Gamma priors with $a = b = 0.1$ over β, λ, and τ to represent equal preferences for each type of exploration.

Results

We used the python package PyMC3 (Salvatier, Wiecki, & Fonnesbeck, 2016) to perform inference. MCMC sampling was performed using the NUTS sampler, with 4 chains of 1000 samples each.

To inspect the range of strategies used by participants we assigned each participant to their most likely group, maximizing $p(g|a_T^i)$ for the i-th participant. Nine groups were assigned at least one participant. We summarize the behavior of each of these groups by their parameter means in Table 1. The largest four groups were assigned 48 out of 69 participants. The largest group has a much larger average τ than other groups, indicating that participants in this group heavily utilized random exploration. The second largest group had a larger average β and λ and smaller average τ, indicating that participants in this group utilize directed in addition to random exploration, using both uncertainty about each action and uncertainty about the reward function maximum. The third largest group also had a relatively large average β and λ, but a smaller average τ than the previous group. This indicates that participants in this group also used both forms of directed exploration, but relied much less on random exploration. The fourth largest group had relatively low average values for all three parameters, indicating that participants in this group did comparatively little exploring, instead choosing actions based on their expected reward. We refer to these groups as *stochastic*, *mixed*, *directed*, and *greedy* respectively.

To better understand how behaviors differed between groups, we measure the distance between participants' actions and both their previous action and their reward function maximum across trials (Figure 1). First, we plot the distribution of the distances between a participant's action and their previous action. Participants across all four of the top groups made a large proportion of their actions in close proximity to their previous action. This proportion was largest for the random group, followed by mixed, directed, and greedy. As we might expect, participants in the random group demonstrated more aggressive exploration with respect to their previous action, while those in the mixed and directed groups were more reserved. In contrast, participants in the greedy group rarely deviated far from their previous action. Next, we plot the median distance from the reward function maximum by trial for each group. For the random and mixed groups, the median distance from the reward function maximum stays level across trials, indicating that participants in these groups favor exploration over converging on the best action. For the directed and greedy groups, the median distance decreases towards zero with the number of trials. While the distance continues to decrease and eventually flattens out for those in the greedy group, the distance for those in the directed group increases after a number of trials, indicating that participants were willing to continue exploring even after the region containing the reward function maximum was located.

If participants were selecting there strategy based on their goal, we would expect the actions of participants in the max-score condition to be best predicted by a strategy that minimizes balances exploration and exploitation, and those of participants in the find-max condition to be best predicted by a strategy that minimizes uncertainty about the reward function maximum. While none of the groups show a preference for the source of uncertainty used to direct exploration (either about rewards of individual actions or the function maximum), these groups do differ in their preference for random and directed exploration. To investigate how reward function complexity and goal determine how people choose between these strategies we compare how well each strategy predicts the actions of participants grouped by experimental condition (Table 2). Participants selecting a strategy in the max-score goal condition are expected to choose a strategy that favors actions with high rewards, while those in the find-max condition are expected to choose a strategy that puts more of an emphasis on exploration. Our results contradict this assumption, with those in the max-score condition best described by the directed strategy and actions of those in the find-max condition best described by the greedy strategy. With function learning being increasingly difficult as function complexity increases, participants in less complex reward function conditions are expected to use this learning to engage in more directed exploration, while those in the more complex reward function conditions are expected to rely more heavily on random exploration. Our results show some evidence for this, as actions of participants in the linear condition were best explained by the directed strategy, while those in the quadratic condition were best explained by the mixed strategy. However, our results also show that the actions of those in the sinc condition were also best described by the directed strategy rather than the mixed or random strategy as we might expect.

	β	λ	τ	*N* Participants
Stochastic	0.29 ± 0.17	6.01 ± 8.87	5.23 ± 3.57	15
Mixed	1.58 ± 1.15	8.56 ± 8.32	1.77 ± 2.43	14
Directed	1.4 ± 1.29	11.19 ± 9.21	0.24 ± 0.22	11
Greedy	0.77 ± 0.44	0.61 ± 0.84	0.14 ± 0.19	8
	0.53 ± 0.27	4.06 ± 5.35	1.3 ± 1.14	7
	1.21 ± 0.55	6.38 ± 9.77	1.32 ± 0.61	6
	0.87 ± 0.33	8.77 ± 9.51	0.22 ± 0.24	4
	3.12 ± 2.33	0.99 ± 0.48	1.82 ± 1.14	2
	1.12 ± 0.39	6.98 ± 6.09	0.89 ± 0.63	2

Table 1: Mean parameters for each group

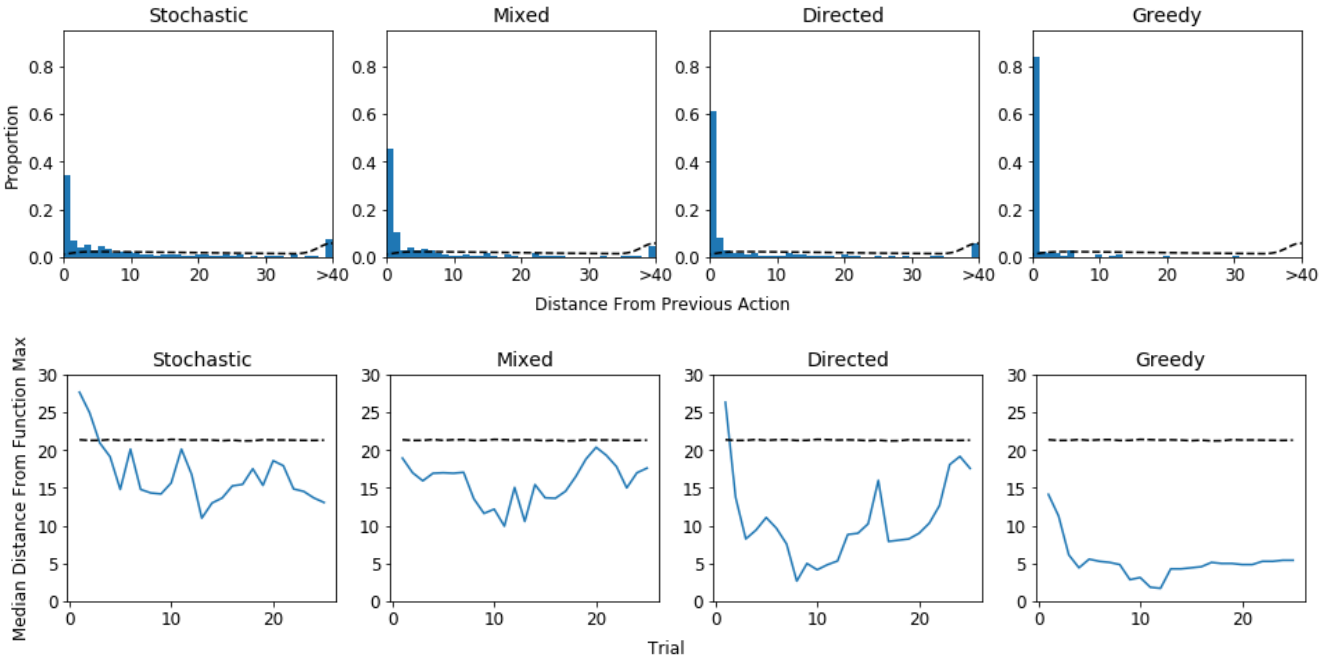

Figure 1: Distance of each action from the previous actions (top) and the reward function maximum (bottom) for the top four groups compared to a random baseline (dashed lines) reflecting uniform random action selection averaged over all conditions.

Discussion

In this study we compared behavior in an explore-exploit task across different reward function complexities and goals. While previous studies have characterized behavior in these tasks as some combination of exploitation and directed and random exploration, it was uncertain how these components might vary with different environments. Additionally, while previous studies only considered uncertainty about individual actions in directed exploration, it had yet to be established how measures of global uncertainty, such as uncertainty about the function maximum, might also be used by people to guide exploration.

Participants in this study each completed a CMAB task where the underlying reward function was either linear, quadratic, or sinusoidal, and their goal was to either maximize their score across all trials (max-score) or to find the best action (find-max). We found that behavior could be described by a relatively small set of strategies, characterized by varying exploration parameters. We found some evidence that strategy was impacted by reward function complexity, as participants in the linear condition were better described by a directed exploration strategy while those in the quadratic condition were better described by a mixed strategy utilizing both stochastic and directed exploration. However, those in the sinc condition were also best explained by a directed strategy, suggesting that these participants relied less on stochastic exploration than those in the quadratic condition despite their relatively complex reward function. Finally, we found that global uncertainty was indeed a measure used in directed exploration alongside uncertainty about individual actions, though we did not find evidence that preference for one form of uncertainty over the other was determined by goal. However, this could have been due to participants underestimating the complexity of the sinc reward function by only exploring around local maxima. Accounts of how reward function complexity influences strategy selection should also consider perceived complexity.

While we were able to describe a wide range of exploration behaviors, it is likely that alternative strategies exist. For example, some have suggested that people approach explore-exploit tasks in two qualitatively different phases, starting with a "pure exploration" phase, designed to reveal what options are most rewarding, and switching to a "pure exploitation" phase focusing on the most rewarding options (Steyvers & Wagenmakers, 2009). Another possibility is that some people do not utilize information about the reward function at all, instead exploring locally as often observed in ecological search strategies (Hills, 2006). A complete account of the types of strategies that people utilize under different circumstances should include a wider array of possible sources for guiding exploration.

References

Audibert, J.-Y., & Bubeck, S. (2010). Best Arm Identification in Multi-Armed Bandits. In *COLT - 23th Conference on Learning Theory*.

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, *47*(2), 235–256.

Borji, A., & Itti, L. (2013). Bayesian optimization explains human active search. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 26* (pp. 55–63).

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non of abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 968-986.

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, *173*, 34-42.

Hennig, P., & Schuler, C. J. (2012, June). Entropy search for information-efficient global optimization. *J. Mach. Learn. Res.*, *13*(1), 1809–1837.

Hills, T. T. (2006). Animal foraging and the evolution of goal-directed cognition. *Cognitive Science*, *30*(1), 3-41.

Ishwaran, H., & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, *96*(453), 161-173.

Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). Population of linear experts: knowledge partitioning and function learning. *Psychological Review*, *111*(4), 1072.

Lee, M. D., Zhang, S., Munro, M., & Steyvers, M. (2011, June). Psychological models of human and optimal performance in bandit problems. *Cogn. Syst. Res.*, *12*(2), 164–174.

Li, L., Chu, W., Langford, J., & Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on world wide web* (pp. 661–670). New York, NY, USA: ACM.

Lieder, F., Helen, A. A., & Griffiths, T. L. (2017). Strategy selection as rational metareasoning..

Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015). A rational model of function learning. *Psychonomic bulletin & review*, *22*(5), 1193–1215.

Mockus, J. (1974). On bayesian methods for seeking the extremum. In *Proceedings of the ifip technical conference* (pp. 400–404). London, UK, UK: Springer-Verlag.

Navarro, D., Griffiths, T., Steyvers, M., & Lee, M. (2006, 04). Modeling individual differences using dirichlet processes. *Journal of Mathematical Psychology*, *50*, 101-122.

Rasmussen, C. E., & Williams, C. K. I. (2005). *Gaussian processes for machine learning (adaptive computation and machine learning)*. The MIT Press.

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in python using PyMC3. *PeerJ Computer Science*, *2*, e55.

Schulz, E., Konstantinidis, E., & Speekenbrink, M. (2018). Putting bandits into context: How function learning supports decision making. *Journal of Experimental*

Function	Goal	Stochastic	Mixed	Directed	Greedy
All	Find Max	-108.95	-98.70	-99.36	**-97.43**
All	Max Score	-108.06	-99.37	**-88.55**	-92.21
Linear	All	-109.63	-101.13	**-88.38**	-92.01
Quadratic	All	-105.12	**-90.35**	-93.46	-92.05
Sinc	All	-111.19	-106.83	**-99.86**	-100.67
Linear	Find Max	-110.32	-99.10	-96.36	**-94.35**
Linear	Max Score	-108.93	-103.17	**-80.41**	-89.66
Quadratic	Find Max	-105.58	**-91.32**	-97.46	-95.76
Quadratic	Max Score	-104.71	-89.45	-89.76	**-88.62**
Sinc	Find Max	-111.25	-106.37	-104.43	**-102.32**
Sinc	Max Score	-111.14	-107.30	**-95.28**	-99.01

Table 2: Average log likelihoods per participant. For comparison, the log likelihood for one participant under a model that assumes judgments are made uniformly at random is -109.55.

Psychology, *44*(6), 927-943.

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.

Srinivas, N., Krause, A., Kakade, S., & Seeger, M. (2010). Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th international conference on international conference on machine learning* (pp. 1015–1022). USA: Omnipress.

Steyvers, M., & Wagenmakers, E.-J. (2009). A bayesian analysis of human decision-making on bandit problems..

Stojic, H. (2016). *Strategy selection and function learning in decision making*. Unpublished doctoral dissertation, Universitat Pompeu Fabra.

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.

Wang, Z., & Jegelka, S. (2017). 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.

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, *2*, 915-924.