Learning and decisions in contextual multi-armed bandit tasks

Eric Schulz[1](eric.schulz.13@ucl.ac.uk), **Emmanouil Konstantinidis**[2](em.konstantinidis@gmail.com), & **Maarten Speekenbrink**[1](m.speekenbrink@ucl.ac.uk)
[1]Department of Experimental Psychology, University College London, London, WC1H 0AP
[2]Department of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA 15213

Abstract

Contextual Multi-Armed Bandit (CMAB) tasks are a novel framework to assess decision making in uncertain environments. In a CMAB task, participants are presented with multiple options (arms) which are characterized by a number of features (context) related to the reward associated with the arms. By choosing arms repeatedly and observing the reward, participants can learn about the relation between context and reward and improve their decision strategy. We present two studies on how people behave in CMAB tasks. Within a stationary environment, we find that participants are best described by Thompson Sampling-based Gaussian Process models. This decision rule incorporates probability matching to the expected outcomes derived from a rational model of the task and it is especially well-adapted to non-stationary environments. In a dynamic CMAB task we again find that participants are best described by probability matching of Gaussian Process expectations. Our findings imply that behavior previously referred to as "irrational" can actually be seen as a well-adapted strategy based on powerful inference algorithms.
Keywords: Decision Making, Learning, Exploration-Exploitation, Contextual Multi-Armed Bandits

Introduction

Multi-armed bandit tasks have proven a useful framework to study learning and decision making (e.g., Steyvers et al., 2009). In a multi-armed bandit task, participants repeatedly choose between multiple options (arms) which have an associated reward and only the reward of the chosen option can be observed. Performing well in these tasks requires a fine balance between exploration (choosing arms in order to learn about their associated rewards) and exploitation (choosing arms which are thought to provide the maximum reward). In standard multi-armed bandit tasks, there is no additional information about the rewards that can be expected from an arm. In real life, such information is often present. For instance, when choosing a restaurant to eat in, there are various cues to the quality of the food on offer, such as the number of customers, the price of the dishes, the location of the restaurant, etc. These features provide contextual information that allows people to form expectations about the satisfaction the restaurant will provide. Contextual multi-armed bandits (Li et al., 2010) are a natural extension of classic multi-armed bandits and it is surprising that not much is known about learning and decision making in these tasks.

In what follows, we will introduce the Contextual Multi-Armed Bandit (CMAB) task and assess how participants perform in two different versions thereof. The experimental tasks can be approached as both a contextual bandit as well as a restless bandit (in which the average rewards associated with the arms vary over time) by ignoring contextual information, but are designed such that only taking the context into account will lead to above-chance performance. We will show that humans are able to learn well within the CMAB and are best described by sensitive exploration-exploitation behavior based on probability matching of choices to the predictions of non-parametric Bayesian models (Srinivas et al., 2009). These models do not try and learn one particular parametric structure, but rather a distribution over different generating mechanisms in a particular environment (see Gershman & Blei, 2012). Thompson sampling, a form of probability matching, offers a simple yet powerful way to balance exploration and exploitation, especially in non-stationary environments (Agrawal & Goyal, 2012; Speekenbrink & Konstantinidis, 2015). Our second experiment shows that the evidence for our model is even more pronounced in a dynamic environment where participants' choices influence future outcomes.

Contextual multi-armed bandits

A CMAB task can be seen as a game in which in each round $t = 1, \ldots, T$, an agent observes a context $s_t \in \mathcal{S}$ from a set $\mathcal{S}$ of possible contexts and has to choose an action $a_t \in \mathcal{A}$ from a set $\mathcal{A}$ of possible actions. Afterwards, she receives a reward $y_t = f(s_t, a_t) + \epsilon_t$ and it is her task to take those actions that produce the highest reward. The expected reward depends on the context, such that the agent has to learn the underlying function f; sometimes, this may require the agent to choose an action which is not expected to give the highest reward, but one that might provide useful information about f, thus choosing to explore rather than exploit.

For an agent who ignores the context s_t, the task would appear as a restless bandit task, as the rewards associated with an arm will vary over time due to the changing context. Learning the function f will make these changes in reward predictable and choosing the optimal arm easier. As it is not given that participants will learn the function, we will compare models of their behavior which are either *context-blind* and only learn based on direct feedback of the chosen arms, or *contextual* and learn the function relating context to reward. All models are based on inferring a predictive distribution of the reward $y_{k,t+1}$ on trial $t+1$ associated with arm k from the previous rewards $y_{1:t} = (y_1, \ldots, y_t)$, chosen arms $a_{1:t}$, and contexts $c_{1:(t+1)}$. For all models consid-

ered here, this predictive distribution is a normal distribution

$$p(y_{t+1}|y_{1:t}, a_{1:t}, c_{1:(t+1)}) = \mathcal{N}(M_{t+1}, V_{t+1}) \quad (1)$$

but the models differ in how they compute the mean M_{t+1} and variance V_{t+1}.

Learning

Context-blind learning Context-blind models only respond to the observed outcomes over time thereby ignoring the context completely.

μ-tracking The first context-blind model is based on tracking the mean μ_k reward associated to each arm k. The Bayesian μ-tracking model computes, on each trial, the posterior distribution of the mean and was implemented by a mean-stable version of the Kalman filter described next (by setting $\sigma_\zeta^2 = 0$).

Kalman filter Unlike the model above, the Kalman filter is a suitable model for tracking a time-varying mean. It is based on the following structural model

$$\begin{aligned} \mu_{k,t} &= \mu_{k,t-1} + \zeta_{k,t} & \zeta_{k,t} &\sim \mathcal{N}(0, \sigma_\zeta) \\ y_{k,t} &= \mu_{k,t} + \epsilon_{k,t} & \epsilon_{k,t} &\sim \mathcal{N}(0, \sigma_\epsilon) \end{aligned}$$

The mean of the predictive reward distribution of an arm k is computed as

$$M_{k,t} = M_{k,t-1} + \delta_{k,t} K_{k,t}[y_t - M_{k,t-1}] \quad (2)$$

where $\delta_{k,t} = 1$ if arm k was chosen on trial t, and 0 otherwise. The "Kalman gain" term is computed as

$$K_{k,t} = \frac{S_{k,t-1} + \sigma_\zeta^2}{S_{k,t-1} + \sigma_\zeta^2 + \sigma_\epsilon^2}$$

where $S_{k,t}$ is the variance of the posterior distribution of the mean reward, computed as

$$S_{k,t} = [1 - \delta_{k,t} K_{k,t}][S_{k,t-1} + \sigma_\zeta^2] \quad (3)$$

The variance of the predictive distribution is

$$V_t = S_t + \sigma_\zeta^2 + \sigma_\epsilon^2 \quad (4)$$

When fitting the model to participants' behavior, prior means and variances were initialized to $M_{k,0} = 0$ and $S_j(0) = 1000$, while σ_ζ and σ_ϵ were estimated by maximum likelihood.

Contextual learning The contextual models learn the functions f_k that map the context to the rewards. We will consider two contextual models: linear and Gaussian Process regression.

Bayesian linear regression Linear regression assumes the expected reward of an arm is an additive function of the m attributes of the context $s_t = (s_{1,t}, \ldots, s_{m,t})$:

$$y_{kt} = f_k(s_t) + \epsilon_{k,t} = \beta_0 + \sum_{i=1}^{m} \beta_i s_{i,t} + \epsilon_{k,t}$$

Bayesian linear regression starts with a prior distribution on the parameters β_i, $i = 0, \ldots, m$ and, from the contexts $s_{1:t}$ and rewards $y_{1:t}$ infers the posterior distribution over these parameters. These can then be used to compute the predictive reward distribution (1), with mean

$$M_{k,t} = \frac{1}{\sigma^2} {s_{t+1}}^\top A^{-1} S y \quad (5)$$

and variance

$$V_{k,t} = {s_{t+1}}^\top A^{-1} s_{t+1} \quad (6)$$

where $A = \sigma^{-2} S S^\top + \Sigma^{-1}$, with S being the context and y the outcomes observed so far.

Gaussian Process regression The second class of used models is non-parametric. Instead of postulating a specific parametric form, Bayesian non-parametric models implicitly assume that the function can be represented by an infinite number of parameters and let the data speak directly by the means of Bayesian inference. One example of a non-parametric model in the functional domain is a Gaussian Process (Rasmussen, 2006).

A Gaussian Process (henceforth $\mathcal{GP}$) is a collection of random variables from which every finite marginal distribution is multivariate Gaussian. A Gaussian Process can be expressed as

$$f(s) \sim \mathcal{GP}\big(m(s), k(s, s')\big). \quad (7)$$

where $m(s) = \mathbb{E}[f(s)]$ is the mean function and $k(s, s') = \mathbb{E}[(f(s) - m(s))(f(s') - m(s'))]$ the covariance function. We assumed a squared exponential kernel

$$k(s, s') = \exp\left(-\frac{(s - s')^2}{2\lambda^2}\right) \quad (8)$$

as covariance function with the lengthscale parameter λ. The predictive reward distribution (1) has mean

$$M_{k,t} = K(s_{t+1}, S)[K(S, S) + \sigma I]^{-1} y \quad (9)$$

and variance

$$\begin{aligned} V_t = {} & K(s_{t+1}, s_{t+1}) \\ & - K(s_{t+1}, S)[K(S, S) + \sigma I]^{-1} K(S, s_{t+1}) \end{aligned} \quad (10)$$

where K is the covariance matrix, S is the context seen so far, and σ is the noise level.

Decision strategies

We will consider two strategies to make decisions in a CMAB based on the expected outcomes according to the above learning models: the Upper Confidence Bound strategy and Thompson Sampling.

Upper Confidence Bounds (UCB) The upper confidence bound (UCB) algorithm, which has been shown to perform well in many real world tasks (Krause & Ong, 2011), balances the current expected value and the variance per arm and chooses the arm with the highest upper confidence bound. The UCB-algorithm can be described as a selection strategy with an exploration bonus, where the bonus depends on the 95% confidence interval of the estimated mean reward. As the UCB-algorithm is essentially deterministic while participants' decisions are expected to be more noisy, the following Softmax-transformation was used when fitting the strategy to participants' behavior

$$p(a_t = k) = \frac{\exp\{\gamma(M_{k,t} + 1.96\sqrt{V_{k,t}})\}}{\sum_{i=1}^{n} \exp\{\gamma(M_{i,t} + 1.96\sqrt{V_{i,t}})\}} \qquad (11)$$

The temperature parameter γ governs how consistent participants choose according to the values generated by the different models and was estimated by maximum likelihood.

Thompson Sampling Thompson sampling chooses each arm according to the probability that it provides the highest reward out of all arms in a particular context (May et al., 2012). This is a form of probability matching. The algorithm can be implemented by sampling for each arm a reward from the predictive reward distribution (1) and choose the arm with the highest sampled reward. Even though this model seems relatively simplistic, it can describe human choices in (non-contextual) restless bandit tasks well (Speekenbrink & Konstantinidis, 2015). Whereas psychology has generally viewed probability matching as an inferior decision strategy, Thompson Sampling has been shown to perform well in bandit tasks and can easily adapt to changing environments as it keeps on exploring other options over time.

The probability of an arm to be chosen can be expressed as

$$p(a_t = k) = p(\forall j \neq k : y_{k,t} \geq y_{j,t}) \qquad (12)$$

and computation from the predictive reward distributions is straightforward (see Speekenbrink & Konstantinidis, 2015).

Hypotheses

We conducted two experiments to test the following 3 hypotheses:

$\mathcal{H}1$**:** Participants will manage to learn within the introduced CMAB-setting and therefore be better described by contextual than by context-blind models.

$\mathcal{H}2$**:** Participants will approach contextual learning in a non-parametric fashion, allowing them to potentially learn different types of functions. Therefore, participants will be better described by the Gaussian Process than by the linear regression model.

$\mathcal{H}3$**:** Instead of maximizing output by a deliberate mean-variance trade-off, participants approach dynamic decision making problems using a probability matching heuristic. Thus, they will be best described by Thompson sampling.

Whereas $\mathcal{H}1$ is based on the assumption that participants can learn the true functions in the CMAB setting, $\mathcal{H}2$ follows recent successful attempts to describe function learning as non-parametric by Gaussian Process regression (Griffiths et al., 2009). That participants are better described by probability matching expected outcomes instead of a mean-variance trade-off ($\mathcal{H}3$) has been shown by Speekenbrink & Konstantinidis (2015) in a large model comparison study within the restless bandit setting.

Experiment 1 : Stationary CMAB

The first experiment was designed to test if participants can learn the functions in a stationary contextual bandit task.

Task

In the task, there were four different arms that could be played. In addition, three binary variables, $s_{j,t}$, $j = 1, 2, 3$, were introduced as the general context. These variables could either be on (+) or off (-). The outcomes of the four arms were dependent on the context as follows:

$$\begin{aligned} y_{1,t} &= 50 + 15 \times s_{1,t} - 15 \times s_{2,t} + \epsilon_{1,t} \\ y_{2,t} &= 50 + 15 \times s_{2,t} - 15 \times s_{3,t} + \epsilon_{2,t} \\ y_{3,t} &= 50 + 15 \times s_{3,t} - 15 \times s_{1,t} + \epsilon_{3,t} \\ y_{4,t} &= 50 + \epsilon_{4,t} \end{aligned}$$

with $\epsilon_{k,t} \sim \mathcal{N}(0, 5)$. Thus, the reward was a different linear function $f_k(s_t)$ of the context $s_t = (s_{1,t}, s_{2,t}, s_{3,t})$, producing an outcome $f_k(s_t) + \epsilon_{k,t}$.

On each trial, the probability that a context feature was on was set to $p(s_{j,t} = +) = 0.5$. The functions f_k were deliberately designed such that the expected reward over all possible contexts are identical with $\mathbb{E}[y_{k,t}] = 50$ in order to avoid first order stochastic dominance of context-blind choices. This means that the only way to gain higher rewards than the average of 50 is by learning how the context features influence the rewards. The

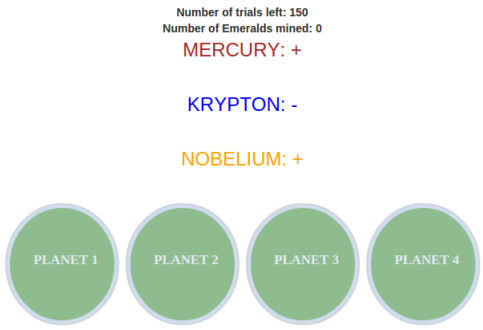

Figure 1: Screenshot of Experiment.

context-blind strategies therefore would not perform better than chance. Moreover, introducing an arm that only returns the overall mean with some added noise (Arm 4) helps us to distinguish even further between contextual and context-blind models. As context blind models only take the outcome into account, they should prefer Arm 4 as it produces the same mean over time, but exhibits less variance and therefore second-order dominates all the other arms. Contextual models on the other hand should tend to never select Arm 4 as taking the context into account will generally lead to outcomes which are better than the overall mean.

Participants

47 participants (26 males, age: $M = 31.9$, $SD = 8.2$) were recruited via Amazon Mechanical Turk and received \$0.3 plus a performance-dependent bonus of up to \$0.5 as a reward.

Procedure

Participants were told that they had to mine for "Emeralds" on different planets. Moreover, it was explained that at each time of mining the galaxy was described by 3 different environmental factors, "Mercury", "Krypton", and "Nobelium", that could either be on (+) or off (-) and had different effects on different planets. Participants were told that they had to maximize the overall production of Emeralds over time by learning how the different elements influence the planets and then picking the planet they thought would produce the highest outcome, given the currently available elements. It was explicitly noted that different planets can react differently to different elements. There were a total of 150 trials and which planet corresponded to which reward function f_k was determined randomly at the start of the experiment.

Results

The average score per round was 66.78 (SD=13.02) and most participants (38 out of 47) performed better than chance (an average score of higher than 50) as is confirmed by a simple t-test against $\mu = 50$, $t(46) = 7.17$, $p < 0.01$. That participants actually do learn over time while also sticking to some exploratory behavior can be see in Figure 2, where the density for higher scores increases and the density for lower scores decreases over trials.

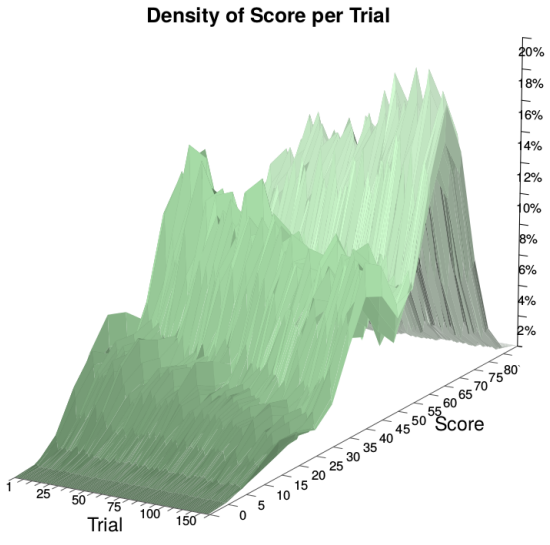

Figure 2: Distribution of obtained rewards (score) over participants by trial in Experiment 1.

Table 1: Average AIC, standard deviations, and the number of participants best fit by the different models.

Model	AIC_{mean}	AIC_{SD}	#best
Random	415.9	0	1
μ-track-UCB	392.1	39	2
μ-track-Thompson	388.1	56	3
Kalman-UCB	390.9	33	2
Kalman-Thompson	375.5*	50	11
Linear-UCB	387.8	34	3
Linear-Thompson	383.0	46	10
GP-UCB	389.4	34	4
GP-Thompson	381.6	42	12*

The overall performance of all models is shown in Table 1. In addition to the aforementioned models, we also included a Random baseline model, which assumes participants decided by random uniform guessing. It can be seen that the contextual models described participants behavior better than the two context-blind models. Altogether, 17 participants were best described by the context-blind models, whereas 29 participants were best described by the contextual models.

Even though the Kalman-Thompson model resulted in the lowest average AIC-value overall, the Gaussian Pro-

cess models described more participants best (16), more than the linear regression models (13) or the Kalman filter (13). The good performance of the Kalman filter might be due to the fact that some people did mostly try to learn on which planet they should mine, which is also indicated by the relative large variance of the two Kalman models. Even though there is only a small difference between the Gaussian Processes and the linear model, it is evermore surprising as the linear model here would be the best description of the underlying system a priori – the task is a linear system after all. What this tells us is that instead of approaching the problem with a fixed parametric representation in mind, participants might indeed apply a learning strategy that is more easily adaptable to other scenarios than a linear one.

Lastly, more people were described best by Thompson sampling than by the UCB strategy (36 vs. 10). This indicates that participants seem to apply this probability matching heuristic.

Intermediate Conclusion

Within a newly introduced task called the Contextual Multi-Armed Bandit task, we have found that participants can best be described by probability matching outcomes of a (close to) rational non-parametric function learning engine. Probability matching used to be referred to as "biased" or "irrational" (Stanovich & West, 2008), but can actually constitute a very sensible strategy, especially in dynamically changing environments (Agrawal & Goyal, 2012). Therefore, one would expect that participants should still be able to perform well even in a dynamically changing environment. The second experiment was designed to test this.

Experiment 2: Dynamic Contextual Multi-Armed Bandit

The second experiment used a similar task as before. However, this time the reward of a given arm (planet) was dependent on how often the particular arm had or had not been chosen previously. The rewards were determined according to the following functions

$$
\begin{aligned}
y_{1,t} &= 50 + 15 \times s_{1,t} - 15 \times s_{2,t} + \epsilon_{1,t} + \zeta_1(t) \\
y_{2,t} &= 50 + 15 \times s_{2,t} - 15 \times s_{3,t} + \epsilon_{2,t} + \zeta_2(t) \\
y_{3,t} &= 50 + 15 \times s_{3,t} - 15 \times s_{1,t} + \epsilon_{3,t} + \zeta_3(t) \\
y_{4,t} &= 50 + \epsilon_{4,t} + \zeta_4(t),
\end{aligned}
$$

where

$$
\zeta_j(t) = \begin{cases} -1, & \text{if } a_{t-1} = j \\ \frac{1}{3}, & \text{otherwise} \end{cases} \tag{13}
$$

This means that every time an arm is chosen its mean reward decreases by 1 point while the means of the unchosen arms increase by $\frac{1}{3}$, thereby creating a dynamic environment in which past choices directly influence future outcomes.

Participants

47 participants (30 males, age: $M = 29.1$, $SD = 8.6$) were recruited via Amazon Mechanical Turk and received \$0.3 plus a performance-dependent bonus of up to \$0.5 as a reward.

Procedure

The procedure was as in Experiment 1, but participants were told that their choices could influence future outcomes.

Results

On average, participants obtained rewards of 59.84 (SD = 9.41). Even though this task was deliberately set up to be more difficult, participants' overall average score was again above chance, $t(46) = 7.17$, $p < 0.01$. In total, 41 out 47 participants performed better than chance. Figure 3 indicates that the evidence of learning was somewhat weaker than before.

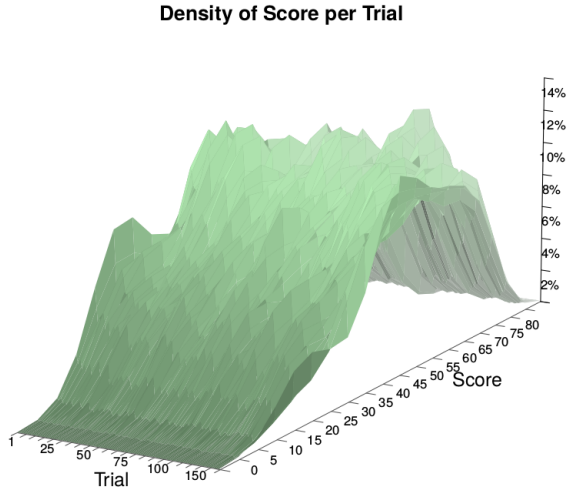

Figure 3: Distribution of obtained rewards (score) over participants by trial in Experiment 1.

While scores tended to increase over trials, this was not as pronounced as in Experiment 1. This might be due to the increase in difficulty of the task, as participants had to both learn a function and take the dynamics of their actions into account.

The overall performance of all models is shown in Table 2. Again, the contextual models described more people best than the context-blind models (14 vs. 30). Thus, even in this more complicated scenario, people seem able to learn about the relation between the context and rewards. The non-parametric models again described more people best than the linear regression models (19 vs. 12) or the Kalman filter (19 vs. 12). Finally,

Table 2: Average AIC, standard deviations, and the number of participants best fit by the different models.

Model	AIC_{mean}	AIC_{SD}	#best
Random	415.9	0	2
μ-tracking-UCB	414.5	10	0
μ-tracking-Thompson	416.8	5	0
Kalman-UCB	385.1	44	5
Kalman-Thompson	387.8	39	9
Linear-UCB	331.0	88	3
Linear-Thompson	321.1	99	9
GP-UCB	349.5	72	3
GP-Thompson	316.9*	104	16*

the Thompson-sampling based strategies described more people best than the UCB strategy (34 vs 11). Overall, the Thompson sampling $\mathcal{GP}$-model described most people best (16) reaching a mean AIC of 316.9.

Discussion and Conclusion

We have introduced the Contextual Multi-Armed Bandit (CMAB) task as a paradigm to investigate decision making in situations where one has to learn contextual functions and simultaneously make decisions according to the predictions of those functions. The CMAB-task here can be seen as a natural extension of past experiments on learning in traditional multi-armed bandit tasks.

In both a stationary and a dynamic task, we found that participants mostly performed above chance and were best described as probability matching to expected outcomes according to a rational Gaussian process function learning model. The above-chance performance shows that participants were able to learn the relation between context and rewards. The good performance of the GP model opens up the field of decision making to a powerful class of general purpose non-parametric learning models. The good performance of the Thompson sampler replicates the results of Speekenbrink & Konstantinidis (2015) in a non-contextual restless bandit task. It shows that probability matching, a behavior often frowned upon as irrational, provides a sensitive strategy that people might actually apply to solve the exploration-exploitation dilemma in a range of situations. This is also what we have confirmed in our second experiment, where participants were even better described by a Thompson sampling algorithm in a more dynamic scenario, where rewards depended on past choices. In conclusion, all of our three main hypotheses were confirmed.

This paper is only a first step into research on CMAB problems. Future studies could try to assess how people behave in scenarios where more context is provided either by creating a multi-context environment (for example, one context per planet) or by providing continuous context variables (for example, values between 0 and 10). Another simple modification could be to check different parameterizations of the underlying functions to differentiate even further between the different candidate models.

Finally, we have only introduced a comparison between a linear model and a Gaussian process in what can be described as an active learning task. In future experiments we aim to try and compare even more elaborate models within this context. Using an exploration-exploitation domain as a platform to compare models against each other might be a useful additional approach to decide among models from a list of many potential candidates (Schulz et al., 2014).

Acknowledgements

ES is supported by the UK Centre for Training in Financial Computing & Analytics. Parts of this work have been presented as Schulz et al. (2015).
Materials can be found at:
https://github.com/ericschulz/contextualbandits

References

Agrawal, S., & Goyal, N. (2012). Thompson sampling for contextual bandits with linear payoffs. *arXiv preprint arXiv:1209.3352*.

Gershman, S. J., & Blei, D. M. (2012). A tutorial on bayesian nonparametric models. *Journal of Mathematical Psychology*, *56*, 1–12.

Griffiths, T. L., Lucas, C., Williams, J., & Kalish, M. L. (2009). Modeling human function learning with gaussian processes. In *Advances in Neural Information Processing Systems*, (pp. 553–560).

Krause, A., & Ong, C. S. (2011). Contextual gaussian process bandit optimization. In *Advances in Neural Information Processing Systems*, (pp. 2447–2455).

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). ACM.

May, B. C., Korda, N., Lee, A., & Leslie, D. S. (2012). Optimistic bayesian sampling in contextual-bandit problems. *The Journal of Machine Learning Research*, *13*(1), 2069–2106.

Rasmussen, C. E. (2006). Gaussian processes for machine learning.

Schulz, E., Konstantinidis, E., & Speekenbrink, M. (2015). Exploration-exploitation in a contextual multi-armed bandit task. In *International Conference on Cognitive Modeling*, (pp. 118–123).

Schulz, E., Speekenbrink, M., & Shanks, D. R. (2014). Predict choice – a comparison of 21 mathematical models. In *36th Annual Conference of the Cognitive Science Society*, (pp. 2889–2894). Cognitive Science Society.

Speekenbrink, M., & Konstantinidis, E. (2015). Uncertainty and exploration in a restless bandit problem. *Topics in Cognitive Science*, *7*, 351–367.

Srinivas, N., Krause, A., Kakade, S. M., & Seeger, M. (2009). Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*.

Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of personality and social psychology*, *94*(4), 672.

Steyvers, M., Lee, M. D., & Wagenmakers, E.-J. (2009). A bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, *53*(3), 168–179.