Exploration-Exploitation in a Contextual Multi-Armed Bandit Task

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

We introduce the Contextual Multi-Armed Bandit task as a method to assess decision making in uncertain environments and test how participants behave in this task. Within an experimental paradigm named Mining in Space, participants see 4 different planets that are described by 3 different binary elements (the context) and then have to decide on which planet they want to mine (which arm to play). We find that participants adapt their decisions to the context well and can best be described by a Contextual Gaussian Process algorithm that probability matches according to expected outcomes. We conclude that humans are well-adapted to contextualized bandit problems even in potentially non-stationary environments through probability matching, a heuristic that used to be described as biased behavior. We argue that Contextual Bandit problems can provide further insight into how people make decisions in real world scenarios.
Keywords: Decision Making, Active Learning, Exploration-Exploitation, Contextual Multi-Armed Bandits

Introduction

A Contextual Multi-Armed Bandit (CMAB) task is a task in which an agent is confronted with mutliple options ("arms" of a bandit) out of which one can be chosen. The context describes the currently available information that can be utilized to choose the best arm to play (Li et al., 2010). This scenario is a good model for many real world problems; from choosing what to eat, to buying clothes in a shop, all the way to finding the right person to befriend; many situations require us to make the right choice in a given context without the chance to actually observe the outcome of unchosen options, constantly trading-off between *exploration* (trying out new things) and *exploitation* (maximizing expected reward). Therefore, contextual bandit tasks might help to shed light on how we make contextual decisions in general and on how we integrate information into our decisions in particular.

Despite a vast amount of research on multi-armed bandit tasks (Steyvers et al., 2009), little is known about participants' behavior in experiments involving contextual bandits. This is remarkable given that contextual bandits provide us with a scenario in which, instead of treating learning and decision making distinctively, participants have to learn a function that maps a context to outcomes and then act according to their predictions of these. In that sense, contextual bandit tasks could be seen as a quintessential scenario of everyday decision making.

In what follows, we will introduce the contextual multi-armed bandit task (`CMAB`) and probe how participants perform in one simple version thereof. The experimental task can be approached as both a contextual bandit as well as a so-called restless bandit (in which the average rewards associated with the arms vary over time) by ignoring information, but is 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 decisions to the estimated outcomes 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 given environment. Moreover, probability matching (also called *Thompson sampling*) offers a simple yet powerful way to balance exploration and exploitation in decisions, especially in non-stationary environments. The main contributions of this paper are threefold:

1. We introduce the `CMAB` as an experimental paradigm and emphasize its importance for psychological research.
2. We model human context learning as non-parametric: instead of relying on an arbitrary set of parametric candidate models, participants seem to learn in a way that represents distributions over generating mechanisms.
3. We show that participants apply a behavior best-described by Thompson sampling/Probability Matching. This behavior has often been referred to as biased and erroneous fallacy. However, it turns out to be a satisfyingly sensible strategy in dynamic environments (see Agrawal & Goyal 2012, for further details).

Definitions and Models

Contextual Bandit Problems

Consider a game in which, in each round $t = 1, \ldots, T$, an agent observes a context $s_t \in S$ from a set of S contexts and has to choose an action $a_t \in A$ from a set of possible actions A. The agent then receives a payoff $y_t = f(s_t, a_t) + \varepsilon_t$. It is the agent's task to take those actions that produce the highest payoff. As the expected payoff depends on the context, 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 payoff, but which might provide more information about f, thus choosing to explore rather than exploit. As the different actions are normally described as playing a bandit's arm and the context provides information that might help to find the right arm to play, these games are called *Contextual Multi-Armed Bandit* tasks.

Different models can be used to learn in a contextual bandit setting. The models applied here broadly fall within two categories: *context-blind* and *contextual* models. Context blind

models ignore the provided context completely and only learn based on direct feedback of the chosen arms. Contextual models do take the context into account and therefore are generally expected to perform better than context-blind models.

We will first describe a general choice rule, then the context blind models, and afterwards the parameterization of the two used contextualized models (linear and Gaussian Process regression) before then describing two different decision rules that can be used for the contextual models.

Choice rule

In the psychological task considered later, the context s_t at time t will be the same for all arms, while the function that maps the context to the (expected) payoff of the arm will vary over arms. The task is therefore to learn functions f_k for each arm that map the context to the payoff and then choose the arm with the highest expected payoff while constantly trading-off between exploration and exploitation. To do so, the models proposed here produce n different values $\theta_{1,t}, \theta_{2,t}, \ldots, \theta_{n,t}$ to compare between the n different arms at a time point t given the current context s_t by some learned function f_k that matches the context s_t to the considered arm k:

$$\theta_{k,t} = f_k(s_t) \tag{1}$$

This could be the mean predicted outcome for every arm or any other value as described below. In order to transform these values to a probability of picking a given arm arm_j, the values are transformed by a softmax rule with inverse temperature parameter γ as in Equation 2.

$$p(\text{arm}_t = k) = \frac{\exp\{\gamma\, \theta_{k,t}\}}{\sum_{i=1}^{n} \exp\{\gamma\, \theta_{i,t}\}} \tag{2}$$

Context-blind Models

Context-blind models ignore the context completely and only respond to the observed outcomes of arms over time.

Random choice The most simplistic context-blind model is a random choice. This model picks every arm with equal probability $p(\text{arm}_t = k) = 1/\#\text{arms}$. As this model does not learn over time, it will provide a baseline against which all the other models can be compared.

μ-tracking The other context-blind model is based on simple mean tracking.

$$\theta_{k,t} = \hat{\mu}_{k,t} = \frac{1}{n} \sum_{\tau=1}^{t} \delta_{\text{arm}_\tau = k} y_\tau \tag{3}$$

where $\delta_{\text{arm}_t = k} = 1$ if arm k is chosen at time t and 0 otherwise.

Contextual Models

The contextual models learn the functions f_k that map the context to the (expected) payoff for each arm. Here, we will consider two contextual models: linear and Gaussian Process regression.

Linear Regression Linear regression is a simple approach to learn each function f_k that relates the contexts s_t to an output $f_k(s_t)$. Each context s_t has values on a total of m attributes, i.e., $s_t = (s_{1,t}, \ldots, s_{m,t})$. The regression model learns a linear function of the context attributes:

$$\hat{f}_k(s_t) = \beta_0 + \sum_{i=1}^{m} \beta_i s_{i,t} + \varepsilon_t \tag{4}$$

Let $s_{1:t} = (s_1, \ldots, s_t)$ denote all the contexts encountered at time t. The regression model is estimated from $s_{1:t}$ and then used to predict new outcomes for each arm given a new contexts at $t+1$. Once the new output has been chosen, the regression model is updated and then used for the next trial with new contexts. As this is a parametric model, it assumes that participants approach the problem in a way that only allows for linear effects of the context. In order for the regression approach to not suffer from matrix deficiencies, 10 pseudo-observations were created from a Normal distribution with $\mathcal{N}(50, 10)$.

Gaussian Process Regression Another class of models is non-parametric. Instead of postulating one concrete parametric form (e.g., a linear one) out of an infinite set of possible forms (a choice that, without any further knowledge, is arbitrary), 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.

A Gaussian Process (henceforth $\mathcal{GP}$) is a collection of random variables from which every finite marginal distribution is multivariate Gaussian. We define a mean function $m(x)$ and the covariance function $k(x,x')$ of a process $f(x)$ as

$$m(x) = \mathbb{E}[f(x)] \tag{5}$$

$$k(x,x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))] \tag{6}$$

A Gaussian process then can be expressed as

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

Even though many different covariance functions exist, within all the examples and calculations presented here the *squared exponential* covariance function with a length scale λ will be used.

$$\text{cov}\big(f(x_p), f(x_q)\big) = k(x_p, x_q) = \exp\left(-\frac{|x_p - x_q|^2}{2\lambda}\right) \tag{8}$$

The lengthscale λ was estimated by using gradient descent.

In the noisy situation that will be analyzed in all of the upcoming situations, the covariance can be written as follows

$$\text{cov} = (y_p, y_q) = k(x_p, x_q) + \sigma_n^2 \delta_{pq}, \tag{9}$$

where δ is Kronecker's δ, which is 1 if $p = q$ and 0 otherwise.

Suppose we have collected observations $\mathbf{y}_t = [y_1, y_2, \ldots, y_t]^\top$ at inputs $\mathbf{x}_t = \{x_1, \ldots, x_t\}$, $y_t = f(x_t) + \varepsilon_t$,

$\varepsilon_t \sim \mathcal{N}(0,\sigma^2)$, then the posterior over f is a $\mathcal{GP}$ with mean $m_T(x)$, covariance $k_T(x,x')$, and variance $\sigma_T^2(x)$:

$$m_T(x) = \mathbf{k}_T(x)^\top (K_T + \sigma^2 \mathbf{I}) y_T \tag{10}$$

$$k_T(x,x') = k(x,x') - \mathbf{k}_T(x)^\top (\mathbf{K}_T + \sigma^2 \mathbf{I})^{-1} \mathbf{k}_T(\mathbf{x}') \tag{11}$$

$$\sigma_T^2(x) = k_T(x,x) \tag{12}$$

where $\mathbf{k}_T(x) = [k(x_1,x),\ldots,k(x_T,x)]^\top$ and $\mathbf{K}_T$ is the positive definite kernel matrix $[k(x,x')]_{x,x' \in A_T}$. The $\mathcal{GP}$ is used in the same fashion as the linear model by always using all the contexts and observed outcomes up to time point t in order to make predictions for time point $t+1$. Therefore, a $\mathcal{GP}$ for every arm over time will be estimated given the observed context as shown in Equation 13.

$$\hat{f}_j(s) \sim \mathcal{GP}\big(m(s), k(s,s')\big). \tag{13}$$

As a Gaussian Process is a non-parametric model for function learning, its application represents the assumption that participants do not a priori expect one parametric form of a function, but rather learn the form by the observed data over time. The Gaussian Process was initialized by the use of 10 pseudo-observations as in the regression approach described before.

Sampling strategies

Let us now look at the different algorithms that can be used to apply the two contextual models in a `CMAB`. Sampling strategies here mean different ways by which one could come up with a choice of an arm, given the estimated expected outcomes at a given time.

Upper Confidence Bounds The upper confidence bound algorithm estimates a trade-off between the current expected value and the variance per arm and optimistically picks the arm with the highest upper confidence bound. This algorithm has been shown to perform well in many real world contextual bandit tasks (Krause & Ong, 2011). The way a UCB-sampling agent would select an arm is described in Algorithm 1.

Algorithm 1 Upper Confidence Bands Sampling

Require: Context $\mathbf{s}$; Models $\mathcal{M}_{j,t-1}$
  for $t = 1,2,\ldots,T$ **do**
    Choose $\text{arm}^*_{jt} = \text{argmax}\ \mu(\mathcal{M}_{j,t-1}(\mathbf{s})) + 1.96\sigma(\mathcal{M}_{j,t-1}(\mathbf{s}))$
    Sample $y_t = f(\text{arm}^*_{jt}) + \varepsilon_t$
    Update $\mathcal{M}_{j,t-1} \rightarrow \mathcal{M}_{j,t}$
  end for

The trade-off is based on a confidence interval approximation based on a normal distribution and therefore the trade-off parameter is set to 1.96, marking the 95% confidence interval. The UCB-algorithm can be seen as a selection strategy with an exploration bonus, where the bonus depends on the confidence interval of the estimated mean return. As we will need probability estimates to model participants choices later on, the estimates for arm^*_{jt} were fed into the softmax equation described above.

Thompson Sampling Thompson sampling chooses each arm according to the (subjective) probability that it provides the highest payoff out of all the available arms, given the context (May et al., 2012). This is a form of probability matching. The algorithm can be implemented by sampling for each arm a payoff according to the learned models of the arms, and then choose the arm with the highest sampled payoff. Even though this model seems very simplistic, it can perform reasonably well in contextual bandit tasks and can describe human choices in (non-contextual) restless bandit tasks well (Speekenbrink & Konstantinidis, 2014). Whereas psychology has looked at probability matching as an inferior strategy of decision making for a long time, it has been shown to perform well in many restless bandit tasks and can easily adapt to changing environments as it still keeps on exploring other options over time.
An agent following the Thomson sampling algorithm would pick the next arm as described in Algorithm 2.

Algorithm 2 Thompson Sampling

Require: Contexts $s_{1:T}$; Models $\mathcal{M}_j$
  for $t = 1,2,\ldots,T$ **do**
    for $\text{arm}_{k,t}, k = 1,\ldots,n$ **do**
      Sample $y^*_{k,t-1} \sim \mathcal{M}_{k,t-1}(s_t)$
    end for
    Choose $\text{arm}_t = \arg\max_k y^*_{k,t}$
    Sample $y_t = f(\text{arm}_t) + \varepsilon_t$
    Update $\mathcal{M}_{j,t-1} \rightarrow \mathcal{M}_{j,t}$
  end for

Main advantages of Thompson sampling are (1) that it does not rely on additional parameter tuning, and (2) that it can adapt to many diverse environments. The probability of an arm to be chosen was calculated as shown in Equation 14.

$$p(\text{arm}_t = k) = p(\forall j \neq k : y^*_{k,t} \geq y^*_{j,t}) \tag{14}$$

This means that each arm is predicted to be chosen by its probability to produce the highest outcome at a given time.

Summary of all models

Taking all of the models (context-blind and contextual) and choice rules together results in the models shown in Table 1.

Class	Algorithm	Description
Context-blind	Random	Picks at random
	μ-tracking	Picks tracked mean
Linear	UCB	Picks upper confidence band
	Thompson	Probability matching
Gaussian Process	UCB	Picks upper confidence band
	Thompson	Probability matching

Table 1: Summary of all used models

Experiment : Contextual Bandit Task

The experiment was designed to test if participants are able to learn in a contextual bandit task. It used a relatively simple description of the context s and the different arms. Within this first CMAB experiment we focused on a task with three binary context variables that could either be on (+) or off (-) and 4 different arms.

Contextual Bandit setting

The outcomes of the different arms in dependency of the context are shown in Equations 14-17.

$$y_{1,t} = 50 + 15 \times s_{1,t} - 15 \times s_{2,t} + \varepsilon_{1,t} \quad (15)$$
$$y_{2,t} = 50 + 15 \times s_{2,t} - 15 \times s_{3,t} + \varepsilon_{2,t} \quad (16)$$
$$y_{3,t} = 50 + 15 \times s_{3,t} - 15 \times s_{1,t} + \varepsilon_{3,t} \quad (17)$$
$$y_{4,t} = 50 + \varepsilon_{4,t}, \quad (18)$$

with $\varepsilon_{k,t} \sim \mathcal{N}(0,5)$. This means that each arm reacted differently to the context $s_t = (s_{1,t}, s_{2,t}, s_{3,t})$ through linear functions, producing an outcome $f_k(s_t) + \varepsilon_{k,t}$ as described before.

For all different contexts, the probability of being + was set to $p(s_{j,t} = +) = 0.5$. The different arms were deliberately set up such that all the expected values are the same, $\mathbb{E}[y_{k,t}] = 50$ over time in order to avoid first order stochastic dominance of context-blind choices[1]. This means that the only way to gain higher values than the individual bandits' averages is by learning how the different factors influence the arms within every trial. The 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 (at the end) almost never select Arm 4 as taking the context into account will generally lead to better outcomes than the simple mean alone.

Methods

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. None of the participants were excluded from the remaining analysis.

Design 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. The total number of trials was fixed to be 150 and the experiment was well-received on Mechanical Turk.[2]

Notice that this task exactly corresponds to the contextualized multi-armed bandit problem described above, where different planets represent different arms and different elements represent the context. This means that a good strategy would involve a trade-off between learning the 4 different functions describing how the elements influence each planet and then maximizing the expected outcome by choosing the right planet (arm) at a given time and context. A screenshot can be seen in Figure 1.

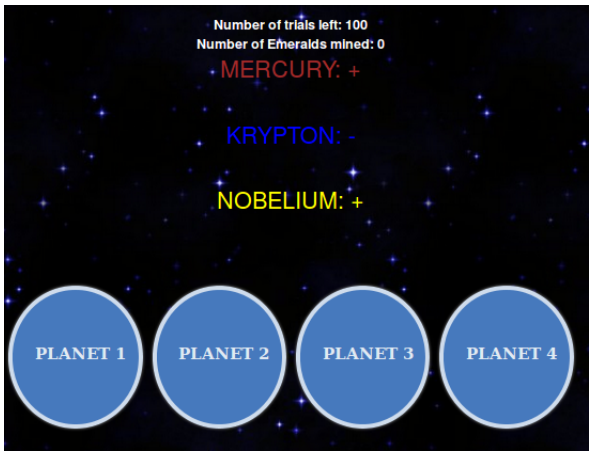

Figure 1: Screenshot of the Experiment

Which planet corresponded to which of the pay-off functions described above was assigned randomly before the start of the experiment.

Analysis All models were fitted by maximum likelihood. We assessed the ability for each of the 6 models to predict participants' choices over all trials and calculated Akaike's "An Information Criterion" (AIC) by finding the best inverse temperature parameter γ through a combination of golden section search and successive parabolic interpolation provided by the R-function `optimize` for all continuous outcomes (the UCB and the μ-tracker) or by using the estimated probabilities directly (for Thompson sampling). The AIC here is based on the log-likelihood of the predicted probabilities for eah chosen arm over all trials.

Hypotheses

Based on our conjectures above, we hypothesized the following 3 findings *a priori*:

[1] Situations only containing - or + were not used

[2] Search for Eric Schulz on Turkopticon

1. Participants will be able to learn how the context depends on the outcomes and therefore will be generally better described by contextual than by context-blind models.
2. Instead of one particular parametric strategy, participants will approach the problem in a non-parametric way allowing them to potentially learn different types of functions, if need arose. Therefore, participants will be better described by the Gaussian Process than by the linear model.
3. Instead of maximizing output by a deliberate mean-variance trade-off, participants approach dynamic decision making problems by utilizing a probability matching heuristic. Thus, they will be better described by the Thompson sampling choice rule than by the Upper Confidence Band approach.

Results

Figure 2 shows the raw data for each participant over all 150 trials.

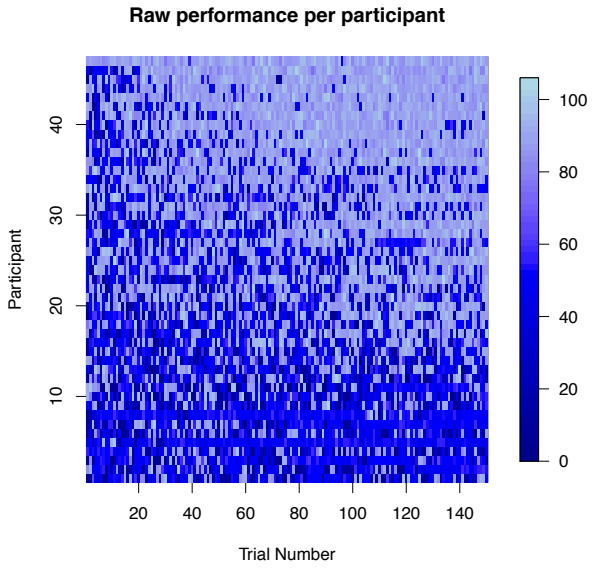

Figure 2: Obtained payoff for each participant at each trial.

In Figure 2, participants are ordered in ascending order according to their mean overall performance. It can be seen that almost all participants received higher payoffs towards the end. Moreover, some participants (the top half) seem to learn the functions very well and then consistently produced high scores over time. In the lower half, however, there are a few participants who do not seem to learn the functions too well.

Most participants also performed better than chance (an average score of higher than 50) as is displayed in the histogram of average rewards per participant shown in Figure 3.
Indeed, performing a simple t-test against $\mu = 50$ confirmed that most participants performed above chance with $t(46) = 7.17$, $p < 0.01$.

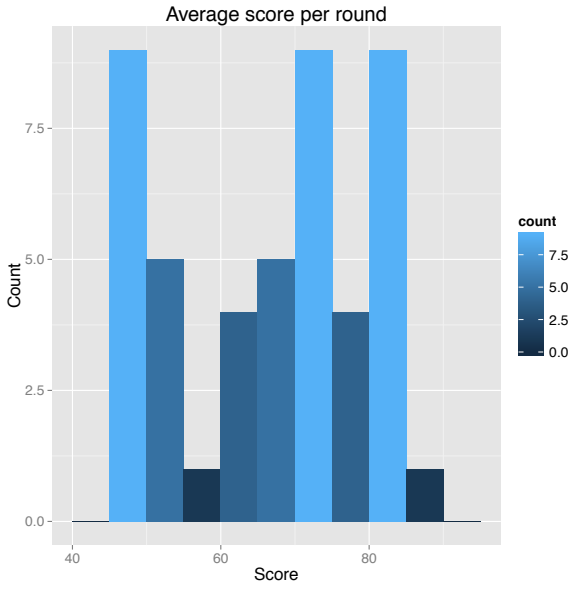

Figure 3: Average payoff per round. More participants per count are marked by a lighter blue.

Even though some participants performed below chance, we did not exclude any of them from the analysis described next as we did not want to bias our results in favor of the contextual models. The overall performance of all models is shown in Table 2.

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	5
μ-tracking	412.9	5	6
Linear-UCB	387.8	34	4
Linear-Thompson	383.0	46	15
GP-UCB	389.4	34	3
GP-Thompson	381.6*	42	18*

The 5 participants that were best described by the Random model were also among the participants who performed at chance level as shown in Figure 3.
It can clearly be seen that the contextual models described participants behavior better than the two context-blind models. Taken together, only 7 participants were best described by the context-blind models, whereas 40 participants were best described by the contextual models.

The Gaussian Process models described more participants best than the linear regression models (21 vs. 19). Even though this is only a small difference, 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 the probability matching algorithm of the Thompson sampler than by the expectation-variance-trade-off calculation of the UCB (33 vs. 7). This indicates that participants seem to apply this heuristic. Probability matching has been described as rather dumb in the past. However, in situations where the goal is to trade-off between exploration and exploitation, this heuristic is actually a smart strategy as it keeps exploring while at the same time generating high outcomes (Agrawal & Goyal, 2012).

That participants actually do learn over time while also sticking to some exploratory behavior can be see in Figure 4.

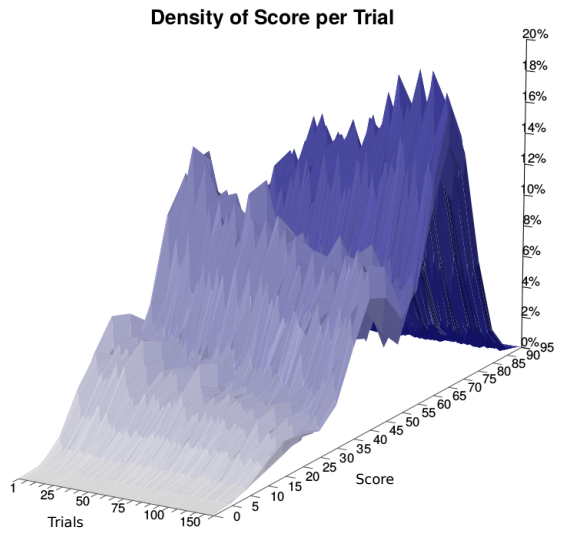

Figure 4: Density of outcome over participants per round.

As participants learn over time, the density for higher scores goes up and the density for lower scores goes down.

Discussion and Conclusion

We have introduced the Contextual Multi-Armed Bandit task as a new paradigm to assess participants' decision making in uncertain environments. Within this task, participants were able to learn the underlying structure well and took the provided context into account. Overall, most participants performed above chance and were best described by a GP-based Thompson sampling algorithm. That participants were best described by a Gaussian Process seems to suggest that – instead of having one specific parametrized representation of the environment– people learn by the means of general effective strategies that can potentially adapt to new or changing environments if required. However, future studies will have to replicate this findings in other domains. The good performance of the Thompson sampler fits well into past findings as Speekenbrink & Konstantinidis (2014) found that Thompson sampling predicts participants' choices well in a restless bandit task. Moreover, this means that probability matching, a behavior that used to be frowned upon as irrational, provides a sensitive strategy that people might actually apply in exploration-exploitation scenarios. In conclusion, all of our three main hypotheses were confirmed. This research can only be seen as a first step into research on contextual bandit problems. Future studies could try to assess how people behave in scenarios where more context is given 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 option could be to assess how participants learn in a multi-context-multi-function environment, that is an environment where the different contexts relate to arms in different ways. As we have found that Thompson sampling can provide a good description of participants' behavior and Thompson sampling is known to be well-adapted towards dynamically changing environments, a future experiments could try to model participants' behavior in dynamic tasks, where the reward structure changes over time or with the number of times a given option has been chosen.

Here, we have introduced a comparison between a linear model and Gaussian process in what can essentially be described as an active learning task. However, in future experiments we aim to try and compare even more elaborate models within this context. Using an active learning domain as a platform for model comparison might be another useful approach to decide among models from a list of seemingly endless contestants (Schulz et al., 2014).

Acknowledgements

ES is supported by the UK Centre for Doctoral Training in Financial Computing & Analytics. Data sets and code are available 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*.

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.

Schulz, E., Speekenbrink, M., & Shanks, D. R. (2014). Predict choice: A comparison of 21 mathematical models. Cognitive Science Society.

Speekenbrink, M., & Konstantinidis, E. (2014). Uncertainty and exploration in a restless bandit task.

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*.

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.