Approximation of Thompson Sampling for Bandit Task Using Nengo

Sixuan Chen (s743chen@uwaterloo.ca)
Department of Psychology, 200 University Ave
Waterloo, ON N2L 3G1 Canada

Abstract

How human balance between exploration and exploitation in a world full of uncertainties is a question studied by many cognitive science researchers using the Bandit Task. Many algorithms has been proposed to explain human's response in this task, including e-greedy, Upper Confidence Bound, Thompson Sampling(Fixed Belief Model) and Dynamic Belief Model. However, few of them consider biological constraints, making them less applicable to explaining human's behavior. In this paper, we discussed how a spiking neuron model could approximate the Thompson Sampling algorithm for the two armed Bandit Task. This model is constructed by the neural engineering framework (Nengo) which can make choice based on its estimation of mean and variance of Bandit reward probability distribution that respects known neurophysiology and neuroanatomy constraints. The results showed that spiking neuron model have the ability to estimate the alpha value of probability distribution with a small root mean squared error less than 0.005, beta value of the probability distribution with root mean squared errors less than 0.5, and potentially make a decision based on those knowledge.

Keywords: Nengo, Exploration-Exploitation, Bandit Task, Probability Estimation

Introduction

The explore-exploit dilemma has been studied using multi-arm Bandit task in both machine learning literature and cognitive science literature. In a standard Bandit task, participants will face multiple bandits which can give them certain amount of reward. Given a limited number of trials, participants need to maximize the reward they obtained. A large menu of algorithm has been developed to explain human's response in this task, including e-greedy, Upper Confidence Bound, Thompson Sampling(Fixed Belief Model), Dynamic Belief Model (Gershman, 2018; Guo & Angela, 2018; Zhang & Angela, 2013).

Many models claim that humans either adopt simplistic policies that requires little amount or memory (e.g. win-stay-lose-shift and e-greedy), or switch between an exploration and exploitation mode either randomly (Daw, O'doherty, Dayan, Seymour, & Dolan, 2006), or using a hybrid model of Thompson sampling and Upper Confidence Bound (Gershman, 2018). However, few of them consider the biological constraints that limit human's computation. For example, could human being remember each amount of reward they received in the past trials? What is the range of reward values that the neuron population could represent, and how complex can the computation in the brain be?

In this paper, we investigated how spiking neuron model could approximate the Thompson Sampling for the two armed Bandit Task. The task involves different aspects of decision making such as representation, valuation, and action selection (Rangel, Camerer, & Montague, 2008). Given inputs generated similar to human behavior task done by (Gershman, 2018), we studied those three aspects of decision making used the neural engineering framework (Nengo) model. Our spiking neuron model make choice based on its estimation of mean and variance of Bandit reward probability distribution that respects known neurophysiology and neuroanatomy constraints. The results demonstrated that spiking neuron model have the ability to estimate the alpha value of probability distribution with a small root mean squared error less than 0.005, beta value of the probability distribution with root mean squared errors less than 0.5. Also, we showed with the help of gamma distribution our model could potentially make a decision based on those knowledge.

Method

Human Bandit Task

In this paper, we build our model by using the same Bandit Task experiment setting as Gershman did in 2018. And we compared the model's choice with human behavior data from this study. Participants played 20 two-armed Bandits in blocks of 10 trials. Every block is independent from previous blocks. A reward points is given to participants if they choose one of the arms. Participants were instructed maximizes their total reward through choosing the arm wisely.

We compared only data from the first experiment in (Gershman, 2018) with our spiking neuron model. In this experiment, one arm produce reward with uncertainty, another always produce reward zero. Arm 1 has the mean reward $\mu(1)$ on each block. The mean was drawn from a Gaussian distribution with mean and variance $\mu_0(1) = 0$ and $\tau_0^2(1) = 10$ (thus each block had a different mean reward for arm 1). The actual mean reward of arm 1 for each trial was randomly drawn from another Gaussian distribution with the mean $\mu(1)$ and variance $\tau^2(1) = 10$. For each trial, whenever participant chose arm 2, they always received a reward of 0 ($\mu_0(2) = 0$ and $\tau_0^2(2) = 0$).

Specifically for this paper, we choose to analyze only the 338 block that arm 1 have a mean of 0, so $\mu(1) = 0$ and $\tau^2(1) = 10$.

Spiking Neurons

There are a plethora of neurons models,including Sigmoid, Tanh, ReLU, leaky integrate-and-fire (LIF) depending the amount of biological detail that is desired. We applied the leaky integrate-and-fire spiking network from Nengo that constructed by the neural engineering framework (Eliasmith & Anderson, 2003). The dynamics of the LIF neuron model are given in the following equation.

$$\frac{dV(t)}{dt} = \frac{J(t)}{C} - \frac{V(t)}{RC}$$

The voltage V changes in response to the input current J, and is dependent on the resistance R and capacitance C of the neuron. The product RC is known as the membrane time constant τ_{RC} which regulate how long it will take for the neurons to reach the spiking threshold. The constant τ_{ref} regulate how long the neuron will stay in refractory period which is a period that the neuron cannot spike again even there is input current.

The neuron used in our model are all leaky integrate-and-fire neurons with maximum firing rate uniformly distributed between 100 Hz and 200 Hz, $\tau_{\text{ref}} = 2$ms and $\tau_{RC} = 20$ms, intercepts adjusted automatically as input range change.

Representation

Since we are using spiking neurons, a population of neurons are required to represent a continuous value (Stewart, Bekolay, & Eliasmith, 2012). The neuron model have two parameters for each neuron: a fixed background input current bias and a fixed neuron gain factor α, which scales the neurons' inputs. These values are randomly chosen to produce a highly heterogeneous population of neurons.

$$J = \alpha x + bias$$

The resulting firing rate is given by function G that was shown to take the following form:

$$G[J] = \begin{cases} \frac{1}{\tau_{ref} - \tau_{RC} ln(1-\frac{1}{J})} & J > 1 \\ \\ 0 & otherwise \end{cases}$$

Using the values and functions mentioned above, Figure 1 demonstrates 16 LIF neurons that generated following the intercept and maximum rate distributions described above. Because participant's mean reaction time is 692 ms with a variance of 8042 ms, we input the each reward value x as a some constant vector lasting 500 ms. For each neuron, we randomly chose a gain vector and a background current. Denote the random gain vector by p, the input current for a neuron is given by the following equation.

$$J = \alpha(p \cdot x) + bias$$

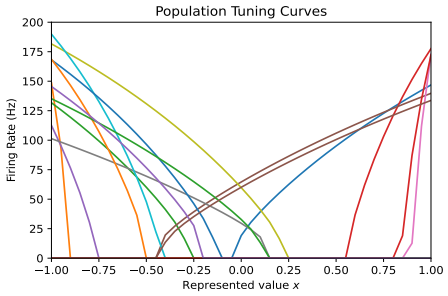

Figure 1: Population Turning Curves of LIF neurons

From on this, Nengo gives a highly distributed multidimensional state representation that produces firing patterns that match recordings from several brain areas including sensory and motor cortex (Georgopoulos, Schwartz, & Kettner, 1986). A detailed description of this representation can be found in literature written by Stewart et al (Stewart, Bekolay, & Eliasmith, 2011) and Eliasmith and Anderson (Eliasmith & Anderson, 2003).

Thompson Sampling

Thompson Sampling is different from many reinforcement learning algorithm for Bandit Task such as e-greedy, Upper Confidence Bound. It does not make a estimate of the mean reward, instead it build up a probability model from the obtained rewards, and then samples from this probability distribution to choose an action.

The advantage of using Thompson sampling is that the model not only provide accurate estimate of the the possible reward obtained, but also a level of confidence in this reward. As participants collect more samples, this confidence will also increase. Updating participants' beliefs based upon cumulative evidence is also known as Bayesian Inference. With a Gaussian likelihood, Thompson Sampling can be addressed as Fixed Belief Model which have been studied a lot by cognitive science researcher (Guo & Angela, 2018).

Participant's initial estimate is known as the prior probability and, after seeing some data, participants could adjust their estimate, forming the posterior probability. Those prior and posterior distributions can be viewed as conjugate distributions. In addition, the prior can be considered as the conjugate prior of the likelihood distribution (assumed to be the ground truth distribution of the data). For each type of conjugate prior, a set of hyper-parameters allows the prior to move towards the posterior when more and more data comes.

Specifically, for unknown mean and unknown variance reward draw from a Gaussian distribution,for each

update we have:

$$\mu_0 \to \frac{\nu\mu_0 + n\bar{x}}{\nu + n}$$
$$\nu \to \nu + n$$
$$\alpha \to \alpha + \frac{n}{2}$$
$$\beta \to \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{n\nu}{\nu + n}\frac{(\bar{x} - \mu_0)^2}{2}$$

where:
μ_0 is the estimated mean of the prior distribution.
α is the gamma's shape parameter.
β is gamma rate parameter.
$\bar{x}$ is the mean of the sample reward.
ν is the number of observations used to form the estimated, prior, mean.
n is the number of times this bandit has been used since the creation of the prior.
x_i is the reward received at each test 'i' of this bandit.

To determine the unknown variance we considered n to be the total number of samples taken right from the start of the experiment and the prior was considered to be the very initial guess. But, it's also possible to consider the prior to be the posterior that was calculated at the previous time step since we are only observing a single new reward value in the update function, thus n becomes 1. Additionally, there is only a single sample value then $\bar{x} = x$ and $\frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 0$ for each update.

To adjust the values to those a neuron population in Nengo can represent with fair accuracy, we scale down the reward by a factor of 10, therefore the variance by a factor or 100. The mathematical model of Thompson sampling with unknown mean and unknown variance for estimate the probability distribution with ground truth $\mu_0 = 0$ and $\tau_0^2 = 0.1$ give us the results shown in Figure 2:

One block of input reward value for our spiking neuron model is a some vector of 5 second (500 ms * 10 = 5000ms). Because we cannot have access to a list of ten separated x value, a vector of calculated mean values of previous trial was feed into the neuron model. Additionally, we give the neuron model a input of step function with 0.1,0.2,0.3... 1.0, each value for 500 ms as a memory of number of current trial number. And, we will scale this number up by a factor of 10 later in the computation connections between input neuron population and beta neuron population. Because of this Nengo setting, it is very hard to have access to previous alpha and beta values of the last trial. Therefore, we approximate the

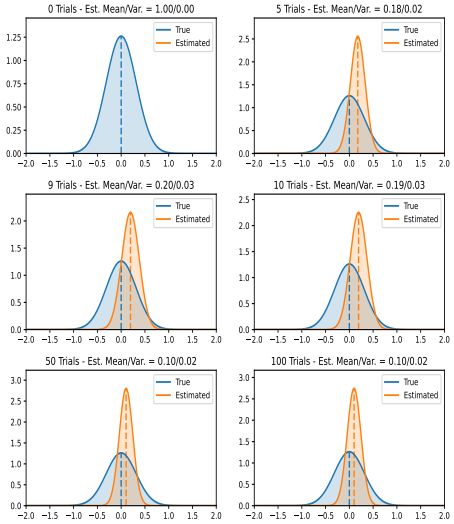

Figure 2: Original Thompson Sampling

Thompson sampling with the following adjustment:

$$\alpha \to \nu\frac{n}{2}$$
$$\beta \to \nu\frac{n\nu}{\nu + n}\frac{(\bar{x} - \mu_0)^2}{2}$$

After those adjustment, our appreciated Thompson sampling still can give a good estimate of ground truth variance shown in Figure 3.

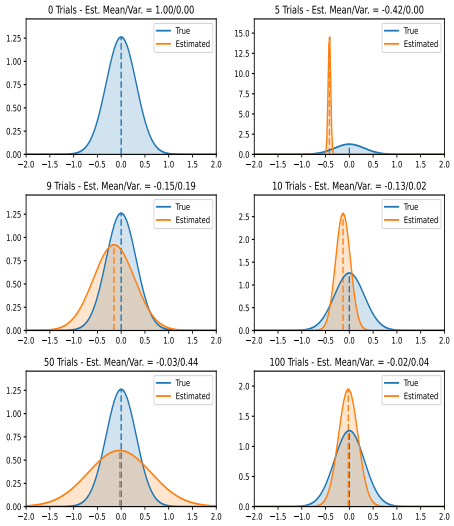

Figure 3: Approximated Thompson Sampling

Results

Nengo Spiking Neuron Model

We first build the LIF spiking neuron model that could estimate the alpha value and beta value for one block of

experiment. This model contains three neuron populations. One 2000 neuron, radius 1 ensemble represents three inputs including reward value per trial, mean reward value so far and current trial number. Another 2000 neuron radius 1 ensemble, is used to represent the alpha value which receives input of the current trial number. Finally, a neuron population represents the beta value which does a computation with those three input. Because we do not know the range of beta values, we run two experiment to find the best neuron number and radius for it.

The model was tested with 500,1000,2000,3000,4000 neuron number and radius 1 each for 20 trial. Increased number of neuron should increase the accuracy (decrease the root mean squared error) since we can average between more neuron to estimate the input. However, our results showed that root mean squared error of beta population is best at 1000 number of neuron. (Figure 4)

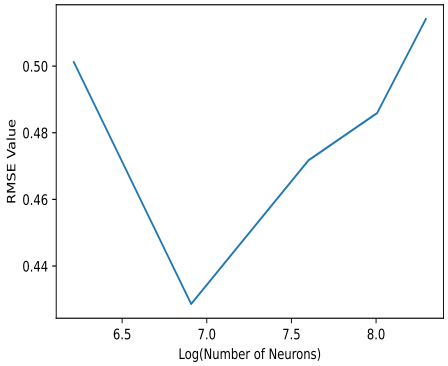

Figure 4: Changing Number of Neurons

The model was then tested with 1000 neuron number and radius 0.6,0.8,1,0,1.2,1.4 each for 20 trial. Decreased radius number should increase the accuracy (decrease the root mean squared error) as the firing rate intercept can be less sparse. However, our results showed that minimum root mean squared error of beta population is at 1 radius. (Figure 5)

Building the model with above setting, below is one block of alpha value estimated by the alpha neuron population shown in Figure 6.

The beta neuron population in the same model also give us a beta estimation shown in Figure 7.

Preceding with the above model, we add two neuron populations in to do action selection based on the probability distribution generated by those estimated alpha and beta value. One for gathering information another to make a action selection. Both of them have 2000 neurons and radius 1. Particularly, we will sample one output from the gamma distribution based on current alpha and beta value, if the value is greater than 0, we will

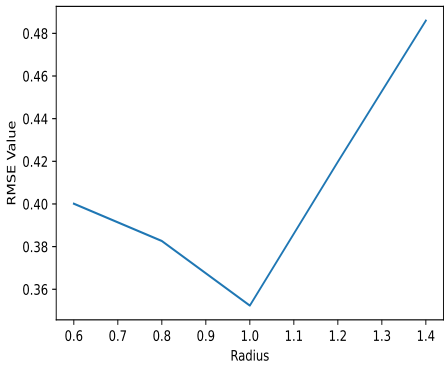

Figure 5: Changing Radius

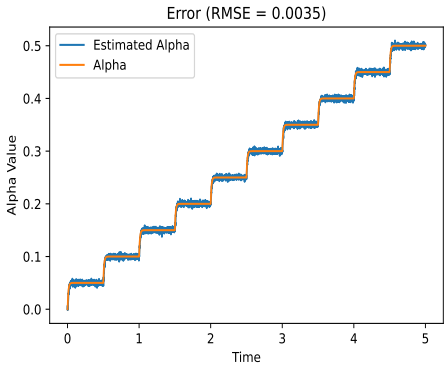

Figure 6: Alpha at Neuron Number 2000, Radius 1

choose arm 1, otherwise it will choose arm 2. With those two neuron populations added, the decision made for one block each trial is shown in Figure 8.

The results is not either 0 or 1, rather the decision values range between 0 and 1 due to the nature of LIF spiking neuron. The figure demonstrate the decision is lingering between choosing arm 1 and arm 2, the higher the decision value the more confident our spiking neuron model want to choose arm 1.

Decision of Human, Nengo and Mathematical Model For the decision values provided by spiking neuron model, we consider value above 0.5 as it choose arm 1, value below 0.5 as they choose arm 2. In order to analyze the performance of our spiking neuron model, we decided to run 338 blocks each one with 10 trials just as human participants did. Also we use a Mathematical Thompson Sampling agent to do the same task with 338 test each one 10 steps. The results is shown in Figure 9.

Discussion

The performance of Nengo is in between a mathematical Thompson Sampling model and Human behavior data, suggesting that Nengo might have adopted the biologi-

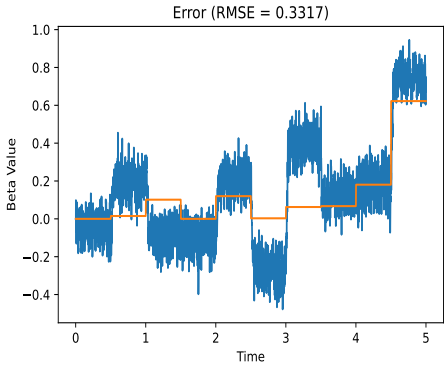

Figure 7: Beta at Neuron Number 1000, Radius 1

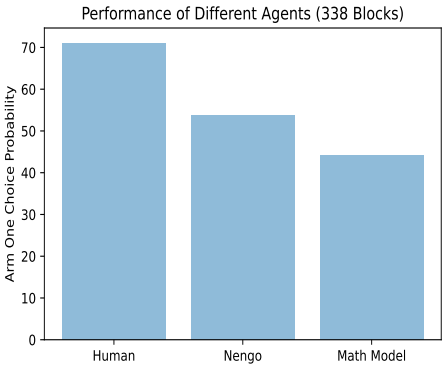

Figure 9: Choice Probability of Different Agents

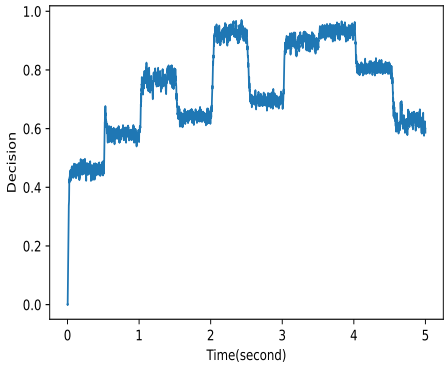

Figure 8: Example Decision Value Made by Nengo

cal limitation that human experienced such that it cannot perform as well as the mathematical model, which has all previous reward information recorded perfectly. It can only estimate the alpha value for gamma distribution with root mean square error below 0.005, but beta value with root mean square error below 0.5. Thus, it cannot estimate the optimal beta value as the mathematical model does.

In addition, Nengo did not shown a preference toward Arm 1 just as human participants demonstrated in the behavior data. This could be due to the nature of context experience human received during their daily life. Previous cognitive science research showed that human learning in the Bandit task was well captured by a Bayesian ideal learning model, the Dynamic Belief Model (DBM), in which humans assume reward rates can change over time even though they are truly stationary.(Guo & Angela, 2018) The "pessimism bias" in the bandit task is well captured by the prior mean of DBM instead of Fixed Belief Model (Thompson Sampling) discussed in here. Indeed, Thompson sampling might only be a good fit for stationary, non-contextual, uncorrelated bandit tasks, because it underestimated the un-stationary real world that human experience in daily life. Many research have demonstrated that human reinforcement learning appears to make use of sophisticated knowledge including structured knowledge about the environment (Rmus, Ritz, Hunter, Bornstein, & Shenhav, 2021) and hierarchically organized priors (Gershman & Niv, 2015) to guide exploration.

In summary, the model we have presented here provides a mechanistic model of how could we compute the beta and alpha values with LIF spiking neurons. It also showed some limitations on the accuracy of the neuron population when performing complex computations. Notice that, we only set the LIF neuron parameter such as max firing rate, τ_{ref} and τ_{RC} with nengo defaults. We did not match them to the neurologically observed values in corresponding areas of the basal ganglia, ventral striatum, or prefrontal cortex that could potentially be where those computation take place in the brain. Also, we did not arrange the connection between neuron populations to correspond to the major known connections in the basal ganglia, ventral striatum, and cortex. Maybe, by matching this value, our Nengo model could approximate human behavior better than the current one, as previous Nengo research showed by Stewart in 2012 that these neurons produce realistic heterogeneous firing patterns similar to spike pattern seen in the ventral striatum of rats performing a two-arm bandit task (Stewart et al., 2012).

Acknowledgments

This Thompson Sampling Mathematical Model code and mathematical details are referred from https://github.com/WhatIThinkAbout/BabyRobot/tree/master

References

Daw, N. D., O'doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879.

Eliasmith, C., & Anderson, C. H. (2003). *Neural engineering: Computation, representation, and dynamics in neurobiological systems*. MIT press.

Georgopoulos, A. P., Schwartz, A. B., & Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science*, *233*(4771), 1416–1419.

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

Gershman, S. J., & Niv, Y. (2015). Novelty and inductive generalization in human reinforcement learning. *Topics in cognitive science*, *7*(3), 391–415.

Guo, D., & Angela, J. Y. (2018). Why so gloomy? a bayesian explanation of human pessimism bias in the multi-armed bandit task. In *Advances in neural information processing systems* (pp. 5176–5185).

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature reviews neuroscience*, *9*(7), 545–556.

Rmus, M., Ritz, H., Hunter, L. E., Bornstein, A. M., & Shenhav, A. (2021). Humans can navigate complex graph structures acquired during latent learning. *BioRxiv*, 723072.

Stewart, T. C., Bekolay, T., & Eliasmith, C. (2011). Neural representations of compositional structures: Representing and manipulating vector spaces with spiking neurons. *Connection Science*, *23*(2), 145–153.

Stewart, T. C., Bekolay, T., & Eliasmith, C. (2012). Learning to select actions with spiking neurons in the basal ganglia. *Frontiers in neuroscience*, *6*, 2.

Zhang, S., & Angela, J. Y. (2013). Forgetful bayes and myopic planning: Human learning and decision-making in a bandit setting. In *Nips* (pp. 2607–2615).