Neural network models of decision making with learning on multiple timescales

Kiyohito Iigaya

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2014
ABSTRACT

Neural network models of decision making with learning on multiple timescales

Kiyohito Iigaya

Animals can quickly learn to make appropriate decisions according to their environment that can change over a wide range of timescales. Yet the neural computation underling the adaptive decision making is not well understood. To investigate basic computational principles and neural mechanisms, here we study simple neural network models for decision making with learning on multiple timescales, and we test our model’s predictions in experimental data. We provide basic network models for value-based decision making under uncertainty.

We first study our basic network model with plastic synapses in dynamic foraging tasks. We find that our model’s behavior can be dramatically different depending on network’s synaptic learning rules and other parameters in the environment. We also discuss a speed-accuracy trade-off. This trade-off cannot be solved by our model with a fixed rate of synaptic plasticity, or by learning
values over a fixed timescale.

One way to overcome this problem is to introduce a wide range of timescales in learning. This can be incorporated by introducing a meta-plasticity into our model, which has been developed theoretically in order to solve memory capacity problems. Guided by our model’s predictions, we find in monkeys performing a dynamic foraging task that they learn values of options over a wide range of timescales, and that they slowly tune the weights of different timescales to make their estimation more accurate. We also find that this computation causes a development of well-known deviations from the optimal behavior in monkeys. We show that this deviation (undermatching) is necessary to improve their harvesting performance within their computational constraints.

Other experiments suggest that animals can tune their timescale of leaning trial by trial. To incorporate these findings, we introduce a neural network that monitors decision making network’s performance. This surprise detection network sends a surprise signal to the decision making network to change the synaptic plasticity depending on the level of uncertainty. We show that this novel coupled network model can tune the range of timescales according to the environment near optimally and it can account for recent experimental findings. Our studies suggest that the learning values over multiple timescales can be a general principle underling adaptive neural computations and that it
can be achieved by simple neural networks with plastic synapses.
# Table of Contents

1 Introduction .......................................................... 1

2 Background ............................................................ 4

   2.1 Neural basis of decision making: from perceptual to value-based decision making. .................................................. 5

   2.2 Theoretical models of value based decision making: reinforcement learning and neural network models ................................................. 11

   2.3 Overview of the thesis ................................................. 16

3 Predicting qualitatively different matching behaviors from neural network models. .................................................. 20

   3.1 Abstract ................................................................... 20

   3.2 Introduction .............................................................. 22

   3.3 Methods and description of the model ............................... 25

      3.3.1 The task ............................................................. 25

      3.3.2 The decision making neural circuit ............................. 25
3.3.3 The plastic synapses ........................................ 26
3.3.4 Mean field analysis ........................................ 28
3.4 Results ......................................................... 34
3.4.1 The three qualitatively different behavioral regimes . . 34
3.4.2 The speed-accuracy trade off ............................... 40
3.5 Discussion ...................................................... 43
3.6 Conclusion ...................................................... 45

4 Multiple timescales of reward integration develop matching behaviors. 54
4.1 Abstract ......................................................... 55
4.2 Introduction ..................................................... 56
4.3 Results .......................................................... 59
4.3.1 The dynamic foraging task ................................. 59
4.3.2 The bias-variance tradeoff ................................. 60
4.3.3 Undermatching reflects integration over multiple timescales 65
4.4 Deviations from the matching law are accompanied by an increase of the harvesting performance .... 67
4.5 Discussion ...................................................... 68
4.6 Supplementary Material ....................................... 71
4.6.1 Neural network model that predicts changes in matching behaviors .................................................. 71
4.6.2 Non monotonic changes in the weights of timescales reflect the experimental schedule: an interplay of consolidation and forgetting. 74

4.6.3 Non-cascade type synaptic models do not account for the experimental results. 79

4.7 Methods 80

4.7.1 The task 80

4.7.2 Analysis 81

4.7.3 Network model with cascade synapses 83

4.8 Conclusion 88

5 Modeling flexible value-based decision making by neural network models. 107

5.1 Abstract 108

5.2 Introduction 109

5.3 Results 112

5.3.1 Synaptic plasticity determines the timescale of reward history integration. 112

5.3.2 The fixed plasticity models face the speed-accuracy trade-off 113

5.3.3 Changing plasticity according to the environment: the cascade model synapses and the surprise detection network 115
5.3.4 The model can self-tune the learning rates near optimally

5.3.5 The distribution of synaptic plasticity over a wide range of timescales is required when the task contains a multi-timescale structure

5.4 Discussion

5.4.1 Relation to the Bayes optimal inference models

5.4.2 Comparison to the multiple stable efficacy synaptic models

5.5 Methods

5.5.1 The decision making network with cascade type synapses

5.5.2 The surprise detection network

5.5.3 Mean field theory

5.6 Conclusion

6 Concluding remarks

Bibliography

A Probability estimation on multiple timescales

A.1 Task and constraints

A.2 Variance

A.2.1 Variance conditional on $p_0$: transient behavior

A.3 Squared bias

A.4 Average squared error, optimal $\alpha$, and undermatching
A.4.1 Time-dependent undermatching slope . . . . . . . . 174
List of Figures

3.1 Model description. A. Decision making network. Each circle represents a population of neurons. As the targets appear, the input population is activated in the same way on each trial. The input is fed through plastic synapses into two excitatory populations which represent the two possible choices. These two populations compete through an inhibitory population and work as a winner-take-all network. The plastic synapses are modified depending on the activity of the pre, post-synaptic neurons and on the outcome of the choice (reward or no reward).

B. Learning rule in rewarded trials in which the population representing the Left target is activated. The synapses to the chosen target are potentiated with a learning rate $\alpha_r$ and those to the other target are depressed with a learning rate $\gamma \alpha_r$.

C. Same as in B, but in unrewarded trials. The synapses to the chosen target are depressed $\alpha_n$ and those to the other target are potentiated $\gamma \alpha_n$. 

vii
3.2 The three different regimes. **A,B,C** Matching regime. **A** Simulated dynamics of the choice probability \( P_L \). The probability of baiting left is set to different values in different blocks of trials (gray lines). The red traces represent \( P_L \) in different runs of simulations with the same parameters. **B** Equilibrium \( P_L \) vs the fractional baiting probability \( r_L/(r_L + r_R) \). **C** The equilibrium distribution of the synaptic weights to the two choice populations in the matching regime at the point indicated by C in panel B. **D, E, F** Perseverative regime. **D** The network tends to choose always the same target, regardless the baiting probability. **E** Two stable equilibrium values for \( P_L \) (top and bottom horizontal red line). The thin red line in the middle represents unstable fixed points of \( P_L \) and separates the two basins of attractions. **F** The distributions of the synapses are now skewed in opposite directions for the two populations of synapses. **G,H,I** Tri-stable regime. **G** \( P_L \) is most of the time close to one of the three stable point (0,0.5,1). **H** For an extended range around a fractional baiting probability of 0.5, there are three stable and two unstable points. **I** The distribution of the synapses for the stable point around \( P_L = 0.5 \).
3.3 How the parameters of the neural circuit and the experimental protocol determine the behavioral regime. Each plot shows for what values of \( \alpha_n/\alpha_r \) (x-axis) and \( \gamma \) (y-axis) the network exhibits the three behavioral regimes (red region=perseverative, blue=tri-stable, green=matching). Different plots correspond to different values of the number of synaptic states \( m \) and the overall reward affluence \( r_{total} = r_L + r_R \). The behavior of the network is shown for \( r_{total} = 0.35 \) (A, D), which is the value used in experiments in which the matching behavior is observed.

As \( r_{total} \) increases, the region with the matching behavior gradually shrinks (B, E with \( r_{total} = 0.7 \) and C, F with \( r_{total} = 1 \)). Notice that the border separating the blue from the green region does not depend on \( m \).
3.4 Speed and accuracy as a function of $m$, the number of synaptic states, and of $\alpha$, the learning rate ($\alpha_n = \alpha_r = \alpha, \gamma = 0$). A Time $\tau$ required to converge to an estimate of the baiting probability vs $m$. Different curves correspond to different values of $\alpha$. $\tau(\alpha, m)$ is approximately $\sqrt{m}/\alpha$. B Standard deviation of $P_L$ vs $m$ for different values of $\alpha$. As $m$ increases, the fluctuations decrease approximately as $1/\sqrt{m}$ and the accuracy of the estimate increases. The initial fractional baiting probability is $\frac{r_L}{r_L + r_R} = 0.1$ and at time zero, it changes to $\frac{r_L}{r_L + r_R} = 0.9$. $\tau$ is estimated as the time it takes to reach $P_L = 0.5$ and the standard deviation of $P_L$ is estimated at equilibrium. The other parameters are $T = 0.05$ and $r_L + r_R = 0.35$. . . . . . . . . . 51
3.5 Optimal learning rates and number of synaptic states for environments with different volatility. The baiting probabilities change at different rates in the three plots (from left to right, the number of trials per block is $s = 10, 100, 1000$). Each plot shows the overall performance of the simulated network (color coded) as a function of the learning rate $\alpha$ ($\alpha_r = \alpha_n = \alpha$) and the number of synaptic states $m$. The performance is the harvesting efficiency, which is defined as the average number of received rewards per trial, divided by the total reward rate. The optimal parameter region is always achieved for a relatively small number of synaptic states ($m < 10$), even in the case of stable environments (right). $T = 0.05$, $\gamma = 0$ and $r_L + r_R = 0.35$
3.6 Increasing the number of synaptic states decreases performance.

A The deviation from the matching law increases as the number of synaptic states $m$ increases and it causes a decrease of the harvesting performance. B As $m$ increases, the difference $I_L - I_R$ between the total synaptic currents injected in the choice populations decreases. C, D This decrease is due to fact that the equilibrium distribution of the two synaptic populations is biased toward the same side. The synaptic current difference is due to the skewness of the distribution. As $m$ increases, the equilibrium distribution becomes progressively more localized around one of the two synaptic bounds, making the difference between $I_L$ and $I_R$ progressively smaller. This leads to an increased deviation from the matching law (A), which deteriorates the performance.
4.1 Task and matching behavior in monkeys. **a** Behavioral protocol: the animal has to fixate the central cross, and after a short delay (Delay), it can make a saccadic movement (Go) toward one of the targets. When a reward is assigned to the chosen target, a drop of water is delivered at the time of the response (Return). The overall maximum reward rate is set to 0.15 rewards per second, which is about 0.35 rewards per trial. The relative reward rates changed in blocks (100 to 200 trials) without warning; ratios of reward rates were chosen unpredictably from the set (8:1, 6:1, 3:1, 1:1). **b** Matching law behavior and deviations from it. Each point represents the blockwise choice fraction against the blockwise reward fraction. The blue line is the diagonal and it represents strict matching. The red line is a linear fit to the datapoints and it is clearly tilted with respect to the diagonal. This type of deviation is called undermatching.

4.2 The bias-variance trade-off.

---

89

90
4.3 a,b Monkeys slowly changed the weights of different timescales over multiple sessions. A model with three distinctive timescales ($\tau_1 = 2$ trials, $\tau_2 = 20$ trials, and $\tau_3 = 1000$ trials) are fitted by changing weights $w_1$ (blue), $w_2$ (red), and $w_3$ (green) with $w_1 + w_2 + w_3 = 1$ on each session. In general animals started with short timescales and weighted progressively more the integrators operating on long timescales. The opposite trend around session 160 is due to the shorting of the experimental sessions. c Deviation from the matching law slowly develops with experience. In the early sessions, monkeys show block-wise matching between choice and reward fraction; however, in later sessions the behavior showed a prominent under-matching. d The deviation from the matching law is caused by the reward integration on a long timescale. The deviation from the matching law computed over the last 50 trials of each block is plotted against the fitted value of $w_3$, the weight of the longest reward integration timescale. Both monkeys show a significant negative correlation between the slope and $w_3$. e Monkeys’ choice behavior reflected reward history over sessions. The reward bias that monkeys experienced in a previous session significantly affect the choice behavior in the current session. Both monkeys’ choice are biased by reward bias of up to around five sessions ago, which is more than 3000 trials on average.
4.4 Monkeys harvesting efficiency improved over experience and it is determined by the changes in the long timescale $w_3$. a Fraction of available rewards collected by monkeys are shown as a function of the weight of the long timescale $w_3$ for days 1 to 10 (blue) and for days 11 to 20 (red). Solid line is a quadratic fit of points. Dotted line is a model’s performance with optimal weights $w_1$ and $w_2$ given $w_3$. The simulation is done on the actual reward schedule used for experiments. Gray area indicates the standard deviation of optimal model’s performance per session. Note that both monkeys became closer to the optimal by changing the weight of long timescale $w_2$. b Monkeys trade-offed bias and variance. The slope of the matching behavior and variance of monkeys choice probability are shown to be correlated. Smaller slope means larger bias; hence the monkeys increased bias to decrease the variance of estimation.
4.5 Neural network model with synaptic plasticity on multiple timescales.

a The network model. The decision is made based on the competition between the two target populations through the inhibitory population. The competition is controlled by the synaptic strength between the input population and the target populations.

b The cascade model of synapses. Each plastic synapse between the input population and the target population can take either or two strength: potentiated or depressed, represented by the strength of 1 and 0, respectively. In addition, each synapse can take one of the metaplastic states with different rates of changing the synaptic strength. The metaplastic transition (vertical arrows) can take place in the same fashion as the normal plastic transition (horizontal and diagonal arrows) but with different probabilities. In general, the lower states are less plastic; thus they have lower probabilities of changing states. In this example, there are three levels of plasticity, and $\alpha_1 > \alpha_2 > \alpha_3$, $q_1 > q_2$. . . . . . . . . . . . . . . . 96
4.6 Synaptic plasticity on multiple timescales accounts for the data.

a Changes in matching behavior is reproduced by the network model. Assuming that the synapses occupied the states equally on day 1, slow transitions to the state with longer timescales introduce a bias toward the mean of the reward on the long timescale. This leads to an increase of deviation from the matching law, which we observed in the experiment. b The model is under the trade-off between the bias (1−slope) and the variance. The plastic synapses (at the top level) create a large variance but a small bias, while the less plastic synapses (at lower levels) induce a small variance but a large bias. As the synapses change the distribution on the plasticity space, the model trades off the bias and the variance of the estimation. This is consistent with our experimental findings.
4.7 Non monotonic changes in the weight $w_3$ of the long timescale $\tau_3$ reflect the experimental schedule on a long timescale. \textit{a,b}, The weight of long timescale $w_3$ correlates with the recent experimental lengths. Daily estimation of $w_3$ is plotted against the mean length of recent experiments. The weight of the long timescale $w_3$ is larger when the animal constantly experienced long experiments with many trials. The mean is taken over 18 experimental days (Monkey F) and 12 experimental days (Monkey G), respectively, as they give the largest correlations. \textit{c,d}, The weight of long timescale $w_3$ anti-correlates with the mean recent inter-experimental-intervals. The weight of the long timescale $w_3$ is smaller when the animal constantly had long inter-experimental-periods. Daily estimation of $w_3$ is plotted against the mean recent inter-experimental-intervals. The mean is taken over 25 experimental days (Monkey F) and 32 experimental days (Monkey G), respectively, as they give the largest magnitude of correlations.
4.8 Changes in the effect of short timescale (learning rate $\eta$, top) and the long timescale (temperature $T$, bottom) in the network model and data. The data generated by simulation of network model and the behavioral data are fitted with the same LNP model. The LNP model has two parameters: the learning rate $\eta$ capturing the reward integration of short timescale (smaller than the block size), and the temperature $T$ reflecting the effect of long time scale reward integration. Both model and data show qualitatively similar trends. That is, the reward integration timescale expands up to around session 150, after which the timescale decreases. This is due to the experimental schedule on Monkey F, and for the model it reflects the interplay between the memory consolidation and forgetting.
4.9 The local matching model with a single timescale can also capture our key findings. Although we showed that the behavior is strongly biased by a very long timescale, we can capture some aspects of the experiment by fitting with the local matching law model with one timescale. a,b The effective learning rate correlates with the slope of the choice fraction vs reward fraction. The effective learning rate can be thought of a weighted average of the multiple timescales. c,d The learning rate is larger when the animal constantly experienced long experiments with may trials. Note that the learning rate $\eta$ is approximately the inverse of the reward integration timescale $\eta = \tau^{-1}$ e,f the learning rate is smaller when the animal constantly had long inter-experimental-periods.

4.10 Non cascade type synaptic model dose not account for the changes in matching behaviors observed in the experiment. If the decision making network has synapses with two independent plasticities, the behavior should not change as opposed to the changes observed in the experiment. Hence the behavior implies the cascade type interactive plasticity model.
4.11 Changes in matching behaviors over the course of experiments.

a Early days of experiments show good matching between choice fraction and income fraction, while later days of experiments show deviations from the matching. b The deviation became prominent after the finishing the fast transition from the previous block of reward fraction, which did not present in early days of experiment. . . . . . . . . . . . . . . . . . . . . . . . . 106

5.1 Speed accuracy trade-off embedded in synaptic plasticity. The dotted line is the ideal choice probability and the colored lines are simulation results. A With less plastic synapses with \( \alpha = 0.002 \), the variance of estimation can be small; however, it becomes hard to adapt to a quick change in the environment. B On the other hand, with very plastic synapses with \( \alpha = 0.2 \), the estimation fluctuates severely, although the adaptation to a new environment is fast. . . . . . . . . . . . . . . . . . . . . . . . . 134
5.2 Model. A Decision making network. Decision is made based on the competition between two excitatory target populations through an inhibitory population. The competition leads to a winner take all process determined by the strength of plastic synapses connecting the input population to the two target populations. After each trial, those synapses are modified according to the learning rule. B Cascade model of synapses used for the decision making network. The synaptic strength is assumed to be binary (depressed and potentiated); but there are meta-plastic states associated with these strength. The transition probability of changing strength is denoted by $\alpha_i$'s, while the transition probability of changing plasticity is denoted by $p_i$'s. As synapses go down to deeper states, they become less plastic and become harder to make transitions. C Binary synapses with different plasticity used for surprise detection network. Synapses have their unique fixed transition rates between the depressed state and the potentiated state.
5.3 Cascade model of synapses with a surprise network can improve the estimation performance under the speed-accuracy trade-off.

A Cascade model of synapses can reduce the fluctuation of estimation as staying in the same context; however, this also makes difficult to learn a new environment. B With a surprise signal from the surprise detection network, the decision making network can achieve both the reduction in fluctuations and the quick adaptation to a new environment. The network parameters are taken as \( \alpha_i = 0.2^i, p_i = 0.2^i, T = 0.1, \gamma = 0 \) and \( m = 10 \), while the total baiting probability is set to 0.4 and the baiting contingency is 9 : 1 on the variable interval schedule. For B the surprise signal is sent when \( P < 0.05 \), which set the relevant transition rates \( \alpha_i = \alpha_1 \) as long as \( P < 0.05 \).
5.4 Surprise network detects a change in environment and increases plasticity in the decision making network. **A** How surprise detection works. Reward history, a sequence of 0 or 1, is integrated by synapses in the surprise signal with different plasticities (different timescales). The surprise detection network computes the averaged difference between two reward rates on different timescales, which is called expected uncertainty. The actual current difference between the two timescales of reward rates (unexpected uncertainty) is compared with the expected uncertainty and it produces a surprise signal when the unexpected uncertainty significantly exceeds the expected uncertainty. **B** As a result of the surprise signal, the decision making network increases its synaptic plasticity, or the transition rates to the more plastic states increase when the network receives the surprise signal.
5.5 Model’s effecting learning rate changes depending on the volatility of environment. A Learning rate (red) changes depending on volatility as observed in experiments. Dotted lines denote change points of contingencies. The model is simulated on a variable rate schedule and with the total reward rate 1 and the contingency is 9 : 1 or 1 : 9 with block size of 30 trials (volatile phase) and 400 trials (stable phase). The network parameters are taken as $\alpha_{ir}^i = 0.5^i$, $\alpha_{nr}^i = 0.5^{i+1}$, $p_{ir}^i = 0.5^i$, $p_{nr}^i = 0.5^{i+1}$, $T = 0.1$, and $\gamma = 1$, $m = 12$, and $h = 0.05$. B Learning rates are self-tuned to the optimal depending on the volatility of the environment. Solid line is the averaged effective learning rate in constant block size. Dotted line is the optimal single learning rate. Note that learning rate of cascade model constantly changes during a single block, which creates a slight deviation from the optimal for small and large block size; however this is an artifact of taking average of entire part of blocks The model is simulated on a VR schedule and with the total reward rate 1 and the contingency is 9 : 1 or 1 : 9 with a fixed block size. The network parameters are taken as $\alpha_{ir}^i = 0.5^{i-1}$, $\alpha_{nr}^i = 0.5^{i-1}$, $p_{ir}^i = 0.5^{i-1}$, $p_{nr}^i = 0.5^{i-1}$, $T = 0.1$, and $\gamma = 1$, $m = 20$, and $h = 0.05$ for the cascade model. For the single timescale model, $\alpha_{ir}^1 = \alpha_{nr}^1$, $\gamma = 1$ and $T = 0.1$. . . . . . . . . . . . . . . . . . . 138
5.6 Cascade model with surprise signal performs better than single timescale models in an environment that changes on multiple timescales. The task is a four-armed bandit task with blocks of 10 trials and 10000 trials on a VR schedule with the total reward rate = 1. The total number of blocks is set to 1000 : 1. In a given block, one of the targets has the reward probability of 0.8, while the others have 0.2. The network parameters are taken as $\alpha_r^i = 0.5^i$, $\alpha_{nr}^i = 0.5^i$, $p_r^i = 0.5^i$, $p_{nr}^i = 0.5^i$, $T = 0.1$, and $\gamma = 1$, $m = 4$, and $h = 0.005$ for the cascade model. The initial state is set to be the most plastic synapses, where all the synapses occupy the most plastic states. For the single timescale model, the parameters are set as $\alpha_r^1 = \alpha_{nr}^1$, $\gamma = 1$ and $T = 0.1$. . . . . 139
5.7 Distribution of synaptic plasticity is required when environment changes on multiple timescales. In this task, a subject needs to find the most rewarding target (shown here in red, although the color is not shown to the subject) from many targets. The most rewarding target with a reward probability of 0.9 (the others with 0.1) changes time to time but within the 'hot spots' indicated by yellow targets (again, the color is now shown to the subject). In this task, having a distribution on synaptic plasticity over a wide range of timescales is significantly beneficial. This is because the less plastic synapses can infer the location of hot spots, while the more plastic synapses can track the local changes of the most rewarding target. Task is on a VR schedule with blocks of 100 trials. In this simulation, the two timescale model is not the cascade model; instead the model has fixed populations of synapses with two different plasticities. The single timescale model is optimized, while the two timescales model is not: \( \alpha^1 = 0.5, \alpha^2 = 0.001, T = 0.05 \) and the synaptic population is the same for both timescales.
Acknowledgments

I’d like to thank five professors at Columbia, who supported my Ph.D. studies. Without them, I couldn’t finish my Ph.D. First, I’d like to thank Stefano Fusi for introducing me to the world of neuroscience. When I met him for the first time, I didn’t know anything about neuroscience; so he gave me a couple of papers to read: his old paper with Daniel Amit and an experimental paper by Bill Newsome’s laboratory. These were my first neuroscience papers, and at that time I could never expect that these two papers would determine my entire Ph.D. thesis, or even that I would eventually analyze the experimental data in one of the papers. It’s been quite an adventure for a physics graduate student, and it entirely owes him. I also want to thank him for taking care of my family and teaching me a lot about italian food and politics. I can’t thank enough Larry Abbott for his heart-warming support and encouragement. Without him, I couldn’t finish my Ph.D., or I couldn’t do neuroscience at all. I am so grateful that I could learn a lot from Larry and be inspired by him during my Ph.D. studies. I hope I can do what I learned from him in the future. I’d also like to thank Ken Miller, who has been always
supportive and open minded about everything. I learned a lot from Ken, who is always fascinated about science. Chapter 5 of this thesis owes his insights and fascination. Let me also thank Misha Tsodyks. He taught me how a scientist should be and how Russian food is design to match vodka shots. He made my summers very enjoyable. Lastly I’d like to thank Tomo Uemura, who has been supporting my family generously as my sponsor at the physics department with his warm heart.

Outside Columbia, I’d like to thank Bill Newsome, Leo Sugrue, Greg Corrado and Yonatan Loewenstein. Without them, I couldn’t write this thesis. I’d like to thank especially Bill for his kind heart and encouragement. Also, I’d like to thank Daeyeol Lee and Mike Kahana, who have been very generous to give me great opportunities to explore new ideas.

I’d also like to thank my friends and colleagues at the theory center. Anthony Decostanzo for teaching me all the biology and the English writing that I could not learn without him (although I still have to learn more, which is clear from this thesis). Yashar Ahmadian for giving me his great insight. The appendix of this thesis entirely owes him. Sandro Romani for always having discussion with me with fresh air. Mattia Riggotti for teaching me neuroscience and restaurants. Pablo Jercog for teaching me the real BBQ. Lorenzo Fontolan for making me laugh. It’s impossible to list everyone; but thank you all.
I’d like to thank my friends in physics department for being great friends. I still feel that the physics department is my home. Thank you Dimtri, Hung, Joon, Junichi, Carlos, Tae Hyun, Junpu, Tsuguo and others. I still remember those days when we were staying late at night at our office to solve problem sets and chat about various things in life. Looking back, it was a great time for me. I also want to thank people at the physics office: Lalla, Lydia, Joey, and Randy for helping me out throughout my graduate studies.

I’d like to thank my parents and my sister Kiyoaki, Hiromi, and Miho, for their warm support throughout my life. I thank my parents in law Hideaki and Maki for their warm heart and hospitality.

Finally, I’d like to thank Emiko and Seiji for their love and patience. Without Emiko, I simply couldn’t finish my Ph.D., and without Seiji, my Ph.D. couldn’t be so memorable. There is no way to express my feeling; I owe them my Ph.D.. Thank you so much.
To Emiko and Seiji
Chapter 1

Introduction

See the future just like the past.

Hideki Yukawa

Whether it is what to eat at a cafeteria or who to spend your life with, choosing the best option is always difficult. This is because almost everything in this world changes in time under uncertainty. Today’s best choice is not necessarily the best choice for tomorrow. There is no a priori the best decision; rather, the best decision has to change through time.

Thus making a better decision heavily relies on our ability to predict the future. How well you can predict the future, paradoxically, relies on how well you can remember the past, and often on how well you can generalize the past. Hence storing the past information is essential for making a better prediction. This may be one of the reasons why animals developed an incredible memory system. Indeed, we can store enormous amount of information in our brain
and we can recall it when it is needed [Kahana, 2012]. Moreover, our memory capacity may go beyond our own experience. Rather, the experience of past generations may be kept in our genes or be inherited as culture in our society.

Yet we are animals, consisting of living cells, which should constrain our computation for storing memory and making predictions. Understanding the constraints, I believe, is the heart of the computational neuroscience and the theoretical neuroscience, which have been pioneered in the last few decades by the people with various backgrounds including many physicists [Dayan and Abbott, 2001; Abbott, 2008]. As an example, Amit and Fusi [Amit and Fusi, 1994] discovered that the biophysically reasonable constraint of bounding synaptic efficacy, where synapses are believed to play a major role in memory function, can cause a catastrophic consequence in memory capacity of neural networks (see the following chapters for more detail). This leads to a number of studies of how animals may avoid this problem within the constraint of bounded synapses and thus provides us an opportunity to understand our unique way of storing memory (for example [Fusi et al., 2005; Fusi and Abbott, 2007; Roxin and Fusi, 2013]).

Keeping this in mind, in this theses I investigate biological ways to accumulate evidence for making better decisions under uncertainty. In other words, I try to understand how we cope with our biological constraints to produce competitive decision behaviors. Biological ways may not be the optimal for
a particular task, but it should be sufficiently good (in many cases close to optimal) for any task that we would ever face in our lives. With this spirit, in this thesis we develop computational models and also analyze experimental data. We hope that our work makes a small step, hopefully to the right direction, toward the understanding of yet to be discovered principles of adaptive decision making behaviors and the neural mechanism behind them.

The thesis is organized as follows. First, I provide some historical backgrounds. Second, I introduce a neural network model of decision making and I show that our model can reproduce various qualitatively different choice behaviors, some of which are observed in experiments. I also address how the biological model with bounded synapses can cause a problem in a typical foraging task. Third, I show how animals overcome this problem within biological constraints by analyzing experimental data and neural network models. Finally, I extend the neural network model and show that the model can reproduce and give insights of recent experimental findings of highly adaptive animal behaviors.
Chapter 2

Background

Warm the old, find a new.

Confucius

Neural basis of decision making, now often called as neuroeconomics, is the main theme of this thesis. This interdisciplinary field has just begun to be studied by an increasing number of people with various backgrounds including neuroscience, psychology, economics, electrical engineering, computer science, and physics. There are more unknowns than knowns in this new exciting field. Yet it is worthwhile to review some historical backgrounds to navigate readers to the following chapters. Due to the diversity of the field, it is impossible to review all the literature; thus I focus on some important developments made in the experimental neuroscience, followed by some theoretical models of adaptive decision making.
2.1 Neural basis of decision making: from perceptual to value-based decision making.

The study of neural basis of decision making initially focused on a perceptual decision making [Newsome et al., 1989; Salzman et al., 1990; Salzman and Newsome, 1994; Shadlen et al., 1996; Parker and Newsome, 1998; Leon and Shadlen, 1998; Gold and Shadlen, 2001; 2007], as opposed to a value-based decision making, which will be the main scope of this thesis. The former is the decision making primarily depending on current sensory evidence, whereas the latter involves the value of evidence that can change over time. This I believe is merely a historical reason, as decision making was initially studied as an extension of perceptions on one hand and motor controls on the other hand. In particular, the early studies of perceptual decision making mostly developed in the visual-oculomotor pathway. This is because it contains two of the most extensively studied systems in neuroscience: the primate visual system and the eye movement control system. As a result, researchers from the both sides in the end converged to study the same area in the middle of the sensory-motor pathway called the parietal cortex, leading to a historical debate weather this area belongs more to the sensory system or to the motor system [Robinson et al., 1978; Andersen et al., 1987; Mazzoni et al., 1996; Colby and Goldberg, 1999].
In any case, it was no doubt that the parietal cortex transforms visual information to decision information that triggers an execution of a new action. It was Shadlen and Newsome [Shadlen et al., 1996] who found a signal of accumulation of evidence strongly related to the perceptual decision in the area called lateral intraparietal area (LIP) at parietal cortex. In their experiment, a monkey sees stochastically moving dots and makes a saccadic eye movement toward the direction to which a fraction of dots is coherently moving in order to receive a reward. They found that the activities of single neurons in LIP covary with the coherence of the motion, along with the actual probability of choice. Importantly, this signal in LIP is not purely sensory or motor; rather it appeared as a signal integrated over time that is required for decision making. This finding suggests that those neurons in LIP represent the evidence used for a computation necessary for decision making [Gold and Shadlen, 2001].

While Shadlen and Newsome found a signal of accumulation of evidence in a perceptual decision task, our decision, however, is not always purely perceptual. It also depends on contexts that determine the value of actions. For example, an umbrella is very valuable to carry on a rainy day, while it is less so on a bright sunny day. Thus our decision can depend on the value of actions that changes time to time. This type of decision is called the value-based decision, which is the main focus of this thesis. It was Platt and Glimcher [Platt and Glimcher, 1999] who found LIP neurons signal the probability of reward
or the magnitude of reward that they will obtain, hence they may encode values of actions that animals plan to take. In their first experiment, a monkey is instructed to make a saccadic eye movement to the target that may result in a reward. They varied the amount of reward delivered to the monkey for each instructed response, or they also varied the probability that each possible response would be instructed. In another experiments, monkeys were free to choose one of two targets associated with different amounts of reward. In both experiments, the firing rate of some neurons in LIP during the planning period are modulated depending on the magnitude of reward that they will obtain, or the value of action that animals about to take.

These Platt and Glimcher’s findings were seminal, as they firstly related monkey’s value based free choice decisions to the activities of neurons in LIP. Yet, the puzzling but interesting part is that monkeys seemed to behave non-optimally. In their free choice task, monkeys should always choose more rewarding target to optimally collect rewards; however, monkeys rather match their rates of response to the amount of reward obtained from different choices. This is known as Herrnstein’s matching law and has been well studied in

\[1\]

This actually allowed the researchers to study the cases for choices to less rewarding target. Strictly speaking, the monkeys behaviors were overmatched, where the slope between the response fraction to a particular target and the fraction of total amount of reward expecting from target was more than 1. In this case, overmatching performs better than the behavior predicted by the pure matching law.
psychology since it was proposed by Richard Herrnstein [Herrnstein, 1961].

This apparently non-optimal behavior can be a problem for neuroscientists. Since the major goal of the studies of neural basis of decision making, around then started to be called ‘neuroeconomics’, is to find the neural signal of internal variables that are represented in the brain for decision making, it is important to design a well controlled experiment where the behaviors are predicted by experimental variables. The relationship is easily understood if the behavior is well described by the optimal strategy, as for example the Bayesian estimation theory provides variables necessary to produce the behavior. However, if one assumes that the goal of animal behavior is to maximize the reward that they obtain, observed non-optimal behavior following the matching law can be a transient phenomenon dominated by unknown variables or noise, which are hardly inferred from the behavior itself. Thus it is desirable to make a situation where the observed behavior is optimal.

Along this line, two attempts were made after Platt and Glimcher. The first is to let a monkey play a competitive game with a computer opponent, where the optimal strategy becomes inherently probabilistic. This is known as mixed strategy game in the game theory and was pioneered to apply to neurophysiological studies by Daeyoal Lee and his colleagues [Lee et al., 2004; Barraclough et al., 2004] and also by Paul Glimcher and his colleagues [Dorris and Glimcher, 2004]. In a simple mixed strategy game called matching penny
game, a monkey has to match her choice to computer’s. The optimal strategy for this game is to make a choice randomly, which monkeys eventually adapted in a response to computer’s strategy. Lee and his colleagues found that the single neuron activities of dorsolateral prefrontal cortex [Barraclough et al., 2004], anterior cingulate cortex (ACC) [Seo and Lee, 2007] and LIP [Seo et al., 2009] are found to be modulated by the values used for decisions. Dorris and Glimcher also found in a mixed strategy game called the inspection game that neurons in LIP may encode the utility of action [Dorris and Glimcher, 2004] (see also [Sugrue et al., 2005] for another interpretation.)

The other approach is to design a task in which Herrnstein’s matching law close to optimal. In fact, Herrnstein’s matching law was known to be close to the optimal behavior in a particular experimental protocol called concurrent variable interval (VI) schedule. This approach was taken by Bill Newsome’s group [Sugrue et al., 2004; Corrado et al., 2005] and Glimcher’s group [Lau and Glimcher, 2005; 2008]. Here I review this approach and key findings, as it will be the main focus of my thesis.

In [Sugrue et al., 2004; Lau and Glimcher, 2005], monkeys are free to choose either of two targets resulting in obtaining a reward or no reward. On VI schedule, a reward is given with a fixed rate rather than with a fixed probability. Thus not choosing a particular target increases the probability of obtaining a reward from the target. A new feature of this experiment in
contrast to the traditional matching law experiments in psychology is that the environment continuously changes. While the total reward rate is fixed to about 0.3 reward per trial, the ratio the reward rates from the two targets change every about 100 trials. This required animals to keep updating their estimation of values.

In this experiment, the original Herrnstein’s matching law, called global matching, is not suitable to predict the behavior, as it assumes the computation of income over the entire experimental session. Rather, they suggested that animals do 'local matching', where animals achieve the matching law on a local timescale. They found that the timescale fitted to monkeys are close to the timescale that maximizes the reward harvesting performance of their model. Also, they found that the activities of neurons in LIP are correlated with the variables called local 'fractional income', the fraction of reward obtained from a choice decided by the total amount of reward obtained from two choices.

In a subsequent analysis, however, they also found in the same experiment that the monkeys actually used multiple timescales of reward history to compute the values of actions [Corrado et al., 2005]. Interestingly, in their analysis they did not find an clear computational advantage of using multiple timescales in their task. They also showed that monkeys choice behavior was deviated from the one predicted by the matching law on local block
scales. Rather, their behaviors were close to the chance level called the under matching. If the monkeys are deviated from the matching law, and if the monkeys used multiple timescales to compute the values of actions, what is the computational goal of the monkeys and how they try to achieve? What is the biological mechanism behind the behavior? I address these questions in this thesis; but before that, I review theoretical background of learning and decision making in particular. For more backgrounds of the theoretical neuroscience and its spirit, see, for example, [Dayan and Abbott, 2001; Abbott, 2008].

2.2 Theoretical models of value based decision making: reinforcement learning and neural network models

To better understand animal behaviors and the underling neural mechanism, we need to build a model that captures the observed behavior. Then we can compare the neural signal with the variables in the model that are necessary for the computation. One popular way is to build a Bayesian inference model that optimally learns task related variables and compare those variables with neural activity (in [Ma et al., 2006] authors proposed that such a computation can be done in population of neurons). On each trial, the Bayesian learner updates
its estimation of unknown variables $V$ conditioned on the past observations $X$ by the Bayes law:

$$P(V|X) \propto P(X|V) P(V),$$

(2.1)

which is nothing but a property of probabilities. This approach is becoming very popular in computational neuroscience and systems neuroscience including the studies of decision making. One of the reasons that this approach has become popular is that animal behaviors, especially human behaviors, are often very close to the ones predicted by the optimal Bayesian learner. Thus people can assume that animals are optimized to solve the task and find correlation between task specific variables and neural activity. This approach, however, has a limitation especially when actual behaviors are not Bayes optimal (in fact, Beck et al. [Beck et al., 2012] recently argued that most computation in our brain should be sub-optimal).

Another popular way is to use learning and decision models developed in other fields such as psychology and machine learning. A notable example is reinforcement learning [Sutton and Barto, 1998; Dayan and Abbott, 2001; Dayan and Niv, 2008], which developed initially in behavioral psychology to explain a simple animal learning in a conditioning task under rewards and punishments; but now an expanding field in computer science dealing with abstract models. The main reason that this approach became so popular in neuroscience owes to Schultz et al. [Schultz et al., 1997; Dayan and Balleine,
2002; Glimcher, 2011]. In a instrumental conditioning paradigm, Schultz and his collaborators found in monkeys that the activities in dopaminergic neurons in a midbrain area called ventral tegmental area (VTA) are consistent with the critic signal required for reinforcement learning.

In particular, they found the firing rate of dopaminergic neurons is consistent with the ‘prediction error’ signal in the temporal difference (TD) learning. In TD learning, the goal of computation is to predict a discounted sum of all future rewards \( V_t \)

\[
V_t = E \left( \sum_{i=0}^{\infty} \gamma^i r_{t+i} \right),
\]

(2.2)

where \( E() \) is the expectation value, \( r_t \) is the reward at trial \( t \), usually 0 or 1, and \( \gamma \) is the discount factor. Equation (2.2) can be rewritten as

\[
V_t = E (r_t + \gamma V_{t+1}).
\]

(2.3)

Now the difference between the true value \( V_t \) and the expected value \( \hat{V}_t \), called the prediction error \( \delta_t \), is

\[
\delta_t = r_t + \gamma \hat{V}_{t+1} - \hat{V}_t,
\]

(2.4)

which is essential for TD learning and is found in activities dopaminergic neurons [Schultz et al., 1997].

The effect of dopaminergic neurons are also directly studied by direct external stimulation (brain stimulation reward, or BSR), where researchers found
CHAPTER 2. BACKGROUND

that midbrain stimulation can modulate the subject choice [Shizgal, 1997].

Now various learning and decision behaviors are suggested to be characterized by the combination of different classes of reinforcement learning that may be performed in different brain areas [Daw et al., 2005; Dayan and Niv, 2008; Daw et al., 2011], while the roles of various neuromodulators, inspired by the finding of prediction error signals, are also investigated [Daw et al., 2002; Yu and Dayan, 2005; Niv et al., 2007; Doya, 2008; Dayan, 2012].

While the reinforcement learning is powerful and tells us what kind of algorithm the brain might employ in learning and decision process, it does not specify the actual neural circuit mechanism to implement the algorithm. To understand how our brain can implement such an algorithm, we need a dynamical system described by a biophysical neural network model. Traditionally, biophysical network consisted of binary neurons and synapses were studied in associative memory, most known as Hopfield network [Hopfield, 1982; Amari, 1971]. This network is a simple spin glass type of network, where a memory is encoded as an attractor in the dynamical system, and the retrieval is achieved as an activation of specific attractor. More precisely, the network consists of neurons \( \xi_i \) (\( i = 1, 2, 3, \ldots, N \)) and synapses \( J_{i,j} \) connecting between them. Memory \( \mu \) is assumed to be an activity pattern of neurons \( \xi^\mu_i \), where \( \xi^\mu_i \) is either \(-1\) or \(1\) depending on if it is activated or not. It can be shown that if the network changes synaptic strength on each memory presentation
to the network $\xi_i^\mu$ according to $J_{i,j} \rightarrow J_{i,j} + \xi_i^\mu \xi_j^\mu$, the network’s capacity $p$ (how many patterns the network can store without completely deleting other patterns) scales with the number of synapses as [Hopfield, 1982]

$$p \sim N_{\text{syn}}.$$  \hspace{1cm} (2.5)

However, later Amit and Fusi [Amit and Fusi, 1994] found that this nice scaling breaks down with a realistic model of synapses, or if one bounds the range of synaptic strength, for example $J_{i,j} = \pm 1$. Assuming a stochastic learning in synapses, it becomes

$$p \sim \log \left( q \sqrt{N_{\text{syn}}} \right),$$  \hspace{1cm} (2.6)

where $q$ is the probability of potentiation or depression of synapses on each memory encoding event, which is related to the plasticity of synapses. This is a major focus of this thesis in the context of learning for decision making.

The idea of using attractor dynamics in neural systems was firstly applied to decision making by Xiao-Jin Wang [Wang, 2002] who showed that a simple attractor network can exhibit a decision making process that is observed in number of perceptual experiments. This biophysically implementable network consists of three groups of neurons 1) input neurons, which are activated by neurons in sensory areas (ex. looking at targets) 2) targets neurons, which are clustered corresponding to different targets. They receive inputs from input neurons and strongly excitatory connected within the same cluster. Different
cluster of neurons are connected through 3) inhibitory neurons, which try to shut down the activity of feeding neurons. This architecture allows a winner-take all process among different target neurons, leading to the decision of winning cluster of neurons. This network is very general and does not assume much about specific area of brain.

This model was extended to apply to value based decision making [Soltani and Wang, 2006; Fusi et al., 2007]. The key is to introduce plasticity into the synapses between input neurons and target neurons so that they can encode probabilistic stimulus [Rosenthal et al., 2001; Seung, 2003]. In [Soltani and Wang, 2006], they show that plastic synapses between the input neurons and target neurons can encode the values of targets, if one allows stochastic reward based Hebbian learning in those synapses (hedonistic synapses [Seung, 2003]). Synapses can take desecrate and bounded values and stochastically modified, where the probability of changing states can be related to the plasticity of synapses. In [Fusi et al., 2007], Fusi et al. found that allowing different plasticity in synapses reproduce monkeys context based decision behaviors.

2.3 Overview of the thesis

The field of neural basis of decision making, now called neuroeconomics, is still a very new field and far from its goal of understanding the neural mechanism of decision making and revealing the new principle of decision behaviors. To
fill the gap, we first need to better understand computations underling animal’s stochastic behavior. As described above, however, animal behaviors are sometimes not optimal [Beck et al., 2012] or cannot be accounted by simple fitness models [Kahneman and Tversky, 1979]. This type of behaviors was usually avoided to study in neuroscience, as the source of the non-optimality is hard to be determined.

I believe, however, that non-optimal behaviors should reflect the uniqueness of our neural system, physical constraints in our brain. Thus understanding non-optimal behavior should provide a critical understanding of the uniqueness of our brain as a whole. By doing so, we can finally find the neural variables that should be computed in our brain. Thus the main goal of this thesis is to find the right model to describe apparently sub-optimal decision making behavior in a changing world.

The thesis is organized as follows. In chapter 3, we illustrate a basic neural network model of decision making with plastic synapses, which will be extended in the following chapters. We show that the basic neural network model with simple synapses can behave qualitatively different according to the uncertainty of environment. We also reveal a speed-accuracy trade-off (or bias-variance trade-off) embedded in a simple synaptic plasticity model, which will be the main issue for the following chapters.

In chapter 4, we study how animals attempt to solve the bias-variance
trade-off in a dynamic foraging task. Previously researchers found that monkeys may track local values of actions, explained by the local matching or reinforcement learning. We show, however, this type of local model is unlikely to solve the bias-variance trade-off. Rather, we expect animals to use multiple timescales of information, as we find that combining the local and global values can avoid the bias-variance trade-off. Interestingly, the combining different timescales of values lead to an apparently sub-optimal behavior, which in fact are commonly observed and puzzled scientists for a long time. In particular, the monkeys performing a dynamic foraging task showed a deviation from the matching law [Corrado et al., 2005; Lau and Glimcher, 2005]. We show that this deviation can be caused by the properties in neural network to integrate reward information over a wide range of timescales, although previously these deviations tend to be explained as network noise [Soltani and Wang, 2006] or imperfect computation [Loewenstein, 2008]. We analyze experimental data obtained by Bill Newsome’s laboratory [Sugrue et al., 2004] to confirm our hypothesis. Also, we show that our neural network model can reproduce the experimental results of tuning the weights of different timescales of reward integration if we introduce synaptic plasticity taking place over multiple timescales.

We continue investigating our network model in chapter 5, where we focus on a trial by trial changes of synaptic plasticity to fully solve the speed-
accuracy trade-off discussed in previous chapters. Recent evidence suggests that animals dynamically modify timescales of memory they rely on. To incorporate this finding, we introduce an additional network that monitor the performance of decision making network over multiple timescales. This network sends signals to the decision making network to increase synaptic plasticity. As a result, the network can optimally adapt the plasticity and the timescale of reward integration. We also show that in a complex task, multiple timescales of reward integration will significantly benefit the decision making performance.
Chapter 3

Predicting qualitatively different matching behaviors from neural network models.

In this chapter we introduce our basic network model and the matching law, which will be more discussed in the following chapters. We find that different types of foraging behaviors can be generated from the same network model.

The work in this chapter is published in [Iigaya and Fusi, 2013].

3.1 Abstract

The matching law constitutes a quantitative description of choice behavior that is often observed in foraging tasks. According to the matching law, or-
ganisms distribute their behavior across available response alternatives in the same proportion that reinforcers are distributed across those alternatives. Recently, a few biophysically plausible neural network models have been proposed to explain the matching behavior observed in the experiments. Here we study systematically the learning dynamics of these networks while performing a matching task on the concurrent variable interval (VI) schedule. We found that the model neural network can operate in one of three qualitatively different regimes depending on the parameters that characterize the synaptic dynamics and the reward schedule: 1) a matching behavior regime, in which the probability of choosing an option is roughly proportional to the baiting fractional probability of that option 2) a perseverative regime in which the network tends to make always the same decision 3) a tri-stable regime, in which the network can either perseverate or choose the two targets randomly approximately with the same probability. Different parameters of the synaptic dynamics lead to different types of deviations from the matching law, some of which have been observed experimentally. We finally show that the performance of the network depends on the number of stable states of each synapses and that bistable synapses perform close to optimal when the proper learning rate is chosen. As our model provides a link between synaptic dynamics and qualitatively different behaviors, this work provides us with insight into the effects of neuromodulators on adaptive behaviors and psychiatric disorders.
3.2 Introduction

One of the most extensively studied foraging behaviors is known as matching behavior, where animals allocate their responses among the reward sites proportional to the relative abundance of reward at each site [Herrnstein, 1961; Herrnstein RJ, 1997]. This type of foraging behavior has been observed across a wide range of species including pigeons, rats, monkeys and humans [Herrnstein, 1961; Herrnstein RJ, 1997; Gallistel, 1994; Gallistel et al., 2001; Sugrue et al., 2004; Corrado et al., 2005; Lau and Glimcher, 2005; 2008; Rutledge et al., 2009]. Although it is widely observed, it is still unclear under which conditions the subjects follow the matching law.

Recently, several theoretical works explain the matching behavior observed in experiments [Sugrue et al., 2004; Lau and Glimcher, 2005] with different models (see e.g. [Soltani et al., 2006; Loewenstein and Seung, 2006; Loewenstein, 2008; 2010; Neiman and Loewenstein, 2013; Sakai and Fukai, 2008a; 2008b; Simen and Cohen, 2009; Katahira et al., 2012]). Here we studied systematically the dynamics of an extended version of the neural network model proposed in [Soltani et al., 2006]. The model network reproduces the matching behavior observed in experiments in which monkeys are trained to choose between two visual targets that are rewarded with different probabilities. We found that the same model can operate in qualitatively different regimes. The richness of the behavior may explain why matching is observed only under cer-
tain circumstances and it can give interesting indications on how alterations of the network parameter due to neurological disorders may affect matching behavior. The model is based on the decision making network first introduced in [Wang, 2002]. Two populations of recurrently connected excitatory neurons, which represent two decisions, compete through a population of inhibitory neurons. The network exhibits a winner-take-all behavior as only one of the two excitatory populations can win the competition. This decision model can be complemented with dynamic synapses to reproduce the matching behavior observed in experiments. The synapses that weight the inputs to the two decision populations are continuously updated depending on the outcome of the choice of the subject. Eventually the distribution of the synapses encode some estimate of the probability that a choice will be rewarded [Rosenthal et al., 2001; Soltani and Wang, 2006; Fusi et al., 2007].

The model has been shown to reproduce several interesting features of the matching behavior observed in recent experiments [Sugrue et al., 2004; Lau and Glimcher, 2005]. However, the analysis of the dynamics was usually restricted to the minimal models that could generate the behavior observed in specific experiments. Here we extend the model of [Soltani and Wang, 2006] by considering more general synapses with multiple states and a different updating rule. Our model has a rich behavior that we studied systematically with a mean field approach. We derived analytically the probability of choosing a
target, which depends in a complicated way on the reward and choice history. This probability can approximate the matching law, but it can also converge to different stable solutions that represent other dynamical regimes. Moreover, we considered a more general graded synaptic model with $m$ states and hard boundaries. This allows us to predict the effects on the behavior that some neuromodulators may have when they change the amplitude of the synaptic modifications (the synaptic changes are proportional to $1/m$). We finally studied how the learning rates can affect the performance of the network when it performs a dynamical foraging task, in which the probability of reward changes with a certain frequency. Fast synapses are obviously good at adapting at new environments, but bad at generating accurate estimates of the probability of reward. Slow synapses are bad at adapting and good at integrating reward and choice history. As the number of synaptic states increases, the synapses become slower and, although the integration time increases, the performance can deteriorate, even when the obvious negative effects of slow transients are not considered. As a consequence, the optimal harvesting performance can be achieved with synapses with a relatively small number of stable states.
3.3 Methods and description of the model

3.3.1 The task

On each trial, the subject selects a left (L) or a right (R) target. The selection leads to an outcome that depends on the reward schedule. In our study we considered the discretized concurrent variable interval (VI) schedule. Each target is either in a baited or in an empty state. If the subject then selects a baited target, it receives a reward and the state of the chosen target returns to empty. Otherwise the subject does not receive any reward. In this case, if the other target is baited, it then remains baited until the subject selects it. Before the beginning of each trial, each target is baited with some probability ($r_L$ for the left target, $r_R$ for the right target). The state of target (baited or empty) is not known to the subject. The VI reward schedule is designed to encourage the subject to “explore” and sample also the target that is baited with a lower probability. The optimal strategy in a stationary environment is to follow the matching law [Sakai and Fukai, 2008a].

3.3.2 The decision making neural circuit

The neural circuit that operates the selection is basically the decision making network proposed in [Wang, 2002; Soltani and Wang, 2006; Fusi et al., 2007] and illustrated in Fig. 3.1. An input, activated at the beginning of each
trial, is weighted and then injected into two populations of excitatory neurons that represent the two choices. These populations excite themselves through recurrent connections and compete through a mutual inhibition mediated by a population of inhibitory cells. The synaptic couplings are chosen so that there are only two stable patterns of activity in the presence of the external input. These two patterns correspond to the two possible decisions (selecting the Left or the Right target). As the neurons receive also a background noisy input, the decision is probabilistic. The probability of choosing the left target $P_L$ depends on the difference between the synaptic input currents $I_L - I_R$ to the two decision populations and it is well fitted by a sigmoid ([Soltani and Wang, 2006]):

$$P_L = 1 - P_R = \frac{1}{e^{-\frac{I_L - I_R}{T}} + 1},$$

(3.1)

where $T$ is a parameter called temperature that depends on the amplitude of the noise.

### 3.3.3 The plastic synapses

The afferent currents $I_L$ and $I_R$ are proportional to the average synaptic weights that connect the population of neurons representing the input and the two decision populations. The decision bias can be changed by modifying the efficacy of these synapses [Soltani and Wang, 2006; Fusi et al., 2007]. The current to a neuron that belongs to the decision of selecting the left target can
be written as:

\[ I_L = \sum_{j=1}^{N} w_j^L \nu_j \] (3.2)

where the \( \nu_j \)'s are the firing rates of the \( N \) neurons in the input population. We now focus on the left target, but analogous expressions can be written for the population representing the right target. Assuming that the input population is uniform \( \nu_j = \nu \), we can simplify the expression of the current:

\[ I_L = \sum_{j=1}^{N} w_j^L \nu = \nu N \langle w \rangle_L \] (3.3)

where \( \langle w \rangle_L \) is the average synaptic weight to the left population. Here we can assume \( \nu N = 1 \) without any loss of generality, as the choice probability \( P_L \) depends only on the ratio between the difference of currents and the temperature and hence we can reabsorb the constant \( \nu N \) in the temperature \( (T/\nu N \rightarrow T) \).

We assume that each synaptic weight can vary between 0 and 1. The synapses can be potentiated or depressed by a fixed amount \( \Delta w = \pm 1/(m-1) \), where \( m \) is the total number of stable synaptic states [Fusi and Abbott, 2007]. Bistable synapses correspond to the case in which \( m = 2 \).

At the end of each trial, the synapses are modified stochastically depending on the activity of the pre and post-synaptic neurons and on the outcome (i.e. whether the subject receives a reward or not. See Fig. 3.1 B,C). The synapses connecting the input population (always active once the targets are on the screen) to the decision population corresponding to the chosen target (active
at the end of the trial) are potentiated \((w \rightarrow w + \Delta w)\) stochastically with probability \(\alpha_r\) in case of reward, while they are depressed \((w \rightarrow w - \Delta w)\) stochastically with probability \(\alpha_n\) in case of no-reward. Conversely, the synapses between the population input and the decision population of the unchosen target are depressed with probability \(\gamma \alpha_r\) in case of reward, and potentiated with probability \(\gamma \alpha_n\) in case of no reward. Synaptic modifications that would bring the synapse outside the boundaries are ignored (hard bounds). The probabilities of modification determine the learning rate. The scheme of synaptic modifications is similar to the one proposed in [Soltani and Wang, 2006; Fusi et al., 2007] and it biases the choice toward the rewarded target by increasing (decreasing) the probability of choosing a target that has been rewarded (no-rewarded).

### 3.3.4 Mean field analysis

The average synaptic current to a decision population (say Left) can be rewritten as:

\[
I_L = \langle w \rangle_L = \sum_{k=1}^{m} k \frac{k-1}{m-1} \rho^k_L
\]

where the sum extends over all the discrete values of the synapses. The synaptic efficacies are \(w(k) = \frac{k-1}{m-1}\) and \(\rho^k_L\) is the fraction of synapses between the input and the Left decision population whose synaptic efficacy is equal to \(w(k)\). The synaptic strength distribution \(\rho^k_L\) changes every time the synapses
are updated according to the following equation:

$$\rho^k_L(t + 1) = Z_L^k(t)\rho^l_L(t),$$  \hspace{1cm} (3.5)

where $t$ indicates time expressed in number of trials and the $Z_L(t)$ is the following matrix:

$$
\begin{pmatrix}
1 - q^+_L(t) & q^+_L(t) & 0 \\
q^+_X(t) & 1 - (q^+_L(t) + q^+_L(t)) & \vdots & 0 \\
0 & q^+_L(t) & q^+_L(t) & 0 \\
\vdots & 0 & \vdots & 1 - (q^+_L(t) + q^+_L(t)) & q^+_L(t) \\
0 & q^+_L(t) & 1 - q^+_L(t) & 1 - q^+_L(t) & \vdots \\
\end{pmatrix},
$$

where $q^+_L(t)$ and $q^+_L(s)$ are respectively the average potentiation and depression rate, which depend on the learning rules and the reward schedule. On the VI reward schedule, they can be written as:

$$q^+_L(t) = \alpha_r P_L(t)b^L_L(t) + \gamma\alpha_n P_R(t)(1 - b^R_L(t))$$

$$q^+_L(t) = \alpha_n P_L(t)(1 - b^L_L(t)) + \gamma\alpha_r P_R(t)b^R_L(t).$$  \hspace{1cm} (3.6)

where $b_L(t)$ is a binary variable which is 1 when the left target is baited on trial $t$. Unfortunately this quantity depends in a complicated way on the reward and the choice history. However, when the baiting probabilities are stationary and $P_L(t)$ changes slow enough to be replaced by its average $\bar{P}_L$ over a certain number of trials, then the expected value of $b_L(t)$ ($\bar{b}_L$) can be approximated
CHAPTER 3. PREDICTING QUALITATIVELY DIFFERENT MATCHING BEHAVIORS FROM NEURAL NETWORK MODELS.

by [Sakai and Fukai, 2008a]:

\[ \bar{b}^L \approx \frac{r_L}{1 - (1 - r_L)(1 - P_L)} \]  

(3.7)

Notice that \( \bar{b}^L \) is usually defined as the return. More generally, the return from a choice \( X \) is the total reward that has been harvested on that choice divided by the number of choices for \( X \). In contrast, the income from choice \( X \) is the total reward that has been harvested on that choice divided by the total number of choices. Using our notation, the income from the left target is \( p_L \bar{b}^L \).

We will discuss below how these quantities are encoded by the statistics of the synaptic weights. Under the approximation \( b_L(t) \sim \bar{b}^L \), the stochastic process for updating the synapses becomes a Markov process and \( Z \) is its transition matrix. This Markov process is homogeneous, as \( Z \) has lost its dependence on \( t \). The distribution of the synapses relaxes to equilibrium, whose properties can be fully characterized. Indeed, the Markov process has a unique equilibrium distribution given by [Fusi and Abbott, 2007]:

\[ \bar{\rho}_k^L = \frac{1 - \bar{q}_L^\uparrow}{1 - \left( \frac{\bar{q}_L^\uparrow}{\bar{q}_L^\downarrow} \right)^m} \left( \frac{\bar{q}_L^\downarrow}{\bar{q}_L^\uparrow} \right)^{k-1}, \]

(3.8)

where \( \bar{q}_L^\uparrow \) and \( \bar{q}_L^\downarrow \) are stationary potentiation and depression rates determined by the Eqs. (3.6,3.7). Notice that both \( \bar{q}_L^\uparrow \) and \( \bar{q}_L^\downarrow \) depend on the behavior of the subject, which is determined by the probability of choosing one of the targets (e.g. \( \bar{P}_L \)). This probability depends in turn on the full distribution of the synapses \( \bar{\rho}_L^k,\bar{\rho}_R^k \), as they determine the total synaptic currents to the two
choice population. This means that the distribution and the choice probability should be determined self-consistently by finding a solution that satisfy simultaneously Eqs. (3.1,3.4,3.6,3.7,3.8). Not surprisingly, the distribution of the synapses that converge to the population representing the choice of the left target becomes a function of the return from left target when $\gamma = 0$, as already shown in [Soltani and Wang, 2006]. Indeed, From Eq. (3.8), the synaptic distribution is a function of the ratio between potentiation and depression rate $q_L/q_r$, which, using Eq. (3.6) can be rewritten as:

$$\frac{\bar{q}^\uparrow_L}{\bar{q}^\downarrow_L} = \frac{\alpha_r}{\alpha_n} \frac{\bar{b}^L}{1 - \bar{b}^L},$$  

(3.9)

Where, as we noted above, $\bar{b}^L$ is the return from target $L$. This is expected, as in the case of $\gamma = 0$ the synapses to the population of neurons representing Left are updated only when the left target is chosen. When $\gamma > 0$, the synaptic distribution becomes also a function of $p_L$ and $p_R$, and hence it may encode the income from the two targets. Note also that Eq.3.9 shows that the dependence of the equilibrium synaptic distribution on $\alpha_r$ and $\alpha_n$ is always through the dependence on term $\alpha_r/\alpha_n$. This is true for any value of $\gamma$ (see Eqs. (3.6,3.8)), and it greatly simplifies the analysis of the dynamical regimes of the network as the independent variables are only $\alpha_r/\alpha_n$ and $\gamma$. We now need to find a self-consistent solution for $P_L$. In general it is not possible to find a closed expression for the VI schedule; however, when the noise is small (i.e. in the limit of $T \to 0$), it is possible to derive analytical expressions. In the matching
regime, the difference between the synaptic currents \( I = I_L - I_R \) should be comparable to the temperature \( T \), so that \( I/T \) is not too large or too small, which are both cases that lead to trivial solutions \( (P_L = 0 \text{ or } 1) \). This implies that, as \( T \to 0 \), the difference between the synaptic current should also decrease at the same rate \( (I = I_L - I_R = \kappa T \to 0) \).

We now consider two solvable cases. First, when there is no interaction between synaptic populations during learning \( (\gamma = 0) \). From Eqs.(3.6,3.7) we can rewrite \( I_L - I_R = \kappa T \) as

\[
(1 - r_L)(1 - r_R)\kappa T P_L^2 + (-(r_L(1 - r_R) + r_R(1 - r_L)) + \kappa T r_L(1 - r_R) - (1 - r_L)P_L + r_L(1 - r_R) - \kappa T r_L = 0. \tag{3.10}
\]

In the limit of \( T \to 0 \) this reduces to

\[
P_L = \frac{r_L(1 - r_R)}{r_L(1 - r_R) + r_R(1 - r_L)} + O(T). \tag{3.11}
\]

Thus for \( T \to 0 \), the choice probability approaches what is determined by the matching law [Herrnstein, 1961]:

\[
\bar{P}_L = \frac{r_L(1 - r_R)}{r_L(1 - r_R) + r_R(1 - r_L)}. \tag{3.12}
\]

Note that this is consistent with our finding that the synaptic distribution becomes a function of the return when \( \gamma = 0 \). In a realistic situation the noise is finite \( (T > 0) \), leading to a deviation from the matching law (undermatching),
which is often observed in experiments [Herrnstein RJ, 1997; Sugrue et al., 2004; Lau and Glimcher, 2005].

In the case $\gamma = 1$, we can obtain the asymptotic slope of the choice probability $P_L$ as a function of the fractional baiting rate. When the reward rate is small, $r_L + r_R \ll 1$, we can linearize the potentiation and depression rate as

$$
\bar{q}_L^\uparrow = \alpha_r \bar{P}_L \frac{r_L}{1,(1-r_L)(1-P_L)} + \alpha_n \bar{P}_R \frac{\bar{P}_R(1-r_R)}{1,(1-r_R)(1-P_R)}
\approx \alpha_r r_L + \alpha_n (\bar{P}_R - r_R) \tag{3.13}
$$

and

$$
\bar{q}_L^\downarrow = \alpha_r r_R + \alpha_n (\bar{P}_L - r_L). \tag{3.14}
$$

The first term in Eq. (3.13), $r_L$, represents the average rate of obtaining reward from Left. The second term, $\bar{P}_R - r_R$, represents the average rate of not obtaining reward when Right is selected [Soltani and Wang, 2006]. Following a procedure similar to the one used in the case $\gamma = 0$, we obtain:

$$
\bar{P}_L = (1 + \frac{\alpha_r}{\alpha_n}(r_L + r_R)\tilde{r}_L + \frac{1 - (1 + \alpha_r/\alpha_n)(r_L + r_R)}{2}. \tag{3.15}
$$

where $\tilde{r}_L = \frac{r_L}{r_L + r_R}$. Equation (3.15) shows that the asymptotic slope is given by $(1 + \frac{\alpha_r}{\alpha_n})(r_L + r_R)$ at $\gamma = 1$. 

CHAPTER 3. PREDICTING QUALITATIVELY DIFFERENT MATCHING BEHAVIORS FROM NEURAL NETWORK MODELS.

3.4 Results

The results of the analysis described in the Methods can be summarized as follows: 1) depending on the parameters of the synaptic dynamics and on the overall reward rate, the model neural circuit can operate in three qualitatively different regimes: one is the widely studied matching regime. The second one is a perseverative regime, where the animal repeatedly chooses the same target regardless of the reward history. The third one is a tri-stable regime, where the animal can either perseverate by choosing repeatedly only one target, or it selects randomly one of the two targets with approximately the same probability. 2) in the matching regime, slow plastic synapses lead to more accurate estimates, but take longer to adapt to environmental changes. This is a speed-accuracy trade-off that is shared by all realistic models. 3) neural circuits with graded synapses with hard bounds have a harvesting performance comparable to the simpler bistable synapses.

3.4.1 The three qualitatively different behavioral regimes

3.4.1.1 The matching regime

Previous studies have shown that the matching behavior observed in experiments can be obtained with the model circuit that we studied [Soltani and Wang, 2006]. A simulation of the network model exhibiting matching behavior is shown in Figure 3.2A. We show in Figure 3.2B that this type of
matching behavior is actually stable in the sense that for any initial condition of the synaptic weights, the model circuit converges to the matching behavior after a sufficient number of trials. More specifically, for a given fractional baiting probability (on the x-axis), the equilibrium choice probability $P_L$ (on the y-axis) is on the red line. This line is “attractive”, in the sense that the combined learning/neural dynamics converge to the point of the red line that corresponds to the baiting ratio for any initial condition. In Figure 3.2B, the matching law (Eq.(3.12)) corresponds to the green line. The neural circuit can only approximate the matching law. We say that the circuit operates in a matching regime or that it exhibits ‘matching behavior’ whenever there is only one stable point for $r_L = r_R$ (at $P_L = 0.5$). The stable solutions of the matching regime are various approximations of the matching law.

The stability of the matching regime depends on both the parameters of the reward schedule and the parameters of the neural circuit. In particular we show in Figure 3.3 that matching behavior is stable when the overall baiting rate (i.e. the sum of the baiting probabilities) is small, the noise is small ($T \ll 1$) and the synaptic modifications prevalently affect the connections to the chosen action ($\gamma \ll 1$).

In the limit case of $\gamma \to 0$ and $T \to 0$ it is possible to derive analytically the choice probability $P_L$: 
CHAPTER 3. PREDICTING QUALITATIVELY DIFFERENT MATCHING BEHAVIORS FROM NEURAL NETWORK MODELS.

\[ \bar{P}_L = \frac{r_L(1 - r_R)}{r_L(1 - r_R) + r_R(1 - r_L)}, \]  

where \( r_L \) and \( r_R \) are the baiting rates for L and R targets respectively. This expression indicates that in this limit the model approaches the matching law.

A matching behavior which deviates from matching law can still be obtained when \( \gamma \) and \( T \) are small enough. It is also instructive to consider the choice probability when the learning rate is the same for chosen and unchosen targets \((\gamma = 1)\) in the limit \( T \to 0 \):

\[ P_L = (1 + \frac{\alpha_r}{\alpha_n})(r_L + r_R)\bar{r}_L + \frac{1}{2} - \frac{1}{(1 + \alpha_r/\alpha_n)(r_L + r_R)} \]

where \( \bar{r}_L = \frac{r_L}{r_L + r_R} \). This shows that the asymptotic slope of \( P_L \) against \( \bar{r}_L \) is given by \((1 + \frac{\alpha_r}{\alpha_n})(r_L + r_R)\) when \( \gamma = 1 \). When \( \alpha_r = \alpha_n \), the model still exhibits matching behavior if the overall baiting rate is small \( r_L + r_R < 1/2 \), but the slope of the choice probability vs the baiting probability ratio is smaller than 1, (under-matching). This is consistent with the experimental observation that when the overall baiting rate is small \( r_L + r_R < 1/2 \), under-matching is observed [Sugrue et al., 2004]. Under-matching has already been studied in other models (for example, see [Loewenstein et al., 2009; Katahira et al., 2012]).

In Figure 3.2C we show the equilibrium distributions of the efficacies of the synapses to the two target populations in the case of a model synapse with 20
states. Both distributions are highly biased toward the depressed states. This is due to the fact that synaptic depression dominates for both the left and the right target populations when the overall baiting probability and $\gamma$ are small. One of the two distributions is more skewed than the other (Left), reflecting the fact that one target is more baited than the other. Note that one could also encode the differences between the baiting probabilities by having two distributions biased toward the opposite ends (one toward the depressed and the other toward the potentiated states). However, this solution would require some tuning in the case of small temperatures, as the difference between the currents to the left and to the right populations should be restricted to vary in a very limited range to have matching behavior (see e.g. [Fusi et al., 2007]). In the case of larger values of $\gamma$, the distributions can be skewed in opposite directions, but the imbalance between the potentiating and the depressing events is always very small.

### 3.4.1.2 Perseveration

Consider now the limit situation in which following reward the synaptic populations to both targets are modified ($\gamma > 0$) and in case of no-reward the synapses remain unchanged. If the model network initially selects the left target, it will keep selecting left indefinitely, as the synapses to the left population can only be strengthened. This is true whether left is the most baited target.
or not. This extreme case illustrates the behavior of the network in the perseverative regime (see Figures 3.2 D, E), in which there is a strong tendency to select repeatedly only one target. Formally, we define the perseverative regime as the one in which there are three fixed points of the learning/neural dynamics when the two choices are baited with equal probability ($r_L = r_R$). The fixed point in the middle is unstable, whereas the other two are stable (Fig. 3.2E) and correspond to the repeated selection of one of the two targets ($P_L = 0, P_L = 1$). If the temperature is large enough and the fractional baiting probability is biased toward 0 or 1, then there is some finite probability that the neural circuit switches from one choice to the other (see e.g. Fig 3.2D).

This perseverative behavior has previously been studied in matching penny games. In these tasks, simulated neural networks that are similar to the model studied here, exhibit perseveration when the random choice behavior, which is the optimal strategy, becomes unstable [Soltani et al., 2006]. This type of behavior is observed in some experiments in the early stages of learning or when the baiting rate is high (see e.g. [Lee et al., 2004]).

3.4.1.3 Tri-stability

In addition to the two regimes described above, we found another qualitatively different behavioral regime that we named tri-stable regime. In this regime (Fig. 3.2G, H), the model either selects the targets with approximately the
same probability \( (P_L \sim \frac{1}{2}) \) or it perseverates at selecting only one target \( (P_L = 0 \text{ or } P_L = 1) \). This behavior can be interpreted as matching with a coarse-grained estimation of the reward rates. The perseverative behavior is observed when the choice probability is initially close to either 0 or 1 \( (P_L \sim 0 \text{ or } P_L \sim 1) \), i.e. in all cases in which there is a strong initial bias toward one target or the other. It is also observed when the baiting rate is analogously biased \( (\text{when } \frac{r_L}{r_L + r_R} \sim 0, \text{ or } \frac{r_L}{r_L + r_R} \sim 1) \).

Formally, the tri-stable regime is characterized by 5 fixed points at \( r_L = r_R \). The two at \( P_L = 0 \text{ or } P_L = 1 \) are stable and correspond to a perseverative behavior. The one in the middle (at \( P_L = 0.5 \)) is also stable and corresponds to a stochastic behavior in which the subject selects the two choices with equal probability. The other two fixed points are unstable.

3.3 shows that the tri-stable regime is obtained in a region of the \( \gamma-\alpha_n/\alpha_r \) plane that separates the perseverative from the matching regime. As one decreases \( \alpha_n/\alpha_r \), the neural circuit switches from the matching regime to the tri-stable regime and then to the perseverative regime. Interestingly the boundary separating the tri-stable from the matching regime does not depend on the number of synaptic states. This is explained by the fact that the transition into the tri-stable regime is characterized by the appearance of two perseverative states. For these states the distributions of the synaptic efficacies depend on the drift determined by the imbalance between potentiation and depres-
sion. This drift is positive for one population (the chosen target) and negative for the other (see Figure 3.2F). These types of distributions have a very weak dependence on the number of synaptic states ([Fusi and Abbott, 2007]). Moreover, it is important to notice that these distributions with opposing drifts can be obtained only if $\gamma$ is sufficiently large.

### 3.4.2 The speed-accuracy trade off

For the VI schedule, the optimal strategy that maximizes the harvesting performance (i.e. total reward accumulate over multiple trials) relies on the ability of the subject to estimate the probability of obtaining a reward.

The accuracy of the estimate depends on the learning rate: slow synapses can integrate evidence on longer time windows, and hence are better in terms of accuracy than fast synapses. However, this is true only in stationary environment. If the environment changes, then slow synapses are disadvantaged as they take longer to adapt to new situations.

For our synaptic model, the learning rate is determined by $\alpha_r$, $\alpha_n$, $\gamma$ and the number of synaptic states, $m$. Slow learning (small $\alpha_r$, $\alpha_n$, $\gamma$ or large $m$) mean a more accurate estimate at the expense of the adaptation time. This behavior is illustrated in Figure 3.4 where we plotted the adaptation time and the accuracy of the probability estimate as a function of $m$ and the learning rates. These quantities are measured in a simulation in which a network
operating in the matching regime starts working in an environment in which
the fractional baiting probability for the left target is \( r_L/(r_L+r_R) = 0.1 \). Then
at time zero \( r_L/(r_L+r_R) \) changes to 0.9 and the network adapts to the new
environment. The adaptation time \( \tau \) is the number of trials it takes to reach
\( P_L = 0.5 \) and the standard deviation of \( P_L \) is estimated at equilibrium. The
adaptation time scales approximately like \( \tau \sim \sqrt{m/\alpha} \), where \( \alpha_R = \alpha_L = \alpha \) and
\( \gamma = 0 \). The amplitude of the fluctuations scales as \( 1/\tau \).

The optimal learning rates in general will depend on the temporal statistics
of the changes in the environment. Figure 3.5 shows the average performance
of the neural circuit on VI schedule as a function of \( m \) and \( \alpha \) for different
lengths of the blocks in which the baiting probability is kept constant. The
shorter the blocks, the more volatile is the environment. The performance
is estimated by measuring the harvesting efficiency, defined as the average
number of rewards per trial divided by the total reward rate. As expected,
the peak performance shifts toward circuits with slower learning dynamics as
the environment becomes more stable.

Interestingly, Fig. 3.5 shows that the optimal number of synaptic states \( m \)
is always close to 2. This means that increasing the complexity of the synapse
by increasing the number of synaptic states does no improve significantly the
harvesting performance. Eventually, for large enough \( m \), the performance
actually decreases. When \( m \) increases above optimal, the estimate keeps be-
coming more accurate because the fluctuations decrease. However, there are other two effects that disrupt the performance: the first one is more obvious and it is due to longer adaptation times. The second one is more subtle and it is explained in Figure 3.6. As $m$ increases, the distribution of the synaptic weights becomes more localized around one of the boundaries. This decreases the difference between the total synaptic current $I_L$ to the left population and the total synaptic current $I_R$ to the right population. As a consequence, the matching behavior shows a more prominent undermatching ($P_L$ becomes closer to 0.5 for every fractional baiting probability. This deviation from the optimal behavior leads to a decrease in the performance. When the environment is volatile, the disruptive effects of longer adaptation times dominate the decrease in the performance. However, in stable environments, undermatching is the main cause of performance degradation.

It is important to notice that 1) $\alpha$ and $m$ both affect the adaptation time in a similar way, however, the effects on the equilibrium distribution are significantly different. 2) in the case in which the subject has only to estimate probabilities (e.g. on the concurrent variable rate (VR) schedule), an increase in $m$ may lead to strong overmatching and hence it is qualitatively different from the VI schedule (see [Ostojic and Fusi, 2013]).
3.5 Discussion

We analyzed a model of a decision making neural circuit that exhibits matching behavior. The whole analysis has been performed in a matching task with discrete variable interval (VI) schedule, in which the two targets are baited with some probability. We found that the same neural circuit has three qualitatively different behaviors depending on the parameters of the synaptic dynamics and on the parameters of the reward schedule. It is already known that matching behavior can be observed only under restricted conditions. For example the total baiting rate should be sufficiently small (typically $r_L + r_R \sim 0.35$). For larger rates, our model predicts that the subject either perseverates or it chooses randomly with equal probability the two targets.

Our analysis can also predict the effects of drugs that affect the learning rates ($\alpha_r, \alpha_n, \gamma$) or change how strongly the synapses are modified at every update (when the synapse has $m$ states, the synaptic modification is proportional to $1/m$). For example dopaminergic drugs used to treat Parkinson disease increase the learning rate from positive outcomes (our $\alpha_r$) [Rutledge et al., 2009]. Patients that are treated with these drugs exhibit a lower tendency to perseverate, which, in our language, would correspond to a transition from the tri-stable regime to the matching regime. A detailed analysis of the data would be required to establish whether the observed perseveration is compatible with the behavior of our network in the tri-stable regime. If that will
be confirmed, then it will be possible to understand what parameter changes cause the perseveration in the untreated patients. This will probably require studying en extension of the model proposed here in which $\gamma$ is different for positive and negative outcomes, but the formalism will be the same.

Our models also showed, not surprisingly, that the learning rates can significantly affect the performance. It is well known that optimal learning rates vary depending on the volatility of the environment [Behrens et al., 2007; Nassar et al., 2010; 2012]. In our analysis we assumed for simplicity that the learning rates are fixed, but it is likely that they actually change dynamically to adapt more rapidly to new environments. There could be biophysical mechanisms to modify the learning rates in individual synapses (see e.g. [Fusi et al., 2005; Clopath et al., 2008]) or system level changes in which different brain areas operate concurrently on different timescales (see e.g. [Roxin and Fusi, 2013]). All these mechanisms will be investigated in future studies.

The number of synaptic states also affects the performance. Our analysis shows that the optimal performance is always achieved for a relatively small number of synaptic states. This result seems to contradict previous studies on memory, which show that synaptic complexity can greatly extend memory lifetimes without sacrificing the amount of information stored per memory [Fusi et al., 2005]. However, we need to consider that 1) the multi state synapses that we analyzed are relatively simple and they are not representative of all
types of complex synapses. On the contrary, the analyzed multistate synapses are not among the most efficient for solving a memory problem [Fusi and Abbott, 2007]; 2) we are considering a problem in which memory is an essential component as it is needed to estimate probabilities, however, our problem is inherently different from the typical benchmarks used to assess memory capacity. In these benchmarks memories are random and uncorrelated, and hence they are presented for storage only once. Then typically the memory strength decays as the synaptic distribution relaxes to equilibrium. In contrast, in a probability estimation problem the equilibrium distribution contains information about the quantity to be estimated. As a consequence, the speed of convergence to equilibrium is not the limiting factor for the performance. Instead the fluctuations around equilibrium can strongly affect the ability to estimate probabilities (see also [Ostojic and Fusi, 2013]).

3.6 Conclusion

To conclude, we showed that our model of decision making can produce qualitatively different foraging behaviors depending on the synaptic learning rules and environment. We also showed that synapses inevitably introduce a speed-accuracy trade off in a typical foraging task, which cannot be solved by simply increasing the number of efficacy states that synapses take. We need something different. The question is how animals dove this speed accuracy trade
off, which will be addressed in the following chapters.
Figure 3.1: Model description. A. Decision making network. Each circle represents a population of neurons. As the targets appear, the input population is activated in the same way on each trial. The input is fed through plastic synapses into two excitatory populations which represent the two possible choices. These two populations compete through an inhibitory population and work as a winner-take-all network. The plastic synapses are modified depending on the activity of the pre, post-synaptic neurons and on the outcome of the choice (reward or no reward). B. Learning rule in rewarded trials in which the population representing the Left target is activated. The synapses to the chosen target are potentiated with a learning rate $\alpha_r$ and those to the other target are depressed with a learning rate $\gamma \alpha_r$. C. Same as in B, but in unrewarded trials. The synapses to the chosen target are depressed $\alpha_n$ and those to the other target are potentiated $\gamma \alpha_n$. 

\[ \text{Left Target} \quad \begin{cases} \text{ex} & \text{Right Target} \\ \text{inh} & \text{Left Target} \end{cases} \]
Figure 3.2 (following page): The three different regimes. **A, B, C** Matching regime. **A** Simulated dynamics of the choice probability $P_L$. The probability of baiting left is set to different values in different blocks of trials (gray lines). The red traces represent $P_L$ in different runs of simulations with the same parameters. **B** Equilibrium $P_L$ vs the fractional baiting probability $r_L/(r_L + r_R)$. The red line represents the stable equilibrium $P_L$. The green line represents the matching law. **C** The equilibrium distribution of the synaptic weights to the two choice populations in the matching regime at the point indicated by C in panel B. **D, E, F** Perseverative regime. **D** The network tends to choose always the same target, regardless the baiting probability. The chosen target depends on the initial conditions. Occasionally, for strongly biased baiting probabilities, the network switches target (see vertical red lines). **E** Two stable equilibrium values for $P_L$ (top and bottom horizontal red line). The thin red line in the middle represents unstable fixed points of $P_L$ and separates the two basins of attractions. **F** The distributions of the synapses are now skewed in opposite directions for the two populations of synapses. **G, H, I** Tri-stable regime. **G** $P_L$ is most of the time close to one of the three stable point (0,0.5,1). **H** For an extended range around a fractional baiting probability of 0.5, there are three stable and two unstable points. **I** The distribution of the synapses for the stable point around $P_L = 0.5$. Parameters: **A** $m = 2$, $T = 0.1$, $\frac{\alpha_n}{\alpha_r} = 1$, $\gamma = 1$, $r_R + r_L = 0.35$, **B** $m = 50$, $T = 0.1$, $\frac{\alpha_n}{\alpha_r} = 0.1$, $\gamma = 0.1$, $r_R + r_L = 1$, and **C** $m = 50$, $T = 0.1$, $\frac{\alpha_n}{\alpha_r} = 0.01$, $\gamma = 1$, $r_R + r_L = 1$. 
CHAPTER 3. PREDICTING QUALITATIVELY DIFFERENT MATCHING BEHAVIORS FROM NEURAL NETWORK MODELS.
Figure 3.3: How the parameters of the neural circuit and the experimental protocol determine the behavioral regime. Each plot shows for what values of $\alpha_n/\alpha_r$ (x-axis) and $\gamma$ (y-axis) the network exhibits the three behavioral regimes (red region=perseverative, blue=tri-stable, green=matching). Different plots correspond to different values of the number of synaptic states $m$ and the overall reward affluence $r_{total} = r_L + r_R$. The behavior of the network is shown for $r_{total} = 0.35$ (A, D), which is the value used in experiments in which the matching behavior is observed. As $r_{total}$ increases, the region with the matching behavior gradually shrinks (B, E with $r_{total} = 0.7$ and C, F with $r_{total} = 1$). Notice that the border separating the blue from the green region does not depend on $m$. 
Figure 3.4: Speed and accuracy as a function of $m$, the number of synaptic states, and of $\alpha$, the learning rate ($\alpha_n = \alpha_r = \alpha, \gamma = 0$). A Time $\tau$ required to converge to an estimate of the baiting probability vs $m$. Different curves correspond to different values of $\alpha$. $\tau(\alpha, m)$ is approximately $\sqrt{m}/\alpha$.

B Standard deviation of $P_L$ vs $m$ for different values of $\alpha$. As $m$ increases, the fluctuations decrease approximately as $1/\sqrt{m}$ and the accuracy of the estimate increases. The initial fractional baiting probability is $\frac{r_L}{r_L + r_R} = 0.1$ and at time zero, it changes to $\frac{r_L}{r_L + r_R} = 0.9$. $\tau$ is estimated as the time it takes to reach $P_L = 0.5$ and the standard deviation of $P_L$ is estimated at equilibrium. The other parameters are $T = 0.05$ and $r_L + r_R = 0.35$. 
Figure 3.5: Optimal learning rates and number of synaptic states for environments with different volatility. The baiting probabilities change at different rates in the three plots (from left to right, the number of trials per block is $s = 10, 100, 1000$). Each plot shows the overall performance of the simulated network (color coded) as a function of the learning rate $\alpha$ ($\alpha_r = \alpha_n = \alpha$) and the number of synaptic states $m$. The performance is the harvesting efficiency, which is defined as the average number of received rewards per trial, divided by the total reward rate. The optimal parameter region is always achieved for a relatively small number of synaptic states ($m < 10$), even in the case of stable environments (right). $T = 0.05$, $\gamma = 0$ and $r_L + r_R = 0.35$. 

$\alpha$, $m$, $10$ trials, $100$ trials, $1000$ trials
Figure 3.6: Increasing the number of synaptic states decreases performance.

A The deviation from the matching law increases as the number of synaptic states $m$ increases and it causes a decrease of the harvesting performance. B As $m$ increases, the difference $I_L - I_R$ between the total synaptic currents injected in the choice populations decreases. C, D This decrease is due to fact that the equilibrium distribution of the two synaptic populations is biased toward the same side. The synaptic current difference is due to the skewness of the distribution. As $m$ increases, the equilibrium distribution becomes progressively more localized around one of the two synaptic bounds, making the difference between $I_L$ and $I_R$ progressively smaller. This leads to an increased deviation from the matching law (A), which deteriorates the performance.
Chapter 4

Multiple timescales of reward integration develop matching behaviors in the bias-variance trade-off.

In the previous chapter, we showed that the simple synaptic plasticity model faces a speed-accuracy trade off in a dynamic foraging task due to a limited ability of reward integration. In this chapter we show that animals solve this trade-off by having multiple timescales of reward history integration, which could be achieved by complex synapses. This work is a collaboration with William T. Newsome’s laboratory at Stanford (Greg S. Corrado, Leo P. Sugrue,
CHAPTER 4. MULTIPLE TIMESCALES OF REWARD INTEGRATION
DEVELOP MATCHING BEHAVIORS.

and William T. Newsome) and Yonatan Loewenstein at Hebrew University. All the experiments were conducted by Bill Newsome’s laboratory [Sugrue et al., 2004]. The large part of this chapter was presented at the Cosyne meeting [Igaya et al., 2013] and a manuscript for publication is in preparation.

4.1 Abstract

From neurons to behaviors, a number of evidence suggests that adaptations take place over multiple timescales according to a power law; yet its function and mechanism remain unclear. To address this issue, we study monkeys performing dynamic foraging task, and we find that the monkeys used a wide range of timescales to integrate the reward history to compute the values of options. We find that this type of computation can nicely solve the bias-variance trade-off in a probability estimation task by optimizing the weights of different timescales (how much one should rely on different timescales) according to the environment. Indeed, monkeys are found to use multiple timescales, including the ones longer than the session length to stabilize their estimation. Moreover, monkeys slowly change the relative weights of their integration timescales over the course of experimental sessions, by which they gradually improved their harvesting performance. As monkeys slowly developed a bias in their choice, which comes from the long timescale reward integration, a commonly observed deviation from the matching law called undermatching became more promi-
nent. Thus we can explain the mechanism behind the undermatching not as a failure of biological brains, but as a clever computation to improve the harvesting performance. We also show that a biophysical model of changing synaptic plasticity that was originally introduced to solve a memory problem can predict the observed changes in behaviors. Our study provides a insight of computations over multiple timescales in a general dynamic decision making task.

4.2 Introduction

The matching law constitutes a quantitative description of choice behavior that is often observed in foraging tasks. According to the matching law, organisms distribute their behavior across available response alternatives in the same proportion that reinforcers are distributed across those alternatives ([Herrnstein, 1961; Herrnstein RJ, 1997]). This type of behavior has been observed across a wide range of species including pigeons, rats, monkeys and humans [Herrnstein, 1961; Herrnstein RJ, 1997; Gallistel, 1994; Gallistel et al., 2001; Sugrue et al., 2004; Corrado et al., 2005; Lau and Glimcher, 2005; 2008; Rutledge et al., 2009]. Although the matching law provides a surprisingly accurate description of the observed behavior, choice often deviates from the strict matching. For example, one of these deviations, known as undermatching, reveals itself as response proportions that are less extreme because the
preferences are closer to indifferent than they should be. These deviations are sometimes interpreted as a failure of the animal, which could be caused by poor discrimination between alternatives [Baum, 1974], by noise in the neural mechanisms underlying decision making [Soltani et al., 2006] or by an imbalance in the learning mechanisms [Loewenstein, 2008].

Here we analyzed an experiment in which monkeys were trained to perform a dynamic foraging task [Sugrue et al., 2004] in which they had to track the changing values of alternative choices through time. The behavioral task was designed so that the optimal probabilistic strategy that maximizes cumulative reward is to follow the matching law. We show that the animal exhibit a significant deviation from the matching law in the form of undermatching. Paradoxically, we observed that this deviation becomes more prominent as the animal becomes more experienced. This deviation from optimal behavior should lead to a decreased harvesting performance, but we actually observed that the overall performance increases as the behavior deviates more strongly from the matching law.

This paradox is solved if one considers that the profitability of the alternative choices changes continuously in the task that we considered. Indeed the choice values are changed periodically in an unpredictable way, and the animal has continuously to update them by integrating the rewards received for each choice. In this non-stationary situation when we decide we should
properly weight our past experiences, possibly in a way that depends on the volatility of the environment [Behrens et al., 2007]. Indeed, if the environment is stable, it is beneficial to consider a large number of past experiences to better estimate the value of competing alternatives. The value of each choice would then be proportional to the number of experiences in which that choice was rewarded. If the environment is volatile, then we should consider only a certain number of recent experiences, as previous ones may not be informative about the current choice values. As a consequence, there is an optimal way of weighting our recent and old experiences. Old experiences, when considered, introduce a bias in our choices which, in our experiment, manifests itself in the form of undermatching. Although this is deviation from optimal behavior, one has to consider that these old experiences also decrease the fluctuations in our estimate of the choice value. The overall effect is that the harvesting performance can increase when more old experiences are taken into account, as observed in the experiment. This simple argument shows that the observed deviations from the matching law might actually be a consequence of a more sophisticated strategy which leads to an overall better harvesting performance.

One quantitative way of describing how past experiences are weighted is to make an assumption about how rewards are integrated over time. A good descriptive model, which reproduces some experimental data [Sugrue et al., 2004; Corrado et al., 2005] is to assume that the most recent experience is the most
important one, and previous ones have a weight that decreases exponentially with the age of the experience (expressed as the number of experiences occurring between the experience under consideration and the present time). The exponential decay is characterized by a timescale which determines how many past experiences are integrated to estimate the value of each choice. It was previously reported that at least two timescales are needed to describe the behavior observed in the experiment that we analyze [Corrado et al., 2005]. Here we show that there is at least a third timescale which should be introduced. This third timescale is significantly longer than the periods of time over which the environment is stable, or in other words, the periods in which the choice values are constant. Integrating over this long timescales leads to the observed undermatching and it contributes to reduce the variance of the choice value estimates, increasing the overall harvesting performance. We finally show that the monkey performance is close to optimal despite the apparent deviation from the matching law.

4.3 Results

4.3.1 The dynamic foraging task

On each trial, the monkey is free to choose between two color targets by making saccadic movements (see Figure 4.1a). Rewards are assigned to the two
colors randomly, at rates that remain constant for a certain number of trials (typically 100-200 trials). Once the reward is assigned to a target, that target is said to be baited, and the reward remains available until the associated color is chosen. This persistence of assigned rewards means that the probability of being rewarded increases with the time since a color was last chosen, and ensures that matching approximates the optimal probabilistic strategy in this task. The reward rates were periodically modified in a random and unpredictable way.

The matching law describes fairly accurately the behavior of both monkeys, as already reported in [Sugrue et al., 2004] and shown in Figure 4.1b,c where we plotted the fraction of times the monkey choose one target versus the fraction of times that target was rewarded. All datapoints are around the diagonal (blue). However, there is some clear deviation from the matching law, which becomes more evident by comparing a linear fit (red line) of the datapoints to the diagonal. This is a clear signature of undermatching as the choices of the animals tend to be closer to indifferent (choice fraction close to 0.5) than what they should be.

4.3.2 The bias-variance tradeoff

We are going to show that this deviation from matching law is actually due to the integration of reward over timescales which are significantly longer than
the block size. Over long timescales there is no imbalance between the two targets, as they are equally rewarded on average. Hence, if this integrator contributes to the decision, the choices should be closer to balanced than what they would normally do in a stationary environment. What is the computational advantage of considering the reward stream over so long timescales? How can the deviation from the matching law caused by the long timescale integrator lead to an improved harvesting performance? In this section we present a simple model that illustrates the bias-variance trade-off already mentioned in the introduction and answers these questions. This model also defines the theoretical framework for understanding the data analysis that will be presented in the next sections.

Consider the dynamic foraging task that we analyzed. One way of modeling the decision process is to integrate rewards for each choice over a certain number of trials. The mechanism is described schematically in Figure 4.2a. The two boxes represent two exponential integrators, one for each target. The top one integrates the reward stream for the green target, whereas the bottom one integrates the reward stream for the red target. The output of each integrator is called local income and it is basically an estimate of the value of a particular choice based on a certain number of experiences. The number of trials over which reward is integrated is \( \tau \), the time constant of the exponential integrator. The decision is then assumed to be the result of the comparison
between the local incomes for the two targets. Specifically, the choices are stochastic with a probability of choosing the green target that is given by:

\[ P_G = \frac{I_G}{I_G + I_R} \]

where \( I_{G/R} \) is the local income for Green/Red target.

The statistics of the decisions generated by the model clearly depend on the timescale \( \tau \). If it is short (Figure 4.2a-c), then the average choice fraction rapidly tracks the reward fraction. Every time the reward fraction changes, the behavior rapidly adapts (Figure 4.2b). However, the disadvantage of short \( \tau \)s, is that the estimated reward fraction can fluctuate wildly. This is evident in Figure 4.2b and it is shown for various reward fractions in Figure 4.2c, where we plotted the choice fraction vs the reward fraction, as in Figure 4.1b,c. In this plot the average (solid line), is very close to the diagonal, indicating that the model has a behavior that follows the matching law, but the fluctuations are very large (shaded area). The case with a very long \( \tau \), which is the other extreme situation, is illustrated in Figure 4.2d-f. In this case the integrators estimate the value of each choice over multiple blocks and, as a consequence, the local “incomes” are constant and balanced. Figure 4.2f shows that under-matching is extreme and the fluctuations are negligible.

Intermediate situations can be constructed in several ways. One possibility
is to choose an intermediate timescale $\tau$, but we will show that it is possible to do better, and that anyway models with a single timescales cannot reproduce the behavioral data observed in the experiment. A simple model that can readily incorporate the necessary timescales is schematically represented in Figure 4.2g. For each choice there are two integrators, each operating on different timescales. One has a relatively short $\tau$, which is assumed to be shorter than the typical blocks. The other one has a long $\tau$ so that it integrates the reward streams over multiple blocks. The values estimated by these integrators are then weighted ($w$ for the long timescales, and $1 - w$ for the short one) to produce the local income, which is then used as previously to generate the choices. As expected, this model has an intermediate behavior, shown in Figure 4.2h,i for a particular choice of $w$. Increasing $w$ would increase the deviation from the matching law, which is caused by a bias in the choice value estimates, but it would also decrease the fluctuations. It is natural to ask whether there is an optimal value of the weight $w$. Figures 4.2j-l show that there is a value of $w$ that maximizes the accuracy of the choice value estimates (i.e. it minimizes the mean squared error) and this value depends on the volatility of the environment (i.e. the number of trials per block). In Figure 4.2j we show the bias and the variance as a function of $w$ (please see Appendix for analytical calculation for a simple example). The variance decreases and the bias increases as the weight of the long timescale integrator increases. The
mean squared errors of the estimates vs $w$ are plotted in Figure 4.2k for two sizes of the blocks of trials in which the reward fraction was constant. Not surprisingly, for more stable environments (dashed line), the optimal weight of long timescales should be larger. Finally we show the relation between $w$ and undermatching but plotting the slope of a linear fit to the choice fraction vs reward fraction curve (red lines in Figures 4.2c,f,i) as a function of $w$. A slope close to 1 would indicate that the behavior follows the matching law. As $w$ increases the slope decreases (undermatching), causing the behavior to deviate from strict matching.

In the ideal case, $w$ should be tuned to the volatility of the environment. Any change in $w$ would be revealed by a change in the effective time constant over which the local income is estimated. This modification would be accompanied by a change in the degree of undermatching whenever the time constants of one or more integrators are sufficiently long. In the next section we will show that undermatching is in fact caused by a change in the integrative properties of the mechanisms underlying decision making. We will then show that this adaptability actually enables the animals to maximize the harvesting performance.
Undermatching reflects integration over multiple timescales

Undermatching could be due to several reasons, already discussed in the Introduction. Here we show that in the case of our experiment 1) the relative weight of the integrators change with experience and with the schedule of the experiments 2) the weights changes are correlated with the predicted change in undermatching.

In order to estimate the relative weights of integrators that operate on different timescales we need to introduce a simple model for describing the observed behavior. In the previous section, for simplicity we discussed a model with two timescales. Now we are going to introduce a model with three integrators per choice (Figure 4.3a) as we know that within each block we can reproduce the observed behavior with two timescales which are shorter than the typical block size [Corrado et al., 2005]. We then need a third timescales that is significantly longer then the block size. The first two timescales have been determined in a previous study and they are $\tau_1 = 2$ and $\tau_2 = 20$. The third one is difficult to estimate and we will assume that it is $\tau_3 = 1000$. Our analysis does not strongly depend on the exact value of $\tau_3$. The three timescales are weighted by $w_1, w_2, w_3$ respectively (only two are independent as we assume that $w_1 + w_2 + w_3 = 1$).

In Figure 4.3b we show the estimates of the three weights as a function of
the session number for two monkeys. For each session we estimated the weights by fitting the model of Figure 4.3a to the data. Although the estimates are rather noisy, it is clear that for both monkeys the weights change in time. For monkey F the relative weight of long timescales progressively increases with experience, until session 160. Then the schedule of the experiment changed and the fast timescales become rapidly dominant (see also below for a more detailed explanation). For monkey G the contribution of the fastest timescale clearly tends to decrease. For now we will not try to explain how the $w_k$ vary as the main point we intend to make is that the relative weights of the timescales vary in time and seem to reflect an adaptive process.

Given these significant variations we can now determine whether there is the predicted correlation between weights and undermatching. In Figure 4.3c,d we plotted the choice fraction vs reward fraction for two periods that comprise multiple sessions (the periods are indicated with colored bars in Figure 4.3b). Initially (block) both monkeys seem to follow the matching law more accurately than in later sessions, when they are more experienced (red). Indeed the linear fit (solid lines) clearly indicates a deviation from the diagonal that changes significantly between the two periods. This is exactly what we were expecting from the model. A more systematic analysis is shown in Figure 4.3e, where we demonstrate that undermatching is correlated with the weight $w_3$ of the integrator that operates on the longest timescale. Undermatching is again
estimated by computing the slope of the choice fraction vs reward fraction curve. This slope is plotted on the y-axis as a function of \( w_3 \). Undermatching becomes more prominent (smaller slope) as \( w_3 \) increases. This clearly indicates that at least an important component of undermatching is explained by a change in the effective timescale over which the local income is estimated.

Further evidence for integration over adaptive multiple timescales is illustrated in Figure 4.3e for both monkeys. We estimated the correlation between the choice fraction in the current session with the reward fraction in a previous session. If one of the integrators operates on timescales that are longer than a block size, then we should detect a significant correlation for at least the previous session. The figure shows that for both monkeys this is the case, as the correlation is significantly different from zero for session lags up to 4-5.

4.4 Deviations from the matching law are accompanied by an increase of the harvesting performance

Following the matching law is the optimal strategy in the stationary case. The monkeys seem to deviate from the matching law more prominently as they become more experienced and this may sound paradoxical. However, as shown with a simple model, in non-stationary environments as the one of
the experiment, these deviations may actually reflect a better strategy. This actually seems to be the case as the harvesting performance of the animals correlates with the weight of the long timescale $w_3$, and hence increases as the animal deviates more from the matching law. We illustrate this correlation in Figure 4.4a,b where we plot the harvesting performance (total received reward) vs $w_3$ for two periods. As the monkeys become more experienced, $w_3$ increases, undermatching becomes more prominent and the harvesting performance increases. The explanation of our simple model is that the fluctuations decrease when $w_3$ increases. This is in fact the case also for the behavior of the monkey, as illustrated in Figure 4.4b. The fluctuations of the choice fraction systematically decrease when $w_3$ increases and undermatching becomes more prominent.

4.5 Discussion

Deviations from the matching law are sometimes interpreted as failures due to limitations of our cognitive or perceptual systems. Some other times they are attributed to the inability of the experimentalist to access all the relevant variables. We showed that they may actually reflect a sophisticated strategy to deal with the variability and unpredictability of non-stationary environments. The decisions of the monkeys seem to be based on reward integration on multiple timescales, some of which are surprisingly long. This is certainly
compatible with previous observations on the behavior of primates [Fusi et al., 2007; Bernacchia et al., 2011] and pigeons [Aparicio and Baum, 2009]. Interestingly, this process of integration is adaptive as the timescales are tuned to the volatility of the environment.

Integrations over multiple timescales can be implemented in several ways. Although it is difficult to determine the exact mechanism we can already make some general considerations about the properties that this mechanism should have. In our model the effective time scale over which rewards are integrated is modulated by changing the relative weights of multiple integrators that operate on diverse timescales. We show in the Supplementary Material that a model with a single varying timescale would be incompatible with the data. This indicates that there must be processes operating on multiple timescales and that the final choice is the result of non-trivial interactions between these processes. This is not too surprising as there are significant computational advantages when synapses are endowed with metaplasticity [Fusi et al., 2005] or memory systems are partitioned in interacting subsystems that are responsible for preserving memories on different timescales [Roxin and Fusi, 2013]. Both these cases show that multiple timescales are important for memory consolidation, and memory is certainly a fundamental component when integration over time is required.

We used our simple model to study the effects on decision making of
changes in the relative weights of different integrators that operate on multiple timescales. We did not model explicitly the mechanism which controls and tunes these relative weights which is a difficult and interesting problem. We also mostly focused on the stationary behavior and we did not analyze transients induced by sudden changes in the statistics of the choice values. It is important to notice that the long timescales that we studied contribute to the harvesting performance on timescales that are significantly longer than the short timescales of the integrator, but still shorter than the periods over which the environment is stable. We basically ignored the effects of multiple timescales on transients. Other researchers looked specifically at the mechanism which may speed up the convergence to more accurate estimates when a change in the environment is detected [Nassar *et al.*, 2010; 2012; Wilson *et al.*, 2013]. These mechanisms are probably complementary to those that we studied, and in the future we will try to combine all these mechanisms in a unified model.

We showed that the harvesting performance of the monkey increases when the relative weights of the integrators are properly tuned. The improvement is admittedly modest, but this is due to the simplicity of the task that we analyzed. In a two-choice task the harvesting performance varies in a rather limited range when the behavior goes from random (or when the monkey responds in any other way that completely ignores the reward ratios) to optimal.
This is a well known limitation which makes it difficult to establish how close the behavior is close to optimal. In more complex tasks, which may involve multiple choices, the situation could be drastically different. Imagine for example a task in which there are 100 different choices and only one of them is rewarded with a probability that is significantly different from zero, using a schedule similar to the one of the two-choice task that we analyzed. The rewarding targets may change time to time; but within a restricted fraction of targets, say within the same 10 targets. In this case the long timescales contain important information about the possible rewarding targets that should be considered. Ignoring this information would lead to a significant decrease in the performance. We will discuss this issue more in detail in the next chapter of this thesis.

4.6 Supplementary Material

4.6.1 Neural network model that predicts changes in matching behaviors

4.6.1.1 Network model

So far we showed that monkeys used the reward history integrated over a wide range of timescales and they changed the weights of different timescales to improve their harvesting performance in the bias-variance trade-off. What
would be the possible neural mechanism behind this computation? Here we provide a possible example using synaptic plasticity model, which has been originally introduced to solve a general memory problem [Fusi et al., 2005; Fusi and Abbott, 2007]. The neural network model is already introduced in the previous chapter [Wang, 2002; Soltani and Wang, 2006; Fusi et al., 2007]. In this network, the input neurons are activated on each trial. These neurons project into two excitatory neural populations, each representing different actions (selecting the target of green or red). Those two groups of neurons inhibit with each other through a globally projecting inhibitory population. The result is a winner take all process between the two target neural populations. We assume that the winning population represents the network’s decision.

The model’s decision is heavily influenced by the reward history, as the reward information is encoded in the plastic synapses between input and target populations. After each trial, synapses will be updated according to a stochastic Hebbian learning [Seung, 2003], changing the strength and plasticity depending on the current state and transition rates. The synaptic efficacy is assumed to have one of the two strengths (weak or strong). Previously we showed that this switch like synapses can perform as good as any bounded synapses with multiple efficacy states [Iigaya and Fusi, 2013].

In addition to the changes in efficacy, here we introduce a metaplastic transition to the state with the same synaptic efficacy but with different plasticity
levels, or different rates of changing the efficacy [Fusi et al., 2005] (figure 5.2).

This model incorporates a memory consolidation process at a synaptic level with various chemical cascade processes taking place over different timescales. It is previously proved to improve network’s general memory performance as well as to reproduce a well known power low memory decay in time, which is widely observed in many experiments [Wixted and Ebbesen, 1991]. More precisely, we assume that the synaptic strength is binary and can be strong or weak as in the previous chapter; but in addition, we assume that there are multiple states with the same strength but different transition rates to the opposite strength state. Synapses make transitions to less plasticity state with a probability which becomes smaller as the synapses enter less plasticity states. For simplicity, changes in synaptic strength takes place with a full recovery of plasticity (diagonal arrows in figure 5.2).

4.6.1.2 Model results

The network model shows similar behaviors to what we found in the experimental data. As seen in figure 4.6a, the model reproduces the changes in matching behaviors observed in the experiment. In the early days of experiment, the model shows a good matching, as we assume that the synapses equally occupy all the states in the beginning of day 1. As the experiment continues, the distribution of synaptic plasticity changes and more and more
synapses occupy less plastic states. This introduces a bias toward the mean of the reward on the long timescale and increases a deviation from the matching law, which we observed in the experiment. Figure 4.6b shows that the model performs in the trade-off between bias and variance. This is because the plastic synapses create a large variance but a small bias, while the less plastic synapses (at lower levels) induce a small variance but a large bias. As synapses change the distribution on the plasticity space, the model trades-off the bias and the variance of the estimation. This is consistent with our findings in the experimental data.

### 4.6.2 Non monotonic changes in the weights of timescales reflect the experimental schedule: an interplay of consolidation and forgetting.

We showed that changes in the weights of different timescales are not monotonic for both monkeys. In particular, Monkey F shows a rapid reversal of weights around session 150. We hypothesized that this is due to the interplay of memory consolidation and forgetting during and between the experiments. More precisely, the cascade synaptic model can lead to a shift of synaptic distribution toward the less plastic states, which corresponds to the memory consolidation process of reward information. This in fact takes place trial by trial in our model, although we do not focus on the changes on the local
scale in this chapter (see the next chapter for more discussion). After the experiment, on the other hand, monkeys engage in different activities, such as walking or eating food. Since the same network should be partially involved in those activities, the synaptic plasticity in the network also should change. In particular, many synapses gain plasticity, which corresponds to the forgetting process between the experiment. (Surprise signals [Iigaya and Fusi, 2011] that will be introduced in the next chapter may also play a role to increase the synaptic plasticity between the task.) As a results, the distribution of synaptic plasticity shift toward the more plastic state between the experiments.

If this scenario is correct, monkey’s integration timescale becomes longer when she is engaged in a series of long intensive experiments. This is indeed the case as seen in figure 4.7a,b, which shows that the weight of the longest timescales of our model, \( w_3 \) is strongly correlated with the total length of recent experiments. If the monkey experiences long inter-experimental-intervals, on the other hand, the weight of the long timescale should decrease as forgetting wins consolidation. This is also the case in the data, as seen in figure 4.7c,d. Here, monkey’s weight of the longest timescale \( w_3 \) is strongly anti-correlated with the total length of the recent inter-experimental-intervals.

The overall dynamics of changes in timescales of model and data are summarized in figure 4.8. Here both the simulated data and the actual experimental data are fitted by a simple non-linear sigmoid model with a nonlinearity \( T \)
and a single reward integration timescale $\tau$, represented here by the learning rate $\eta = \tau^{-1}$. This type of model is introduced by [Corrado et al., 2005] to analyze this data. Here we chose the model since our network model has the same nonlinearity. The learning rate in this model corresponds to the changes in the weights of fast timescales that are smaller than the block size. Since the reward is balanced on a long timescale, the changes in weight of the long timescale that are longer than block size is captured by the changes in the temperature $T$ here. This is because any changes of synaptic population toward the less plastic state with the timescale on which the reward is practically balanced would be induced equally to both target populations. As a result, the signal, the difference between the synaptic population at less plastic states, would be just renormalized, corresponding to the changes in the temperature $T$ in the sigmoid model. For example, as increasing the weights of long timescale, the weight of the small timescale decreases. This is approximately the same as rescaling the input, if we ignore the effect of slight imbalance of rewards on a long timescale, which is achieved by changing $T$. Large $T$ means that there is more bias on a long timescale.

Note that this type of changes in $T$ is different from the effect of network noise, although in the mean field theory they are similar. If the noise increase, the variance of estimation should also increase. This leads to a decrease in harvesting performance. In other words, increasing noise does not compensate
the bias-variance trade-off. This is not what we found in the data. As $T$ increases, the performance also increases along with a decrease in the variance of estimation. This means that the changes in $T$ term in this experiment is the rescaling of the signal, not the increasing the noise. This is consistent with our model of multiple timescales of synaptic plasticity. As in this experiment the long timescales are well balanced hence by taking the difference of values those contribution from the long timescales vanishes. The only effect that we observe is a decrease in weights of short timescales. This is a rescaling of the signal, leading to an undermatching effect.

Previously Soltani and Wang [Soltani and Wang, 2006] proposed that the observed under matching effect can be due to the noise in the neural network. More precisely, the incoherence of activity in target neurons of each pool. This indeed predicts an undermatching effect; however, this is not consistent with the data, as increasing the noise, the variance should increase hence performance should deteriorate. The observed effect is the opposite. Although in the mean field it is the same change in the parameter, the physical meaning and the consequence is different. Still, we show that the biophysical network model with plastic synapses can capture the many aspects of reward integration and changes in behaviors newly found in the experiment. We propose that the changes in the temperature term is due to the changes in the synaptic plasticity.
Figure 4.9 shows that this analysis does not depend on a specific choice of the fitting model. Here we analyzed data with a single timescale local matching model proposed in [Sugrue et al., 2004]. The changes in matching behaviors are shown to be strongly correlated with the changes in the learning rate, as the learning rate should reflect the changes in the weights of different timescales. Also, the non-monotonic changes in the weights of different timescales is now transformed to the non-monotonic changes in the learning rate, which can be explained by the experiment schedule (session lengths and inter-experimental-intervals). Although this single timescale model does not account for the observed changes in the bias of matching, as it just slows down the learning while the data shows there is a very slow component does not adapt within the block size, this reduced model can still capture some aspects of changes in the timescale in the data.

To summarize, here we showed that the non-monotonic changes in the weights of timescales in data reflects task schedule including the length of the task and the length of inter-experimental-intervals. Our network model provides a possible biophysical mechanism behind this change. That is, the interplay of memory consolidation as a result of task experience, and the recovery of plasticity or forgetting as a result of experiences in the inter-experimental intervals.
4.6.3 Non-cascade type synaptic models do not account for the experimental results.

In the previous section we showed that the cascade type model of synapse captures some of the features of observed data, including the changes in the matching behaviors and changes in the integration timescales. There are also other possible ways to achieve multiple timescales of reward integration on plastic synapses. The most natural way is to have independent populations of synapses with different rates of plasticity. As seen figure 4.10, this model dose not reproduce experimental data. This is because there is no interaction between the two populations of synapses. The key computation found in the experiment is changing the weights of a wide range of reward integration timescales. The independently stable plasticity model, however, does not change the rate of plasticity. Hence this class of models with fixed plasticity of parallel independent reward integrations on different timescales cannot account for the observed data.

Also, it is clear from the data that the changes in the matching behavior is not because the monkeys became slow but because they obtained a bias. Figure 4.11 shows that monkey’s smoothed instantaneous choice fraction nicely follows the reward fraction on an early day; however, the choice fraction apparently biased toward the mean of the reward on a long timescale on a later day. After transients of each block of trials, the choice fraction stays at the
value that is far from the reward fraction, which shows that the undermatching effect is not due to a slow transient within a block but it comes with a bias toward the mean of the reward history that cannot be discarded within a block of trials.

4.7 Methods

4.7.1 The task

The full experiment is described in [Sugrue et al., 2004; Corrado et al., 2005]. A monkey is free to choose between two targets, where the color of each target cues the probability that its selection with an eye movement will be rewarded with a drop of juice. Analogous to natural environments, rewards in this task are assigned to the two colors at rates that are independent and stochastic (Poisson probability distribution). Once assigned, a reward remains available until the associated color is chosen (concurrent variable interval schedule). This persistence of assigned rewards means that the likelihood of being rewarded increases with the time since a color was last chosen, and ensures that matching approximates the optimal probabilistic strategy in this task. The bait probability (the rate of assigning a reward to a target) is fixed for a block of trials; but they change abruptly after about 100 trials without any cue. The total reward rate is fixed to about 0.35 reward per trial. Data is collected from
two monkeys (monkey F and monkey G).

4.7.2 Analysis

4.7.2.1 Multiple timescales local matching model

To determine the weights of different timescales, we assume that monkeys perform a local matching with the local incomes $I_G$ and $I_R$ computed by combination of multiple timescales. Local matching is

$$P_R = \frac{I_R}{I_R + I_G}$$

(4.1)

where $P_R$ is the probability of choosing the red target at the current trial. In [Sugrue et al., 2004], the local incomes are assumed to be computed on a single timescale; however, here we assume that the incomes are computed on multiple timescales in parallel:

$$I_{R,i}^t = \left(1 - \frac{1}{\tau_1}\right) I_{R,i}^{t-1} + \frac{1}{\tau_i} r^{t-1},$$

(4.2)

where $I_{R,i}^t$ is the local income from target Red on trial $t$ ($t = 1, 2, 3, ...$) computed over the timescale of $\tau_i$, and $r^{t-1}$ is 1 (0) when the target was rewarded (no-rewarded) at $t-1$. The local income is a weighted sum of different timescales:

$$I_R^t = \sum_{i=1}^{m} w_i I_{R,i}^t,$$

(4.3)

where the weights $w_i$'s are normalized so that $w_1 + w_2 + .. = 1$. In our analysis, the total number of timescales are set to be 3 and the timescales are set to
CHAPTER 4. MULTIPLE TIMESCALES OF REWARD INTEGRATION
DEVELOP MATCHING BEHAVIORS.

be \( \tau_1 = 2 \) trials, \( \tau_1 = 20 \) trials, and \( \tau_1 = 1000 \) trials. The first two timescales are set to be consistent with the findings in [Corrado et al., 2005]. The last timescale is set to be larger than the block size. Our results are insensitive to varying the choice of exact timescales. We estimated the set of weights \( w_1, w_2, w_3 \) for each monkey on each session by minimizing the mean squared error between the probability of choice predicted by the model and the experimental data.

4.7.2.2 Correlation between choice bias and reward bias

To see the effect of reward history on multiple sessions, we estimated a choice bias of monkeys on each session by the model of LNP model with one timescale presented in [Corrado et al., 2005]. In this model, the choice probability is assumed to be a function of difference in the values of two choices:

\[
P_R = \frac{1}{e^{-\frac{v_R-v_G+\delta}{\tau}} + 1}
\]

where \( V_G \) and \( V_R \) are the values computed on a single timescale \( \tau \),

\[
V_G = \left(1 - \frac{1}{\tau}\right)V_G + \frac{1}{\tau_i}r^{i-1},
\]

\( T \) and \( \delta \) are free parameters. Here \( \delta \) is the estimated bias in choice. The reward bias is computed by taking the fraction of reward obtained from a choice over given session \( r_R/r_R + r_G \). The correlation coefficient is found by comparing the choice bias on session \( s \) and reward bias on session \( s - i \), where \( i = 0, 1, 2, 3, ... \). We compared the values with the results of shuffled data.
4.7.2.3 Comparison of data and neural model

To compare the experimental data with our neural network model, we fitted a sigmoid model with one timescale equation 4.4 with $\delta = 0$ for simplicity. In this model, the reward integration on a timescale less than the block size is effectively fitted by the time constant $\tau$ in equation (4.5). On the other hand, the timescale longer than the block size is incorporated to the magnitude of $T$, as the experiment is design to balance the reward.

4.7.3 Network model with cascade synapses

The network is similar to the one introduced in [Wang, 2002; Fusi et al., 2007; Soltani and Wang, 2006; Iigaya and Fusi, 2013]. In the network, the input population is activated on each trial. This population is fed into two separated excitatory recurrent populations, each of which represent different target Green and Right. Those target populations are mutually inhibited through a inhibitory population. As a result, one of the two target populations will suppress the other, which corresponds to the decision of choosing the winning target. This winner take all process is determined by the synaptic strength between the input population and the target populations. The probability of choosing target red is well approximated by

$$P_R = \frac{1}{e^{\frac{s_R - s_G}{\tau}} + 1}$$

(4.6)
where $S_R$ is the total synaptic strength connecting input population and target population $R$. In our simulation, the synaptic efficacies are assumed to take one of the two values 0 or 1. This assumption of bounded synapses is vital for a biological synapses, though actual number of possible efficacy states is not very sensitive to our findings [Iigaya and Fusi, 2013].

In [Soltani and Wang, 2006], synapses were assumed to be binary and undergo stochastic Hebbian learning, where the synapses targeting rewarded chosen action is potentiated with a certain probability and depressed with a certain probability for synapses targeting no-rewarded chosen action. Those probability represent the plasticity of synapses and the timescale of reward integration is approximately given by $1/\alpha$. Here, we introduce a cascade model of synapses [Fusi et al., 2005], where synapses can undergo metaplastic modification and change plasticity itself $\alpha$. We assumed that synapses can take discrete plasticity states and modification is taken place in the same way as usual changes in strength but different probabilities.

More precisely, the state of a synapse is characterized with two variables, the synaptic strength and the rate of plasticity. For the synaptic strength, we assume that there are only two possibilities, depressed ($-$) or potentiated ($+$). For the rate of plasticity, on the other hand, we assumed that there are three states ($i = 1, 2, 3$). This is just to be consistent with the data analysis. We assume that there are transitions from less plastic to more plastic state with
probability $q_i$. More precisely, a synapse at the depressed (potentiated) state at the level $i$, $i = 1$ or 2, of plasticity, can enter the depressed (potentiated) state at the level $i + 1$ after the targeting action is no-rewarded (rewarded) with a probability of $q_i$, while no transition to deeper state is possible at the least plastic state ($i = 3$). Deeper states (less plastic states) have smaller probabilities of changing synaptic strength $\alpha_1 > \alpha_2 > \alpha_3$.

In our simulation, we used a mean field approximation. In other words, instead of tracking all the synaptic states, we study the distribution of synapses on different synaptic states:

$$\sum_{i=1}^{m} F_i^{(G,R)}^- + \sum_{i=1}^{m} F_i^{(G,R)}^+ = 1,$$

where $F_i^{(G,R)(-,+)}$ is the fraction of depressed ($-$) or potentiated ($+$) synapses with plasticity $\alpha_i$ in the synaptic population targeting green (G) or red (R). Here $m$ is the number of possible plasticity states, which in our case is $m = 3$. Then the probability of choosing target A (A is G or R) becomes

$$P_A = \frac{1}{e^{-S_{A,B}} + 1},$$

where

$$S_{(A,B)} = \sum_{i=1}^{m} F_i^{(A,B)^+}.$$

The learning is as follows. When the network receives a reward after taking an action of choosing A:
CHAPTER 4. MULTIPLE TIMESCALES OF REWARD INTEGRATION
DEVELOP MATCHING BEHAVIORS.

\[
F_{1}^{A^+} \rightarrow F_{1}^{A^+} + \sum_{i=1}^{m} \alpha_{i}^{r} F_{i}^{A^-} - q_{r}^{1} F_{1}^{A^+}
\] (4.10)

\[
F_{1<i<m}^{A^+} \rightarrow F_{1<i<m}^{A^+} + q_{r}^{i-1} F_{i-1}^{A^-} - q_{r}^{i} F_{i}^{A^+}
\] (4.11)

\[
F_{m}^{A^+} \rightarrow F_{m}^{A^+} + q_{r}^{m-1} F_{m-1}^{A^-}
\] (4.12)

and

\[
F_{1}^{B^-} \rightarrow F_{1}^{B^-} + \sum_{i=1}^{m} \gamma \alpha_{i}^{r} F_{i}^{B^+} - \gamma q_{r}^{1} F_{1}^{B^-}
\] (4.14)

\[
F_{1<i<m}^{B^-} \rightarrow F_{1<i<m}^{B^-} + \gamma q_{r}^{i-1} F_{i-1}^{B^-} - \gamma q_{r}^{i} F_{i}^{B^-}
\] (4.15)

\[
F_{m}^{B^-} \rightarrow F_{m}^{A^+} + \gamma q_{r}^{m-1} F_{m-1}^{B^-}
\] (4.16)

where \(\alpha_{i}^{r}\) is the transition probability from state \(i\) to 1 when obtained a reward and \(\gamma\) is the factor for the probability of chaining states for synapses targeting the unchosen action, which is taken to be 0 or very small in our simulation. If the network did not obtain a reward for action A:

\[
F_{1}^{A^-} \rightarrow F_{1}^{A^-} + \sum_{i=1}^{m} \alpha_{nr}^{i} F_{i}^{A^+} - p_{nr}^{1} F_{1}^{A^-}
\] (4.18)

\[
F_{1<i<m}^{A^-} \rightarrow F_{1<i<m}^{A^-} + q_{r}^{i-1} F_{i-1}^{A^-} - q_{r}^{i} F_{i}^{A^-}
\] (4.19)

\[
F_{m}^{A^-} \rightarrow F_{m}^{A^-} + q_{r}^{m-1} F_{m-1}^{A^-}
\] (4.20)
and

\[
F_{1}^{B+} \rightarrow F_{1}^{B+} + \sum_{i=1}^{m} \gamma \alpha_{nr}^{i} F_{i}^{B-} - q_{r}^{i} F_{1}^{B+} \quad (4.21)
\]

\[
F_{1<i<m}^{B+} \rightarrow F_{1<i<m}^{B+} + \gamma q_{r}^{i-1} F_{i-1}^{B+} - \gamma q_{r}^{i} F_{i}^{B+} \quad (4.22)
\]

\[
F_{m}^{B+} \rightarrow F_{m}^{B+} + \gamma q_{r}^{m-1} F_{m-1}^{B+} \quad (4.23)
\]

\[
F_{1}^{B+} \rightarrow F_{1}^{B+} + \sum_{i=1}^{m} \alpha_{nr}^{i} F_{i}^{B-} - q_{r}^{i} F_{1}^{B+} \quad (4.24)
\]

where \( \alpha_{nr}^{i} \) is the transition probability from state \( i \) to 1 when the network obtained no reward.

Between the experimental sessions, we assumed that the network experiences random events:

\[
\frac{\partial F_{1}^{A-}}{\partial t} = -(f_{+} \alpha^{1} + f_{-} q^{1}) F_{1}^{A-} + \sum_{i=1}^{m} \alpha^{i} F_{i}^{A+} \quad (4.25)
\]

\[
\frac{\partial F_{1<i<m}^{A-}}{\partial t} = -(f_{+} \alpha^{i} + f_{-} q^{i}) F_{i}^{A-} + f_{-} q^{i-1} F_{i-1}^{A-} \quad (4.26)
\]

\[
\frac{\partial F_{m}^{A-}}{\partial t} = -(f_{+} \alpha^{i}) F_{1}^{A-} + f_{-} q^{m-1} F_{m-1}^{A-} \quad (4.27)
\]

\[
\frac{\partial F_{1}^{A+}}{\partial t} = -(f_{-} \alpha^{1} + f_{+} q^{1}) F_{1}^{A+} + \sum_{i=1}^{m} \alpha^{i} F_{i}^{A-} \quad (4.28)
\]

\[
\frac{\partial F_{1<i<m}^{A+}}{\partial t} = -(f_{-} \alpha^{i} + f_{+} q^{i}) F_{i}^{A+} + f_{+} q^{i-1} F_{i-1}^{A+} \quad (4.29)
\]

\[
\frac{\partial F_{m}^{A+}}{\partial t} = -(f_{-} \alpha^{i}) F_{1}^{A+} + f_{+} q^{m-1} F_{m-1}^{A+} \quad (4.30)
\]

where \( f_{-} \) and \( f_{+} \) are the rate of depression and potentiation event, respectively.

Transition probabilities \( \alpha^{i} \)'s and \( q^{i} \)'s can be different from the ones during the task. More precisely, \( \alpha^{i} \) can be larger than \( q_{i} \), as it may include the recovery
of plasticity possibly due to the surprise signal that will be discussed more details in the next chapter.

4.8 Conclusion

To conclude, we showed that monkeys dynamically tuned the multiple timescales of memory of reward history and they improved their performance under the bias-variance trade-off. Commonly observed undermatching behaviors were explained by the multiple timescales of reward integration and shown to be co-varied with the weight of the very long timescale. This wide range of timescale can be incorporated by a synaptic plasticity model with different chemical processes with different timescales. One question remains. Do monkeys modify the timescales more locally, for example trial by trial? Recent experimental evidence indeed suggests it. In the next chapter, we propose a biophysical way to achieve the very local modulation of integration timescales, which in our case the rates of synaptic plasticity.
Figure 4.1: Task and matching behavior in monkeys. a Behavioral protocol: the animal has to fixate the central cross, and after a short delay (Delay), it can make a saccadic movement (Go) toward one of the targets. When a reward is assigned to the chosen target, a drop of water is delivered at the time of the response (Return). The overall maximum reward rate is set to 0.15 rewards per second, which is about 0.35 rewards per trial. The relative reward rates changed in blocks (100 to 200 trials) without warning; ratios of reward rates were chosen unpredictably from the set (8:1, 6:1, 3:1, 1:1). b Matching law behavior and deviations from it. Each point represents the blockwise choice fraction against the blockwise reward fraction. The blue line is the diagonal and it represents strict matching. The red line is a linear fit to the datapoints and it is clearly tilted with respect to the diagonal. This type of deviation is called undermatching.
CHAPTER 4. MULTIPLE TIMESCALES OF REWARD INTEGRATION
DEVELOP MATCHING BEHAVIORS. 90

Figure 4.2 (following page): The bias-variance trade-off. The local matching model that integrates reward over a short timescale $\tau_1$ (a) produces a fluctuating estimate of local incomes, and thus the probability of choosing the green target exhibits large fluctuations (b); however, on average, the choice probability matches the blockwise reward fraction (c). Models with a very long integration timescale $\tau_2$ (d) can greatly reduce the fluctuations of the choice probability (e). In this example the choice probability is constant and close to 0.5 because the rewards from the two targets are balanced on long timescales which are longer than blocks. With this long timescale, the choice probability becomes independent of the local reward history, showing a strong deviation from the matching law (f). A model with these two timescales $\tau_1$, $\tau_2$, generates a biased choice probability with a smaller fluctuations (g,h,i). The bias-variance trade-off in the dynamic estimation task can be solved by changing the weights of two integrators with different timescales. j The squared bias (blue) and the variance of estimation (orange) of the model are plotted against the weight of a long timescale. The solid line refer to a block size of 100 trials and the dotted line is when the block size is 10,000 trials. k Squared error of estimation, the sum of bias and variance, is plotted against the weight of a long timescale. The minimum gives the optimal weight. Note that the optimal weight of a long timescale depends on the block size. l Slope of the matching behavior changes when the weight of the long timescales varies.
CHAPTER 4. MULTIPLE TIMESCALES OF REWARD INTEGRATION
DEVELOP MATCHING BEHAVIORS.

Short timescale
$\tau_1 \ll \text{block size}$

Long timescale
$\tau_2 \gg \text{block size}$

Short + Long timescales
$\tau_1$ and $\tau_2$

Bias-variance trade-off

- block = 100 trials
- block = 10000 trials
Figure 4.3 (following page): a,b Monkeys slowly changed the weights of different timescales over multiple sessions. A model with three distinctive timescales ($\tau_1 = 2$ trials, $\tau_2 = 20$ trials, and $\tau_3 = 1000$ trials) are fitted by changing weights $w_1$ (blue), $w_2$ (red), and $w_3$ (green) with $w_1 + w_2 + w_3 = 1$ on each session. In general animals started with short timescales and weighted progressively more the integrators operating on long timescales. The opposite trend around session 160 is due to the shorting of the experimental sessions. c Deviation from the matching law slowly develops with experience. In the early sessions, monkeys show block-wise matching between choice and reward fraction; however, in later sessions the behavior showed a prominent undermatching. d The deviation from the matching law is caused by the reward integration on a long timescale. The deviation from the matching law computed over the last 50 trials of each block is plotted against the fitted value of $w_3$, the weight of the longest reward integration timescale. Both monkeys show a significant negative correlation between the slope and $w_3$. e Monkeys’ choice behavior reflected reward history over sessions. The reward bias that monkeys experienced in a previous session significantly affect the choice behavior in the current session. Both monkeys’ choice are biased by reward bias of up to around five sessions ago, which is more than 3000 trials on average.
CHAPTER 4. MULTIPLE TIMESCALES OF REWARD INTEGRATION
DEVELOP MATCHING BEHAVIORS.

\[ \tau_1 = 2 \text{ trials} \quad \tau_2 = 20 \text{ trials} \quad \tau_3 = 1,000 \text{ trials} \]

\[ w_1 + w_2 + w_3 = 1 \]

\[ P_a = \frac{l_a}{l_a + l_r} \]

\[ r_G = -0.63, p < 10^{-5} \]

\[ r_F = -0.58, p < 10^{-6} \]
Figure 4.4 (following page): Monkeys harvesting efficiency improved over experience and it is determined by the changes in the long timescale $w_3$. a Fraction of available rewards collected by monkeys are shown as a function of the weight of the long timescale $w_3$ for days 1 to 10 (blue) and for days 11 to 20 (red). Solid line is a quadratic fit of points. Dotted line is a model’s performance with optimal weights $w_1$ and $w_2$ given $w_3$. The simulation is done on the actual reward schedule used for experiments. Gray area indicates the standard deviation of optimal model’s performance per session. Note that both monkeys became closer to the optimal by changing the weight of long timescale $w_2$. b Monkeys trade-offed bias and variance. The slope of the matching behavior and variance of monkeys choice probability are shown to be correlated. Smaller slope means larger bias; hence the monkeys increased bias to decrease the variance of estimation.
CHAPTER 4. MULTIPLE TIMESCALES OF REWARD INTEGRATION
DEVELOP MATCHING BEHAVIORS.

(a) Monkey F and Monkey G

(b) Monkey F and Monkey G

Harvesting efficiency vs. $w_3$

Matching vs. Undermatching

Slope vs. Fluctuation

Day 1-20 vs. Day 21-40

$r = 0.65$

$p < 10^{-8}$

$r = 0.59$

$p < 10^{-3}$
Figure 4.5 (following page): Neural network model with synaptic plasticity on multiple timescales.  

**a** The network model. The decision is made based on the competition between the two target populations through the inhibitory population. The competition is controlled by the synaptic strength between the input population and the target populations.  

**b** The cascade model of synapses. Each plastic synapse between the input population and the target population can take either or two strength: potentiated or depressed, represented by the strength of 1 and 0, respectively. In addition, each synapse can take one of the metaplastic states with different rates of changing the synaptic strength. The metaplastic transition (vertical arrows) can take place in the same fashion as the normal plastic transition (horizontal and diagonal arrows) but with different probabilities. In general, the lower states are less plastic; thus they have lower probabilities of changing states. In this example, there are three levels of plasticity, and $\alpha_1 > \alpha_2 > \alpha_3$, $q_1 > q_2$. 
CHAPTER 4. MULTIPLE TIMESCALES OF REWARD INTEGRATION
DEVELOP MATCHING BEHAVIORS.

\[ \alpha_1 - \alpha_2 + \alpha_3 \]

When targeting population is rewarded

When targeting population is not rewarded
CHAPTER 4. MULTIPLE TIMESCALES OF REWARD INTEGRATION
DEVELOP MATCHING BEHAVIORS.

Figure 4.6: Synaptic plasticity on multiple timescales accounts for the data. 

Changes in matching behavior is reproduced by the network model. Assuming 
that the synapses occupied the states equally on day 1, slow transitions to the 
state with longer timescales introduce a bias toward the mean of the reward on 
the long timescale. This leads to an increase of deviation from the matching 
law, which we observed in the experiment. The model is under the trade-off 
between the bias (1−slope) and the variance. The plastic synapses (at the top 
level) create a large variance but a small bias, while the less plastic synapses 
(at lower levels) induce a small variance but a large bias. As the synapses 
change the distribution on the plasticity space, the model trades off the bias 
and the variance of the estimation. This is consistent with our experimental 
findings.
Figure 4.7 (following page): Non monotonic changes in the weight $w_3$ of the long timescale $\tau_3$ reflect the experimental schedule on a long timescale. 

**a,b**, The weight of long timescale $w_3$ correlates with the recent experimental lengths. Daily estimation of $w_3$ is plotted against the mean length of recent experiments. The weight of the long timescale $w_3$ is larger when the animal constantly experienced long experiments with may trials. The mean is taken over 18 experimental days (Monkey F) and 12 experimental days (Monkey G), respectively, as they give the largest correlations. 

**c,d**, The weight of long timescale $w_3$ anti-correlates with the mean recent inter-experimental-intervals. The weight of the long timescale $w_3$ is smaller when the animal constantly had long inter-experimental-periods. Daily estimation of $w_3$ is plotted against the mean recent inter-experimental-intervals. The mean is taken over 25 experimental days (Monkey F) and 32 experimental days (Monkey G), respectively, as they give the largest magnitude of correlations.
CHAPTER 4. MULTIPLE TIMESCALES OF REWARD INTEGRATION
DEVELOP MATCHING BEHAVIORS.

(a) Monkey F
Mean recent number of trials (18 days)
\[ r = 0.64 \]
\[ p < 10^{-6} \]

(b) Monkey G
Mean recent number of trials (12 days)
\[ r = 0.77 \]
\[ p < 10^{-12} \]

(c) Monkey F
Mean recent inter-experiment-intervals (25 days)
\[ r = -0.54 \]
\[ p < 10^{-4} \]

(d) Monkey G
Mean recent inter-experiment-intervals (32 days)
\[ r = -0.49 \]
\[ p < 0.004 \]
CHAPTER 4. MULTIPLE TIMESCALES OF REWARD INTEGRATION
DEVELOP MATCHING BEHAVIORS.

Figure 4.8 (following page): Changes in the effect of short timescale (learning rate $\eta$, top) and the long timescale (temperature $T$, bottom) in the network model and data. The data generated by simulation of network model and the behavioral data are fitted with the same LNP model. The LNP model has two parameters: the learning rate $\eta$ capturing the reward integration of short timescale (smaller than the block size), and the temperature $T$ reflecting the effect of long time scale reward integration. Both model and data show qualitatively similar trends. That is, the reward integration timescale expands up to around session 150, after which the timescale decreases. This is due to the experimental schedule on Monkey F, and for the model it reflects the interplay between the memory consolidation and forgetting.
CHAPTER 4. MULTIPLE TIMESCALES OF REWARD INTEGRATION
DEVELOP MATCHING BEHAVIORS.
Figure 4.9 (following page): The local matching model with a single timescale can also capture our key findings. Although we showed that the behavior is strongly biased by a very long timescale, we can capture some aspects of the experiment by fitting with the local matching law model with one timescale. 

**a,b,** The effective learning rate correlates with the slope of the choice fraction vs reward fraction. The effective learning rate can be thought of a weighted average of the multiple timescales. **c,d** The learning rate is larger when the animal constantly experienced long experiments with may trials. Note that the learning rate $\eta$ is approximately the inverse of the reward integration timescale $\eta = \tau^{-1}$ **e,f** the learning rate is smaller when the animal constantly had long inter-experimental-periods.
CHAPTER 4. MULTIPLE TIMESCALES OF REWARD INTEGRATION
DEVELOP MATCHING BEHAVIORS.

---

**A**

**Monkey F**

\[
\text{Learning rate} \quad r = 0.55 \quad p < 10^{-6}
\]

**B**

**Monkey G**

\[
\text{Learning rate} \quad r = 0.53 \quad p < 10^{-6}
\]

**C**

**Monkey F**

\[
\text{Recent mean number of trials per day} \quad r = -0.70 \quad p < 10^{-9}
\]

**D**

**Monkey G**

\[
\text{Recent mean number of trials per day} \quad r = -0.49 \quad p < 10^{-4}
\]

**E**

**Monkey F**

\[
\text{Recent mean inter−experiment−intervals [day]} \quad r = 0.52 \quad p < 10^{-4}
\]

**F**

**Monkey G**

\[
\text{Recent mean inter−experiment−intervals [day]} \quad r = 0.37 \quad p < 0.005
\]
Figure 4.10: Non cascade type synaptic model dose not account for the changes in matching behaviors observed in the experiment. If the decision making network has synapses with two independent plasticities, the behavior should not change as opposed to the changes observed in the experiment. Hence the behavior implies the cascade type interactive plasticity model.
Figure 4.11: Changes in matching behaviors over the course of experiments.

**a** Early days of experiments show good matching between choice fraction and income fraction, while later days of experiments show deviations from the matching. **b** The deviation became prominent after the finishing the fast transition from the previous block of reward fraction, which did not present in early days of experiment.
Chapter 5

Modeling flexible value-based decision making by neural network models with plastic synapses.

In the previous chapter we showed that monkeys accumulate reward evidence over multiple timescales and they dynamically tune their weights of the timescales. As a result, animals improved their foraging performance under the bias-variance trade-off. While in the previous chapter we focused on the slow changes of the weights averaged on each session, here we focus on changes on a more local timescale (trial by trial). In fact, recent experimental
evidence suggests that animals (especially humans) can modify the learning rate (the timescale of reward accumulation) on a trial by trial basis. Is our model capable to reproduce those results? What would be the minimal neural circuit mechanism explaining those behaviors? These are the questions that we address in this chapter. This work is presented at the Society for Neuroscience in 2011 [Iigaya and Fusi, 2011] and a journal manuscript is in preparation.

5.1 Abstract

Recent studies suggest that animals may adapt their decisions trial by trial, by dynamically changing the timescales of learning. Yet its neural mechanism is not clear. Here we propose that dynamic changes in synaptic plasticity can be a mechanism behind the adaptive behavior. We show that a decision making network with the synapses that can change the rate of plasticity can dynamically modulate the timescale of reward accumulation. When the environment is stable, the synapses become less plastic so that the integrating timescale becomes longer and the values encoded in the synapses become consolidated. On the other hand, a rapid unexpected change in the environment is detected by an additional network, which sends a surprise signal to the decision making network. As a result, the synapses in the decision network gains more plasticity, which allows the network to explore the environment to learn a new environment. Importantly, our model does not need to assume
any prior knowledge of reward environment; rather it changes the distribution of synaptic plasticity dynamically according to the environment, independent of the initial condition. Thus it can adapt to any environmental changes including the one that is not expected from the past. We show that our model with changing the levels of synaptic plasticity can perform significantly better than any fixed plasticity model in a simple multi-armed bandit task. We also point out that if the surprise detection network has a malfunction, the system cannot consolidate or update memory, resulting in a highly volatile or a highly perseverative behavior.

5.2 Introduction

We make decisions based on our future forecast, and the forecast is based on our past experience, which is stored as memory in our brain. Since environment changes time to time, we need to keep accumulating and updating our memory in order to make appropriate decisions. The important question here is how much we should update our memory after each action. If the update is large, we rely on the recent trends and discard the long timescale statistics. On the other hand, if the update is small, we can have a good estimate on a long timescale, only by paying the price of ignoring the recent changes even if they are prominent or meaningful. Thus there is a trade-off in how much we should update our belief on each time step.
Indeed, in the classical reinforcement learning theory of the decision making, the future reward is estimated by integrating past reward history on a single timescale (which is characterized by the learning rate $\alpha$, how much you would update the estimation), and this timescale is assumed to be fixed over many trials. Here, this single timescale integration inevitably leads to a problem of speed accuracy trade-off [Soltani and Wang, 2006]. That is, if the environment is stable, one can have a good estimate of future rewards by integrating past on a long timescale; however, this long timescale integration is harmful when the environment is volatile, requiring an integration on a shorter timescale to obtain a reasonable reward estimate.

How do animals solve this speed accuracy trade-off? There are a number of evidence that animals integrate rewards sequence on multiple timescales [Corrado et al., 2005; Fusi et al., 2007; Bernacchia et al., 2011] or they dynamically change the integrating timescale depending on the environment [Behrens et al., 2007; Rushworth and Behrens, 2008; Soltani et al., 2006; Nassar et al., 2010; 2012; Neiman and Loewenstein, 2013]. In [Corrado et al., 2005], monkeys had to allocate their choice depending on the reward schedule, where the monkeys were found to integrate the reward history on multiple timescales. In [Behrens et al., 2007], Behrens et al. showed in a simple two-armed bandit task that humans have a larger learning rate in a volatile phase than in a stable phase, where in the volatile (stable) phase the probability of getting reward from
different actions alternate rapidly (slowly). Moreover, in [Nassar et al., 2010; 2012; Wilson et al., 2013], the learning rate is shown to monotonically decreases over trials in a fixed environment and it jumps up after a change in the environment.

It has also been shown that the observed changes in learning rates are close to the optimal predicted in the Bayesian framework [Gallistel et al., 2001; Behrens et al., 2007; Nassar et al., 2010; 2012; Wilson et al., 2013]; however, the actual biological implementation of the dynamical modulation of integration timescales is not clear. Recent studies suggest that midbrain neurons and cortical neurons show multiple timescales of memory [Bromberg-Martin et al., 2010; Bernacchia et al., 2011; Kim and Hikosaka, 2013]. Also, from single neuron adaptation to memory decays and free recall, increasing evidence suggests that multiple timescales or a power law scaling are involved [Wixted and Ebbesen, 1991; La Camera et al., 2006; Ulanovsky et al., 2004; Fusi et al., 2005; Wark et al., 2009; Lundstrom et al., 2010; Pozzorini et al., 2013; Romani et al., 2013]. Here we show that a neural network with synaptic plasticity model originally proposed to solve a fundamental memory problem [Fusi et al., 2005] can account for the observed adaptive choice behaviors including the changes in learning rates. We also show that a key feature of our model, multiple timescales of reward accumulation, can be a great advantage for a relatively complex decision making task.
In our model, changing learning rate in behavior is a result of changing distribution of plasticity in synaptic population, which naturally incorporate the experimental fact that reward is integrated on multiple timescales [Corrado et al., 2005; Fusi et al., 2007; Bernacchia et al., 2011] and the integration timescale change over trials [Behrens et al., 2007; Soltani et al., 2006; Nassar et al., 2010; 2012]. Indeed, we show that our model can reproduce the key experimental results of changing integrating timescale over different reward blocks [Behrens et al., 2007] and within the same reward block [Nassar et al., 2010; 2012]. Moreover, we show that the integrate timescale of our model is automatically tuned to optimal, and it can perform significantly better than a model with any fixed timescale when the environment changes on multiple timescales.

5.3 Results

5.3.1 Synaptic plasticity determines the timescale of reward history integration.

The basic architecture of the decision making network is described in the previous chapters. The most crucial part of the network [Soltani and Wang, 2006; Fusi et al., 2007; Wang, 2008] described in figure 5.2 is the plastic synapses feeding inputs to decision making neurons representing different actions, as the
decision probability can be well described as a function of the total efficacy of the plastic synapses [Soltani and Wang, 2006]. We assume that the efficacy of the plastic synapses is bounded and binary (weak or strong) [Amit and Fusi, 1994; Fusi and Abbott, 2007] and the efficacy can be changed by a reward modulated stochastic Hebbian learning [Seung, 2003; Soltani and Wang, 2006; Fusi et al., 2007] described in the previous chapters. The essential point here is that synapses can change their efficacy with a certain probability \( \alpha \), as consequence of an action followed by a reward or no reward. This transition probability is closely related to the plasticity of synapses, as synapses with a larger transition probability \( \alpha \) is more vulnerable to change. Thus we can think of \( \alpha \) as a measure of plasticity, which takes value from 0 to 1. Naturally, this also related to the timescale of reward integration; roughly speaking the synapses integrate the reward information on the timescale of \( 1/p \).

5.3.2 The fixed plasticity models face the speed-accuracy trade-off

Although it has been shown that the model synapses can estimate the probabilistic value such as reward probabilities or returns of actions, the estimation is reliable only if the plasticity of synapses are very small (\( \alpha = \) small) [Soltani and Wang, 2006]. On the other hand, if the plasticity is too small, it becomes very difficult to adjust the estimation when there is a sudden change
in environment. This speed-accuracy trade-off is described in figure 5.1. For simplicity, here we assume that each synapse can be either at strong or weak state in their strength. The plasticity of synapses in this case is described by a transition probability from one state to another, denoted by $\alpha$. As seen in figure 5.1, if the plasticity is too high ($p = \text{large}$), the difference in total synaptic strengths of different actions becomes very volatile and the estimation of reward probability becomes unreliable. On the other hand, if the plasticity is small, or $\alpha$ is small, the estimation is accurate; however, this leads to a fatal disability of learning a new value when there is a sudden change.

The heart of this problem lies in the fixed plasticity $\alpha$, or the timescale on which the reward information is integrated by synapses $\tau = 1/\alpha$. The model can obtain a good estimation of a stable value when the timescale of reward integration is large, which in our case small $\tau$ or large plasticity rate $p$. This is only good when the value is stable. When the value is not stable but rapidly changing, it is more important to keep track of local changes in the value. This can be done in our model with a short timescale of reward integration (small $\tau$ or a high plasticity rate $p$).
5.3.3 Changing plasticity according to the environment: the cascade model synapses and the surprise detection network

How animals solve the speed-accuracy trade-off? Experimental studies suggest that they integrate the reward history on multiple timescales rather than a single timescale [Corrado et al., 2005; Fusi et al., 2007; Bernacchia et al., 2011]. Also, other studies showed that animals can change the integration timescale, or the learning rate, depending on the environment [Behrens et al., 2007; Nassar et al., 2010; 2012]. To incorporate these findings to our model, we introduce synapses that can change plasticity $\alpha$ itself in addition to the strength (week or strong) depending on the environment. The model is called the cascade model, which was originally proposed to improve a general memory performance of a neural network [Fusi et al., 2005]. In this model, described in figure 5.2B, the degree of synaptic strength is still assumed to be binary (week or strong); however, there are $m$ states with different levels of plasticity $\alpha_1, \alpha_2, \ldots, \alpha_m$, where $\alpha_1 > \alpha_2 > \ldots > \alpha_m$. There are also transitions from one level of plasticity to another level of plasticity denoted by $p_i$ ($i = 1, 2, \ldots, m - 1$). We also assume that $p_1 > p_2 > \ldots > p_{m-1}$, meaning that entering less plastic states gets harder and harder as go further to less plastic states; however, at the same time the changing synaptic strength also becomes harder as the synapses become less plastic. It has been shown that the logarithmically distributing
transition probabilities (ex. \( \alpha_i = \left(\frac{1}{2}\right)^i \) and \( p_i = \left(\frac{1}{2}\right)^i \)) is sufficient to exploit the power of this model as it efficiently covers a wide range of timescale [Fusi et al., 2005].

As shown in Fig. 5.3A, cascade model synapses can reduce fluctuations of synaptic strength when the network is in a steady environment (where the reward probability is fixed.). This is because more and more synapses make transitions to less plastic states when the network is in a stable environment. This recruits longer integrating timescales, resulting in a consolidation of memory of action values. This consolidation, however, leads to another problem. Namely, when there is a sudden change in the reward environment, the network cannot update its estimation because the synaptic plasticity is very low (figure 5.3A). Thus we find that the naïve cascade model is good for consolidating memory but it is harmful when it has to rapidly update the probability estimation.

To solve this problem, we introduce a network that can detect an unexpected change in the environment and send a surprise signal to the decision making network to increase the synaptic plasticity of the decision making network. In this network, plastic synapses receive reward signal or no-reward signal on each trial and undergo stochastic learning independent of the action. The crucial difference of synapse model of this network and decision making network is that the synapses in surprise network do not change the
plasticity $p_i$’s. Rather, they have fixed plasticity $p_i$’s and integrate reward on fixed timescales ($\tau_i = 1/p_i$’s) (See Fig. 5.2). What the network does is to compare the average difference of synaptic strength of different timescales (expected uncertainty) and the actual difference on current trial (unexpected uncertainty) [Yu and Dayan, 2005]. More precisely, when an unexpected uncertainty significantly exceeds the expected uncertainty, a surprise signal is sent to the decision making network and the synapses increase the plasticity in the network. It is schematically shown in figure 5.4.

5.3.4 The model can self-tune the learning rates near optimally

Here we show that our model can account for the observed flexible changes in learning rates [Behrens et al., 2007; Nassar et al., 2010]. In our model, the effective learning rate can be defined as the transition rate $\alpha_i$’s weighted by synaptic populations. As seen in figure 5.5A, the effective learning rate is large when environment is rapidly changing, where it is small when the environment is steady. This is consistent with the findings in [Behrens et al., 2007]. Also, within a steady environment, the learning rate is largest after the change point and slowly decay over trials. This is consistent with the experimental findings in [Nassar et al., 2010]. Note that our model does not assume a priori any timescale of the environment. Rather, it can dynamically adapt to any
environment by automatically changing the distribution of synaptic plasticity.

Figure 5.5B, shows that our model can tune their plasticity (and hence the learning rate) depending on the volatility of environment (block size) and it is very close to the optimal learning rate of a single learning rate model that maximizes the harvesting efficiency. The effective learning rate of our full model is defined by the average potentiation/depression rate weighted by the synaptic population on each state. Since the population is always shifting its distribution, we take the time average of our simulated data over blocks of trials to compute the effective learning rate given specific block size. On the other hand, the single learning rate model is our network model with a single plasticity level without any surprise signal.

This agreement is remarkable, as we did not assume any specific timescales in our model of plasticity; rather, we assume a wide range of timescales \( 1/\alpha_i \)'s) and synapses themselves make metaplastic transitions to tune the overall learning rate of the network through time.

Does our network with metaplastic synapses collect more rewards than a model with non-metaplastic synapses with an optimal timescale? One can expect this is the case when the task’s timescale changes over time. To see this, we compared the performance of our models in a simple multi-armed bandit task with variable block sizes. Indeed, as seen in figure 5.6, we found that our model with metaplastic synapses can perform significantly better than any
CHAPTER 5. MODELING FLEXIBLE VALUE-BASED DECISION MAKING BY NEURAL NETWORK MODELS.

model with a fixed single synaptic plasticity.

Here we show that the model can actually perform significantly better than non-metaplastic model in a certain task. For example, in a simple multi-armed bandit task with variable block sizes, we found that our model can perform significantly better than any model with fixed single synaptic plasticity as seen in figure 5.6. In this task, the block size is not constant but can take one of the two different lengths. Thus the subject has to dynamically tune the learning rate according to the current block length. Thanks to the self tuning property, the model with cascade synapses with a surprise signal collects significantly more rewards than any single plasticity model with a single integration timescale.

5.3.5 The distribution of synaptic plasticity over a wide range of timescales is required when the task contains a multi-timescale structure

The key point of our synaptic model is that synapses are distributed over different plasticity states. In other words, synapses covers a wide range of timescales at the population level. So far, we only focused on the mean of the synaptic distribution as an effective learning rate and show that it is optimally tuned by the model itself; but here we show that the synaptic distribution over a wide range of timescales is required when a task involves a multi-timescale
structure. For example, imagine a task where a subject has to choose a target from many available targets (say, 25). One of the many targets is much more rewarding than the other ones on each trial. The most rewarding target is fixed for a given block of trials, while it changes to another target at the end of the block. In this task, the most rewarding target is always chosen from a fixed small fraction of targets.

In this task, the subject should learn the fraction of targets (hot spots) that can be the most rewarding target, and on top of that the subject should find the target that is currently most rewarding. Otherwise, the subject has to explore all the targets every time the block changes, which is now a large penalty although in a typical two alternative choice task the penalty was very small.

This inference on two timescales can be achieved in our model of the synaptic plasticity on multiple timescales. As seen in figure 5.7, the decision making network model with two synaptic timescales can perform significantly better than the one with the optimal single synaptic timescale. Here we assume a simple distribution of synaptic plasticity (more plastic : less plastic = 1 : 1) and kept fixed in order to focus on the function of multi-timescale reward history integration. The two-timescale model outperforms, since the synaptic plasticity on a long timescale can learn the hot spots and the one on a short timescale can learn the actual most rewarding target on top of it.
In reality, we always face this type of task involving internal structures, where we have to infer the relevant targets depending on the context that change slowly over time. Even in a very simple task such as the dynamic foraging task [Sugrue et al., 2004; Lau and Glimcher, 2005], monkeys are trained to make saccadic eye movement to one of the two targets. This means that they are well trained and do not usually make a saccade to an empty white space, which can be thought of a result of slow learning by less plastic synapses.

Note that the synaptic plasticity on a long timescale works like a prior in a Bayesian inference framework. In fact, our result suggests that our priors of an environment can be dynamically learned biologically by synaptic plasticity on a long timescale or increasing the number of less plastic synapses. In fact, our finding suggest that the prior term in general should contain information learned over multiple timescales and should change dynamically according to the environment.

5.4 Discussion

5.4.1 Relation to the Bayes optimal inference models

We showed that our biophysical network with plastic synapses can adapt to changes in an environment by changing the effective learning rate of the net-
work by changing the synaptic population over discrete states with different rates of synaptic plasticity. The effective learning rate in our model is a population weighted average of different timescales determined by the rates of synaptic plasticity. The essential computation here is that the model changes the weights of different timescales of reward integration. This idea is also recently developed in a Bayesian change detection model [Wilson et al., 2013].

Starting with a full Bayesian online change point detection algorithm [Adams and MacKay, 2007], Wilson et al. developed a reduced model of the full algorithm. Instead of tracking an actual context length, or a run length, they introduced a probability distribution over discrete run lengths to be tracked over time and showed that the reduced model can work as good as the full change detection model and that it fits human behaviors. The motivation for them to introduce a reduced multiple timescale model is that the full Bayesian model has a huge computational cost that cannot be achieved in our physical brain, as the model requires to update the information about all the past time step. The reduced model, however, keeps track of changes in averaged value on several timescales; thus computational cost is significantly reduced. In their paper they claim that it is more biophysically plausible than the full model.

In this paper, on the contrary, we started with a biophysically plausible network and we explored a way to achieve adaptive decision behaviors. We found that the previously proposed synaptic model with discrete plasticity
states [Fusi et al., 2005] can perform highly flexible value estimations. The flexibility was more improved by a surprise detection network, which is similar to the idea of change point detection, although in our model the change point is found for different timescales and we do not reset all the timescales even if there is a change on some timescales. This allows us to see the benefit of having multiple timescales at the same time when a task involves multiple timescale structures, as our model’s goal is to cover a wide range of timescales that is relevant to the current task, instead of finding change points. Nonetheless, it is interesting that we separately came to propose similar kinds of computation, a learning on multiple discrete timescales, from two opposite sides. From David Marr’s point of view [Marr, 1982], we started from the lower implementation level, whereas Wilson et al. started from the higher levels of normative and algorithmic requirements.

5.4.2 Comparison to the multiple stable efficacy synaptic models

Note that the trade-off discussed in this chapter was also discussed in chapter 3, where we allowed synapses to have multiple stable values of efficacy. This modification, however, did not improve the model’s performance, as the increase in the number of efficacy states is similar to the decrease in the synaptic plasticity. In other words, the multi-stable efficacy model merely change
the learning rate $p$ to $p/n$, where $n$ is the number of synaptic states with different synaptic strength. It was also shown that this type of modification does not improve a simple memory performance [Fusi and Abbott, 2007].

In this chapter we show that meta-plastic changes in synapses significantly improve the model’s performance in dynamic foraging task and that the model’s behaviors are consistent with recent experimental findings. Note that in a simple memory problem, the metaplastic synapses also enhance network’s memory storage performance [Fusi et al., 2005].

5.5 Methods

The model consists of two networks: 1) a decision making network, which makes decisions according the actions values stored in plastic synapses 2) a surprise detection network, in which synapses compute expected uncertainties and unexpected uncertainties on multiple timescales to send a surprise signal to the decision making network when it detect unexpected uncertainty exceeds the expected uncertainty.
5.5.1 The decision making network with cascade type synapses

The decision making network [Soltani and Wang, 2006; Fusi et al., 2007; Wang, 2008] is illustrated in Fig.5.2. In the network, two groups of excitatory neurons that are selective to target stimuli (A or B) receive inputs from sensory neurons on each trial. We can also assume the overlapping of cells in those two decision making populations, which would be incorporated to the noise or temperature $T$ in our model. The target neurons compete with each other though inhibitory neurons that feed back to excitatory populations. As a result, one of the target populations win the competition (winner take all process) [Wang, 2002]. We assume that this leads to the action of choosing the corresponding target.

The resulting decision is probabilistic but its probability can be well fitted by a function of the synaptic efficacies $E_A$ and $E_B$ [Soltani and Wang, 2006] :

$$P_A = \frac{1}{e^{-\frac{E_A-E_B}{T}} + 1},$$

(5.1)

where $P_A$ is the probability of choosing target A and the temperature $T$ is a free parameter describing the amount of noise in the network.

After an action is taken, a reward or no reward is given probabilistically according to task’s reward schedule, which feeds into the decision making network as a reward signal. The reward signal evokes a stochastic Hebbian learning in plastic synapses so that the synapses can change its efficacy, whereas
the efficacy of each synapse is assumed to be bounded [Amit and Fusi, 1994; Fusi and Abbott, 2007]. It is known that this biologically plausible assumption of the bounded synaptic strength leads to a fatal limit on a memory capacity, such as an exponential decay of memory signal [Amit and Fusi, 1994; Fusi and Abbott, 2007]. One of the solutions proposed to solve this problem is to allow synapses to change their plasticity itself, in addition to the efficacy [Fusi et al., 2005]. It is called cascade model and it incorporates chemical cascade processes that take place on multiple timescales in plastic synapses.

In our case, if a synapse has a probability of changing its efficacy $\alpha$ per trial, then the memory of the reward given on that trial can last on the timescale of $1/\alpha$ trials. This means that a decision of decision making neural circuit on a given trial is based on the reward history integrated on the timescale of $1/\alpha$. For example, if $\alpha = 0.1$, the decision is made based on the reward experience on the last 10 trials. This also leads to a fatal problem in our case. For example, if $\alpha$, which from now on we call as a plasticity or a learning rate, is small, the estimate of action values will be very volatile. This is good when the environment is also volatile; however if the environment is stable, one should have a smaller plasticity $p$ so that the timescale of the integration is large enough to make the action value more reliable. In fact, there is a speed-accuracy trade-off of changing synaptic efficacy in the decision making network. This tradeoff has been studied in Bayesian framework and
found that the optimal is to change the learning rate dynamically according to the environment [Behrens et al., 2007]. This corresponds to changing the probability of synaptic plasticity $\alpha$ according to the environment. How do we do that?

One way is to achieve this is to allow synapses to change plasticity $\alpha$. We apply a model introduced in a memory system called the cascade model [Fusi et al., 2005]. Details are schematically described in Fig.5.2 B. In this model, each synapse is in one of the bounded synaptic states that has different efficacy and also different plasticity rate. Say, a synapse that feeds into a target A neuron is at the "depressed-1" state. If a reward is given to target A, then the synapse can be potentiated and go to the state "potentiated-1" with probability $\alpha_1$. Imagine that the synapse actually went through the potentiation and now at the "potentiated-1" state and again the action of choosing target A is rewarded, then the synapse can go to a less plastic (plasticity = $\alpha_2$) state "potentiated-2" without changing its efficacy with a probability of $p_1$. In that state, the synapse can still be depressed after a depressing event (no-reward); however, with a smaller probability (now it’s $\alpha_2 < \alpha_1$.) This cascade model of synapses that include essence of different chemical cascade process can achieve a great performance improvement in memory performance in a simple task and shows power law decay of memory signal rather than exponential.
5.5.2 The surprise detection network

To solve this problem, we introduce a network that can detect an unexpected change in its environment and that the network can send a surprise signal to the decision making network. Then the surprise signal triggers an increase of synaptic plasticity in the decision making network. In the surprise network, plastic synapses receive a reward signal or a no-reward signal on each trial. There is a distribution of synaptic plasticity in this network so that the reward information is integrated on multiple timescales. The model of synapses is similar to the cascade model in decision making network; however, in this network synapses do not change their plasticity rates. Rather, they have fixed rates of plasticity $\alpha_i$’s that do not change over trials. What this network does is to compare the difference between the reward memory traces computed on different timescales and compute an expected uncertainty (an average of the difference over the longer timescale of the comparison) and compare it with the actual difference (an unexpected uncertainty) on the current trial. It is schematically shown in Fig.5.2. The computation is done by a population level; thus we take a mean field approximation; thus for details, see section 5.5.3.
5.5.3 Mean field theory

Instead of tracking all the synaptic states, one can study the distribution of synapses on different synaptic states:

\[
\sum_{i=1}^{m} F_i^{(A,B)-} + \sum_{i=1}^{m} F_i^{(A,B)+} = 1. \tag{5.2}
\]

Then the probability of choosing target A becomes

\[
P_A = \frac{1}{e^{-E_{(A,B)} - E_B}} + 1, \tag{5.3}
\]

where

\[
E_{(A,B)} = \sum_{i=1}^{m} F_i^{(A,B)+}. \tag{5.4}
\]

When the network received a reward after choosing target A:

\[
F_1^{A+} \rightarrow F_1^{A+} + \sum_{i=1}^{m} \alpha_i F_i^{A-} - p_i F_i^{A+} \tag{5.5}
\]

\[
F_{1<i<m}^{A+} \rightarrow F_{1<i<m}^{A+} + p_i F_{i-1}^{A+} - p_{i+1} F_i^{A+} \tag{5.6}
\]

\[
F_m^{A+} \rightarrow F_m^{A+} + p_i F_{m-1}^{A+} \tag{5.7}
\]

and
\[ F_{1}^{B-} \rightarrow F_{1}^{B-} + \sum_{i=1}^{m} \gamma \alpha_{i} F_{i}^{B+} - \gamma p_{1} F_{1}^{B-} \quad (5.9) \]

\[ F_{1<i<m}^{B-} \rightarrow F_{1<i<m}^{B-} + \gamma p_{i} F_{i-1}^{B-} - \gamma p_{i} F_{i}^{B-} \quad (5.10) \]

\[ F_{m}^{B-} \rightarrow F_{m}^{A+} + \gamma p_{i} F_{m-1}^{B-} \quad (5.11) \]

\[ F_{1}^{A-} \rightarrow F_{1}^{A-} + \sum_{i=1}^{m} \alpha p_{i} F_{i}^{A+} - p_{1} F_{1}^{A-} \quad (5.13) \]

\[ F_{1<i<m}^{A-} \rightarrow F_{1<i<m}^{A-} + p_{i} F_{i-1}^{A-} - p_{i} F_{i}^{A-} \quad (5.14) \]

\[ F_{m}^{A-} \rightarrow F_{m}^{A-} + p_{i} F_{m-1}^{A-} \quad (5.15) \]

\[ F_{1}^{B+} \rightarrow F_{1}^{B+} + \sum_{i=1}^{m} \gamma \alpha_{i} F_{i}^{B-} - \gamma p_{1} F_{1}^{B+} \quad (5.16) \]

\[ F_{1<i<m}^{B+} \rightarrow F_{1<i<m}^{B+} + \gamma p_{i} F_{i-1}^{B+} - \gamma p_{i} F_{i}^{B+} \quad (5.17) \]

\[ F_{m}^{B+} \rightarrow F_{m}^{B+} + \gamma p_{i} F_{m-1}^{B+} \quad (5.18) \]

where \( \alpha_{i} \) is the transition probability from state \( i \) to 1 when the network obtains a reward, and \( \gamma \) is the factor determining the probability of chaining states of synapses targeting an unchosen action at a given trial. Similarly, when the network received reward after choosing target A:

\[ \text{where} \quad \alpha_{i} \text{ is the transition probability from state } i \text{ to } 1 \text{ after the network obtained no reward.} \]
CHAPTER 5. MODELING FLEXIBLE VALUE-BASED DECISION MAKING BY NEURAL NETWORK MODELS.

The surprise network integrates a series of rewards on different timescales.

When the network received a reward:

\[ R^+_i \rightarrow R^+_i + \alpha_i R^{B-}_i \]  \hfill (5.20)

\[ R^-_i \rightarrow R^-_i - \alpha_i R^{B-}_i \]  \hfill (5.21)

When the network received no reward:

\[ R^-_i \rightarrow R^-_i + \alpha_{nr}^i R^{B+}_i \]  \hfill (5.22)

\[ R^+_i \rightarrow R^+_i - \alpha_{nr}^i R^{B+}_i \]  \hfill (5.23)

On each trial, the network computes the difference between \( R \)'s on the slower timescales of comparisons:

\[ < R >_{i,j} = < R >_{i,j} + \min(\alpha^i_j, \alpha^j_i)(|R_i - R_j|) < R >_{i,j} \]  \hfill (5.24)

when \( i < j \). This is expected uncertainty [Yu and Dayan, 2005]. On each trial, it compares this value with an unexpected uncertainty \(|R_i - R_j|\), or the actual current difference between pairs of \( R \)'s. When the expected uncertainty \( < R >_{i,j} \) significantly exceeds the current unexpected uncertainty \(|R_i - R_j|\), the surprise network sends a surprise signal to the decision making network so that the synapses in the decision making network can increase the rates of plasticity. Importantly this is done only for the levels of synapses that the surprise is detected and slower levels do not change the rates of plasticity.
This allows the network to keep information on different timescales as long as it is useful. For example, when the p-value \((P)\) that current unexpected uncertainty between \(i\)’th and \(j\)’th level lies in the gaussian distribution with the expected uncertainty as its variance becomes smaller than the threshold \(H (P < H)\), we set

\[
\alpha^k_r \to \alpha_r \\
\alpha^k_{nr} \to \alpha_{nr}
\]

for \(k \leq j\).

### 5.6 Conclusion

In summary, we studied the speed-accuracy trade-off embedded in the rate of synaptic plasticity in neural network models in dynamic inference tasks. To avoid this trade-off, we introduced the metaplastic synapses that can change the rates of plasticity [Fusi et al., 2005], along with a surprise detection network with multiple rates of synaptic plasticity. When an environment is stable, more and more synapses in the decision making network become less plastic to consolidate the information of reward sequence. This leads to an increase of integrating timescale at the population level. On the other hand, when an environment has suddenly changed, or an animal experienced an unexpected change in the environment [Yu and Dayan, 2005], the synapses rapidly
increase their plasticity, thanks to a surprise signal sent by the surprise detection network. As a result, the decision making network can adapt to a new environment quickly and start to consolidate new information about the environment. We also show that our model can reproduce some experimental results and has a significant computational advantage in a complex environment. Our work provides biophysical insights of highly adaptive behaviors that are becoming to be revieled in recent experiments.
Figure 5.1: Speed accuracy trade-off embedded in synaptic plasticity. The dotted line is the ideal choice probability and the colored lines are simulation results. A With less plastic synapses with $\alpha = 0.002$, the variance of estimation can be small; however, it becomes hard to adapt to a quick change in the environment. B On the other hand, with very plastic synapses with $\alpha = 0.2$, the estimation fluctuates severely, although the adaptation to a new environment is fast.
CHAPTER 5. MODELING FLEXIBLE VALUE-BASED DECISION MAKING BY NEURAL NETWORK MODELS.

Figure 5.2: Model. A Decision making network. Decision is made based on the competition between two excitatory target populations through an inhibitory population. The competition leads to a winner take all process determined by the strength of plastic synapses connecting the input population to the two target populations. After each trial, those synapses are modified according to the learning rule. B Cascade model of synapses used for the decision making network. The synaptic strength is assumed to be binary (depressed and potentiated); but there are meta-plastic states associated with these strength. The transition probability of changing strength is denoted by $\alpha_i$’s, while the transition probability of changing plasticity is denoted by $p_i$’s. As synapses go down to deeper states, they become less plastic and become harder to make transitions. C Binary synapses with different plasticity used for surprise detection network. Synapses have their unique fixed transition rates between the depressed state and the potentiated state.
Figure 5.3: Cascade model of synapses with a surprise network can improve the estimation performance under the speed-accuracy trade-off. A Cascade model of synapses can reduce the fluctuation of estimation as staying in the same context; however, this also makes difficult to learn a new environment. B With a surprise signal from the surprise detection network, the decision making network can achieve both the reduction in fluctuations and the quick adaptation to a new environment. The network parameters are taken as $\alpha_i = 0.2^i$, $p_i = 0.2^i$, $T = 0.1$, $\gamma = 0$ and $m = 10$, while the total baiting probability is set to 0.4 and the baiting contingency is 9 : 1 on the variable interval schedule. For B the surprise signal is sent when $P < 0.05$, which set the relevant transition rates $\alpha_i = \alpha_1$ as long as $P < 0.05$. 
Figure 5.4: Surprise network detects a change in environment and increases plasticity in the decision making network. **A** How surprise detection works. Reward history, a sequence of 0 or 1, is integrated by synapses in the surprise signal with different plasticities (different timescales). The surprise detection network computes the averaged difference between two reward rates on different timescales, which is called expected uncertainty. The actual current difference between the two timescales of reward rates (unexpected uncertainty) is compared with the expected uncertainty and it produces a surprise signal when the unexpected uncertainty significantly exceeds the expected uncertainty. **B** As a result of the surprise signal, the decision making network increases its synaptic plasticity, or the transition rates to the more plastic states increase when the network receives the surprise signal.
Figure 5.5 (following page): Model’s effecting learning rate changes depending on the volatility of environment. **A** Learning rate (red) changes depending on volatility as observed in experiments. Dotted lines denote change points of contingencies. The model is simulated on a variable rate schedule and with the total reward rate 1 and the contingency is 9 : 1 or 1 : 9 with block size of 30 trials (volatile phase) and 400 trials (stable phase). The network parameters are taken as $\alpha_i^r = 0.5^i$, $\alpha_{nr}^i = 0.5^{i+1}$, $p_r^i = 0.5^i$, $p_{nr}^i = 0.5^{i+1}$, $T = 0.1$, and $\gamma = 1$, $m = 12$, and $h = 0.05$. **B** Learning rates are self-tuned to the optimal depending on the volatility of the environment. Solid line is the averaged effective learning rate in constant block size. Dotted line is the optimal single leaning rate. Note that learning rate of cascade model constantly changes during a single block, which creates a slight deviation from the optimal for small and large block size; however this is an artifact of taking average of entire part of blocks. The model is simulated on a VR schedule and with the total reward rate 1 and the contingency is 9 : 1 or 1 : 9 with a fixed block size. The network parameters are taken as $\alpha_r^i = 0.5^{i-1}$, $\alpha_{nr}^i = 0.5^{i-1}$, $p_r^i = 0.5^{i-1}$, $p_{nr}^i = 0.5^{i-1}$, $T = 0.1$, and $\gamma = 1$, $m = 20$, and $h = 0.05$ for the cascade model. For the single timescale model, $\alpha_r^1 = \alpha_{nr}^1$, $\gamma = 1$ and $T = 0.1$. 
Figure 5.6 (following page): Cascade model with surprise signal performs better than single timescale models in an environment that changes on multiple timescales. The task is a four-armed bandit task with blocks of 10 trials and 10000 trials on a VR schedule with the total reward rate = 1. The total number of blocks is set to 1000 : 1. In a given block, one of the targets has the reward probability of 0.8, while the others have 0.2. The network parameters are taken as \( \alpha_r^i = 0.5^i, \alpha_{nr}^i = 0.5^i, p_r^i = 0.5^i, p_{nr}^i = 0.5^i, T = 0.1, \) and \( \gamma = 1, m = 4, \) and \( h = 0.005 \) for the cascade model. The initial state is set to be the most plastic synapses, where all the synapses occupy the most plastic states. For the single timescale model, the parameters are set as \( \alpha_r^1 = \alpha_{nr}^1, \gamma = 1 \) and \( T = 0.1. \)
CHAPTER 5. MODELING FLEXIBLE VALUE-BASED DECISION MAKING BY NEURAL NETWORK MODELS.

Task with two block sizes

Harvesting efficiency

Plasticity p

fixed timescale

cascade model + surprise signal
Figure 5.7 (following page): Distribution of synaptic plasticity is required when environment changes on multiple timescales. In this task, a subject needs to find the most rewarding target (shown here in red, although the color is not shown to the subject) from many targets. The most rewarding target with a reward probability of 0.9 (the others with 0.1) changes time to time but within the 'hot spots' indicated by yellow targets (again, the color is now shown to the subject). In this task, having a distribution on synaptic plasticity over a wide range of timescales is significantly beneficial. This is because the less plastic synapses can infer the location of hot spots, while the more plastic synapses can track the local changes of the most rewarding target. Task is on a VR schedule with blocks of 100 trials. In this simulation, the two timescale model is not the cascade model; instead the model has fixed populations of synapses with two different plasticities. The single timescale model is optimized, while the two timescales model is not: $\alpha^1 = 0.5, \alpha^2 = 0.001, T = 0.05$ and the synaptic population is the same for both timescales.
CHAPTER 5. MODELING FLEXIBLE VALUE-BASED DECISION MAKING BY NEURAL NETWORK MODELS.
Chapter 6

Concluding remarks

In this thesis we studied the nature of adaptive decision making and its possible neural mechanism by using the neural network models with different types of plastic synapses. We firstly showed that qualitatively different types of behaviors can be obtained from the same neural network model with plastic synapses that learns the values of actions from the reward history. We also found that the network model faces a trade-off in the dynamic choice task: an accurate estimation of stable environment and a quick adaptation to a changing environment. This problem, formulated as the speed-accuracy trade-off or the bias-variance trade-off, is shown to be overcome by introducing multiple timescales of memory on the reward history information, modeled by the metaplastic synaptic plasticity models. We found that the monkeys performing a foraging task were strongly influenced by the multiple timescales
CHAPTER 6. CONCLUDING REMARKS

of reward history. We also found that the influence from different timescales was tuned over multiple sessions to improve their harvesting performance. We showed that the observed tuning of weights were qualitatively consistent with the predictions of the neural network model with the plastic synapses that can change the rates of plasticity. We further investigated this network model and we showed that the network can optimally tune the reward integration timescales on a trial-by-trial basis if we introduce an additional network to send a surprise signal. The model’s adaptive behaviors were consistent with recent experimental findings. Our work provides an insight into the neural circuit mechanism of the highly adaptive learning and decision behaviors that are observed in experiments.
Bibliography


[Lundstrom et al., 2010] Brian Nils Lundstrom, Adrienne L Fairhall, and Miguel Maravall. Multiple timescale encoding of slowly varying whisker


Appendix A

Probability estimation of binary sequence over blocks of trials with multiple timescale models

Here we describe the bias-variance trade-off in a probability estimation of biased sequence using a simple example, and we show how the model with multiple timescales can perform and create undermatching. This chapter entirely owes Yashar Ahmadian.

A.1 Task and constraints

Consider a subject observes the time series of the flips of a coin ($s_t$ with heads $s = 1$, tails $s = 0$), where the head probability (bias of the coin), $p$, is constant
over blocks of $T$ trials. The bias changes randomly from one block to the next, with the new $p$ independently sampled from some distribution, $\pi(\cdot)$, on the interval $[0, 1]$ (our default choice will be the uniform distribution on $[0, 1]$).

The goal of the observer is to estimate the current head probability, $p$, while the observer has two timescales of sequence integrators. More precisely, the probability estimate, $v_t$ at time $t$, is given by

$$v_t = (1 - \alpha)v_{1,t} + \alpha v_{2,t}$$  \hspace{1cm} (A.1)

where $0 \leq \alpha \leq 1$, and $v_{1,t}$ and $v_{2,t}$ are leaky integrations of the recent history of $s_t$ over the time scales $\tau_1$ (fast) and $\tau_2$ (slow), respectively, i.e.

$$v_{i,t} = (1 - q_i)v_{i,t-1} + q_is_t$$ \hspace{1cm} (A.2)

where the learning rates, $q_i$ are the inverse of time constants. Solving (in steady state, i.e. after many many trials and blocks) we have

$$v_{i,t} = q_i \sum_{n=0}^{\infty} (1 - q_i)^n s_{t-n}. \hspace{1cm} (A.3)$$

Here we can imagine the time-scales, $q_1$ and $q_2$, as fixed hardware constraints, but consider $\alpha$ to be flexible. We want to find the optimal $\alpha$ leading to the minimum possible average square error for the estimator $v_t$, i.e. the $\alpha$ that minimizes the long-time average of $(v_t - p)^2$, given the knowledge of block size $T$ and the internal time-scales, $\tau_i$’s.

We will adopt the following index notation. We use $t$ as the trial index, $n$ as the trial lag (into the past), and $k$ as the block index. We denote the
current block by \( k = 0 \), with \( k = 1, 2, \ldots \) indicating past blocks (so \( k \) is rather the block-lag into the past). Thus \( p_0 \) is the head probability of the current block, \( p_1 \) that of the previous block, and so on. We choose the time origin such that the first trial of the current block \( (k = 0) \) has \( t = 1 \), with trials in past blocks having zero or negative \( t \)'s.

We also adopt the following averaging notations. We denote the average of a quantity, conditioned on knowing the full sequence of block-probabilities \( p_{0:\infty} \), by \( \langle \cdot \rangle \), i.e.

\[
\langle X_t \rangle \equiv \mathbb{E}[X_t|p_{0:\infty}]. \tag{A.4}
\]

We indicate averaging over \( p_{0:\infty} \) by \( [\cdot]_\pi \), i.e.

\[
[X]_\pi \equiv \int X_t(p_{0:\infty}) \prod_{k=0}^{\infty} d\pi(p_k). \tag{A.5}
\]

Finally we indicate averaging over the duration of a block by a bar:

\[
\bar{X} \equiv \frac{1}{T} \sum_{t=1}^{T} X_t \tag{A.6}
\]

Thus we set out to calculate the long-run average square error, which in the above notation is given by

\[
\mathbb{E}[((v_t - p_0)^2)]_\pi \tag{A.7}
\]

and then find the optimal \( \alpha \) that minimizes this cost.

We start by evaluating \( \langle (v_t - p_0)^2 \rangle \) which can be decomposed in the stan-
APPENDIX A. PROBABILITY ESTIMATION ON MULTIPLE TIMESCALES

standard way, into the variance of the estimator and its squared bias, i.e.

\[
\langle (v_t - p_0)^2 \rangle = \langle (v_t - \langle v_t \rangle)^2 \rangle + \langle (\langle v_t \rangle - p_0)^2 \rangle \tag{A.8}
\]

\[
= \langle \delta v_t^2 \rangle + \langle (v_t) - p_0 \rangle^2. \tag{A.9}
\]

A.2 Variance

We will first look at the long-run average variance \(\langle \delta v_t^2 \rangle\). From Eq. (A.1), the conditional variance is given by

\[
\langle \delta v_t^2 \rangle = (1 - \alpha)^2 \langle \delta v_{1,t}^2 \rangle + \alpha^2 \langle \delta v_{2,t}^2 \rangle + 2\alpha(1 - \alpha) \langle \delta v_{1,t} \delta v_{2,t} \rangle \tag{A.10}
\]

where \(\delta v_{i,t} \equiv v_{i,t} - \langle v_{i,t} \rangle\). From Eq. (A.3), \(v_{i,t}\) is a linear combination of independent random variables, \(s_t\) (the latter are independent only when conditioning/fixing \(p_0;\infty\) of course), thus its variance is the sum of the variances of the terms in this sum. The variance of \(s_{t-n}\) in block \(k\) is given by

\[
\langle \delta s_{t-n}^2 \rangle = p_k(1 - p_k), \quad k = -\left\lfloor \frac{t - n - 1}{T} \right\rfloor \equiv k(t - n). \tag{A.11}
\]

Thus

\[
\langle \delta v_{i,t}^2 \rangle = q_i^2 \sum_{n=0}^{\infty} (1 - q_i)^{2n} p_{k(t-n)}(1 - p_{k(t-n)}). \tag{A.12}
\]

Similarly

\[
\langle \delta v_{1,t} \delta v_{2,t} \rangle = q_1 q_2 \sum_{n=0}^{\infty} (1 - Q)^n p_{k(t-n)}(1 - p_{k(t-n)}), \tag{A.13}
\]

where we defined

\[
1 - Q \equiv (1 - q_1)(1 - q_2). \tag{A.14}
\]
APPENDIX A. PROBABILITY ESTIMATION ON MULTIPLE TIMESCALES

Since \( p_k(1 - p_k) \) is the same in all blocks, hence independent of \( k(t - n) \), we can readily calculate the variance averaged over \( p_{0: \infty} \), by summing the infinite geometric series, obtaining

\[
\langle \delta v^2 \rangle_{t} = \frac{1}{2\tau_1 - 1} \left[ p(1 - p) \right]_\pi , \quad (A.15)
\]

\[
\langle \delta v_1, t \delta v_2, t \rangle = \frac{1}{\tau_1 + \tau_2 - 1} \left[ p(1 - p) \right]_\pi . \quad (A.16)
\]

where we used \( q_i^2 \sum_{n=0}^{\infty} (1 - q_i)^{2n} = \frac{q_i^2}{1-(1-q_i)^2} = \frac{q_i}{2-q_i} = \frac{1}{2\tau_i - 1} \) and \( q_1 q_2 \sum_{n=0}^{\infty} (1 - Q)^n = \frac{q_1 q_2}{1-(1-Q)} = \frac{q_1 q_2}{q_1 + q_2 - q_1 q_2} = \frac{1}{\tau_1 + \tau_2 - 1} \). In particular, \( \langle \delta v^2 \rangle_{t} \) is time independent:

\[
\langle \delta v^2 \rangle_{t} = \langle \delta v^2 \rangle_{t} = \left[ p(1 - p) \right]_\pi \left[ \frac{(1 - \alpha)^2}{2\tau_1 - 1} + \frac{\alpha^2}{2\tau_2 - 1} + \frac{2\alpha(1 - \alpha)}{\tau_1 + \tau_2 - 1} \right] . \quad (A.17)
\]

A.2.1 Variance conditional on \( p_0 \): transient behavior

For completeness, we will also calculate the variance conditional on \( p_0 \) as well, obtaining its full transient behavior throughout the block. That is, here we will only average over \( p_{1: \infty} \), but not over \( t \) and \( p_0 \). Going back to Eq. (A.12), we rewrite it by decomposing the sum into sums over blocks:

\[
\langle \delta v^2 \rangle_{t} = q_i^2 \left\{ \sum_{n=0}^{t-1} (1 - q_i)^{2n} p_0(1 - p_0) + \sum_{k=1}^{\infty} p_k(1 - p_k) \sum_{n=t+(k-1)T}^{t+kT-1} (1 - q_i)^{2n} \right\} \quad (A.18)
\]
APPENDIX A. PROBABILITY ESTIMATION ON MULTIPLE TIMESCALES

It helps to rewrite this in the form

\[
\langle \delta v_{i,t}^2 \rangle = q_i^2 \left\{ \sum_{n=0}^{\infty} (1 - q_i)^{2n} p_0 (1 - p_0) + \sum_{n=t+1}^{\infty} \sum_{k=1}^{(k-1)T} (1 - q_i)^{2n} \right\}
\]

(A.19)

(the first term pretends that the probability was \( p_0 \) in the entire past, and the second term corrects for this by adding the difference of the variances accumulated over previous blocks contributed by the true probability, \( p_k \), and the current one, \( p_0 \), respectively). By the geometric series formula the sum over block \( k \) is given by

\[
\sum_{n=t+(k-1)T}^{t+kT-1} (1 - q_i)^{2n} = (1 - q_i)^{2t+2(k-1)T} \sum_{n=0}^{T-1} (1 - q_i)^{2n}
\]

(A.20)

\[
= (1 - q_i)^{2t+2(k-1)T} \frac{1 - (1 - q_i)^{2T}}{1 - (1 - q_i)^2}
\]

(A.21)

\[
= (1 - q_i)^{2t} \frac{1 - (1 - q_i)^{2T}}{1 - (1 - q_i)^2} (1 - q_i)^{2T(k-1)}.
\]

(A.22)

Using \( \frac{q_i^2}{1 - (1 - q_i)^2} = \frac{q_i}{2 - q_i} = \frac{1}{2 \tau_i - 1} \), we then have

\[
\langle \delta v_{i,t}^2 \rangle = \frac{1}{2 \tau_i - 1} \left\{ p_0 (1 - p_0) \right.
\]

\[
+ (1 - q_i)^{2t} \frac{\sum_{k=0}^{\infty} (1 - q_i)^{2Tk} [p_{k+1}(1 - p_{k+1}) - p_0(1 - p_0)]}{\sum_{k=0}^{\infty} (1 - q_i)^{2Tk}} \left\}
\]

(A.23)
Similarly for $\langle \delta v_{1,t} \delta v_{2,t} \rangle$ we have

\[
\langle \delta v_{1,t} \delta v_{2,t} \rangle = q_1 q_2 \left\{ \sum_{n=0}^{\infty} (1 - Q)^n p_0 (1 - p_0) \right. \\
+ \sum_{k=1}^{\infty} [p_k (1 - p_k) - p_0 (1 - p_0)] \sum_{n=t+(k-1)T}^{t+kT-1} (1 - Q)^n \right\} \\
= \frac{1}{\tau_1 + \tau_2 - 1} \left\{ p_0 (1 - p_0) \right. \\
+ (1 - Q)^t (1 - (1 - Q)^T) \sum_{k=0}^{\infty} (1 - Q)^T k [p_{k+1} (1 - p_{k+1}) - p_0 (1 - p_0)] \right\} \\
= \frac{1}{\tau_1 + \tau_2 - 1} \left\{ p_0 (1 - p_0) \right. \\
+ (1 - Q)^t \sum_{k=0}^{\infty} (1 - Q)^T k [p_{k+1} (1 - p_{k+1}) - p_0 (1 - p_0)] \right\} \\
\sum_{k=0}^{\infty} (1 - Q)^T k \right\}
\]

where we used $\frac{q_1 q_2}{1 - (1 - Q)} = \frac{q_1 q_2}{q_1 + q_2 - q_1 q_2} = \frac{1}{\tau_1 + \tau_2 - 1}$. Averaging over $p_1: \infty$ and summing the infinite geometric series over blocks, combining contributions as in Eq. (A.10), and using Eq. (A.17), we then obtain

\[
[\langle \delta v_t^2 \rangle]_\pi = \frac{p_0 (1 - p_0)}{[p(1 - p)]_\pi} \left[ [\langle \delta v_t^2 \rangle]_\pi \right. \\
+ ([p(1 - p)]_\pi - p_0 (1 - p_0)) \left. \left[ (1 - q_1)^{2t} \left( \frac{1 - \alpha}{2 \tau_1 - 1} \right) \right. \right. \\
+ (1 - q_2)^{2t} \left( \frac{\alpha^2}{2 \tau_2 - 1} \right) \left. + (1 - q_1)^t (1 - q_2)^t \left( \frac{2 \alpha (1 - \alpha)}{\tau_1 + \tau_2 - 1} \right) \right. \right\}
\]

\[(A.26)\]

**Interpretation:** the first line gives the steady state value of the variance in the current block if it was infinitely long, and the second line gives the transient memory of variance from previous trials, which wears off for $t \gg \tau_2$. It starts from a value equal to the average variance $[\langle \delta v_t^2 \rangle]_\pi = [p(1 - p)]_\pi \left[ \frac{(1 - \alpha)^2}{2 \tau_1 - 1} + \frac{\alpha^2}{2 \tau_2 - 1} + \frac{2 \alpha (1 - \alpha)}{\tau_1 + \tau_2 - 1} \right]$.  

APPENDIX A. PROBABILITY ESTIMATION ON MULTIPLE TIMESCALES

at \( t = 0 \) and eventually (given an infinitely long current block) relaxes to its steady-state value based on the current \( p_0 \) as opposed to the average, i.e. to

\[
\begin{align*}
p_0(1 - p_0) \left[ \frac{(1-a)^2}{2\tau_1-1} + \frac{\alpha^2}{2\tau_2-1} + \frac{2\alpha(1-a)}{\tau_1+\tau_2-1} \right].
\end{align*}
\]

A.3 Squared bias

First let us calculate \( \langle v_{i,t} \rangle \). Using the notation of Eq. (A.11), from Eq. (A.3) we have

\[
\langle v_{i,t} \rangle = q_i \sum_{n=0}^{\infty} (1 - q_i)^n p_{k(t-n)}
\]

(A.27)

Again we can decompose this over blocks:

\[
\langle v_{i,t} \rangle = q_i \sum_{n=0}^{t-1} (1 - q_i)^n p_0 + \sum_{k=1}^{\infty} p_k \sum_{n=t+(k-1)T}^{t+kT-1} (1 - q_i)^n
\]

(A.28)

and again it helps to rewrite this as

\[
\begin{align*}
\langle v_{i,t} \rangle &= q_i \left\{ \sum_{n=0}^{\infty} (1 - q_i)^n p_0 + \sum_{k=1}^{\infty} (p_k - p_0) \sum_{n=t+(k-1)T}^{t+kT-1} (1 - q_i)^n \right\}. 
\end{align*}
\]

(A.29)
APPENDIX A. PROBABILITY ESTIMATION ON MULTIPLE TIMESCALES

Noting that \( q_i \sum_{n=0}^{\infty} (1 - q_i)^n = 1 \), for the bias component, \( b_{i,t} \equiv \langle v_{i,t} \rangle - p_0 \), we obtain

\[
b_{i,t} \equiv \langle v_{i,t} \rangle - p_0 \quad \text{(A.30)}
\]

\[
= q_i \sum_{k=1}^{\infty} (p_k - p_0) \sum_{n=t+(k-1)T}^{t+kT-1} (1 - q_i)^n \
\quad \text{(A.31)}
\]

\[
= q_i (1 - q_i)^t \sum_{k=0}^{\infty} (p_{k+1} - p_0) (1 - q_i)^{Tk} \sum_{n=0}^{T-1} (1 - q_i)^n \
\quad \text{(A.32)}
\]

\[
= (1 - q_i)^t \left[ 1 - (1 - (1 - q_i)^T) \right] \sum_{k=0}^{\infty} (p_{k+1} - p_0) (1 - q_i)^{Tk} \
\quad \text{(A.33)}
\]

\[
= (1 - q_i)^t \sum_{k=0}^{\infty} (1 - q_i)^{Tk} (p_{k+1} - p_0) \sum_{k=0}^{\infty} (1 - q_i)^{Tk} \
\quad \text{(A.34)}
\]

\[
= -(1 - q_i)^t \delta p_0 + (1 - q_i)^t \left[ 1 - (1 - q_i)^T \right] \sum_{k=0}^{\infty} (1 - q_i)^{Tk} \delta p_{k+1} \
\quad \text{(A.35)}
\]

\[
= -(1 - q_i)^t \delta p_0 + (1 - q_i)^t \left[ 1 - (1 - q_i)^T \right] \sum_{k=0}^{\infty} (1 - q_i)^{Tk} \delta p_{k+1} \
\quad \text{(A.36)}
\]

where we defined

\[
\delta p_k \equiv p_k - [p]_\pi. \quad \text{(A.37)}
\]

Note that we can write the bias, Eq. (A.36), in the form

\[
b_t \equiv \langle v_t \rangle - p_0 = \sum_{k=0}^{\infty} A_k(t) \delta p_k \
\quad \text{(A.38)}
\]

where we defined

\[
A_0(t) \equiv -(1 - \alpha)(1 - q_1)^t + \alpha (1 - q_2)^t \quad \text{(A.39)}
\]

\[
A_k(t) \equiv (1 - \alpha)B_{1,k}(t) + \alpha B_{2,k}(t) \quad (k > 0). \quad \text{(A.40)}
\]
and

\[ B_{i,k}(t) \equiv (1 - q_i)^t \left[ 1 - (1 - q_i)^T \right] (1 - q_i)^T^{(k-1)} \quad (k > 0, \ i = 1, 2). \]  

(A.41)

The bias squared is then given by

\[ b_t^2 = \sum_{k=0}^{\infty} \sum_{j=0}^{\infty} A_k(t) A_j(t) \delta p_k \delta p_j = \sum_{k=0}^{\infty} A_k^2(t) \delta p_k^2 + \sum_{k \neq j} A_k(t) A_j(t) \delta p_k \delta p_j. \]  

(A.42)

Since \( \delta p_k \) are zero-mean independent variables, averaging over them kills the second, off-diagonal term (this is true even if we don’t average over \( p_0 \)) in the above expression. The bias squared averaged over \( p_k \) in the previous blocks (but not on \( p_0 \)) is thus given by

\[
[b_t^2|p_0]_\pi = \delta p_0^2 A_0^2(t) + \left[ \delta p^2 \right]_\pi \sum_{k=1}^{\infty} A_k^2(t) \\
= (p_0 - [p]_\pi)^2 \left[ (1 - \alpha)(1 - q_1)^t + \alpha(1 - q_2)^t \right]^2 \\
+ \left[ \delta p^2 \right]_\pi \sum_{k=1}^{\infty} \left[ (1 - \alpha)^2 B_{1,k}^2(t) \right. \\
\left. + \alpha^2 B_{2,k}^2(t) + 2\alpha(1 - \alpha)B_{1,k}(t)B_{2,k}(t) \right]  
\]  

(A.43)

Now we have

\[
\sum_{k=1}^{\infty} B_{i,k}^2(t) = (1 - q_i)^{2t} \left[ 1 - (1 - q_i)^T \right]^2 \sum_{k=0}^{\infty} (1 - q_i)^{2Tk}  
\]  

(A.45)

\[
= (1 - q_i)^{2t} \frac{[1 - (1 - q_i)^T]^2}{1 - (1 - q_i)^{2T}}  
\]  

(A.46)

\[
= (1 - q_i)^{2t} \frac{1 - (1 - q_i)^T}{1 + (1 - q_i)^T}  
\]  

(A.47)
and (using $1 - Q \equiv (1 - q_1)(1 - q_2)$)

$$
\sum_{k=1}^{\infty} B_{1,k}(t)B_{2,k}(t) = (1 - Q)^t \left[1 - (1 - q_1)T \right] \left[1 - (1 - q_2)T \right] \sum_{k=0}^{\infty} (1 - Q)^{Tk}
$$

$$
= (1 - Q)^t \frac{\left[1 - (1 - q_1)T \right] \left[1 - (1 - q_2)T \right]}{1 - (1 - q_1)T(1 - q_2)T}
$$

yielding

$$
[b_i^2|p_0] = (p_0 - [p]_\pi)^2 \left[(1 - \alpha)(1 - q_1)^t + \alpha(1 - q_2)^t\right]^2
$$

$$
+ \left[\delta p^2\right] \left[(1 - \alpha)^2 \frac{1}{1 + (1 - q_1)^T(1 - q_1)^{2t}} + \alpha^2 \frac{1}{1 + (1 - q_2)^T(1 - q_2)^{2t}} + 2\alpha(1 - \alpha) \frac{1 - (1 - q_1)^T(1 - q_2)^t}{1 - (1 - q_1)T(1 - q_2)T} \right],
$$

for the transient behavior of conditional average bias squared in the current block.

Averaging over $p_0$ yields

$$
[b_i^2] = 2 \left[\delta p^2\right] \left[(1 - \alpha)^2 \frac{1}{1 + (1 - q_1)^T(1 - q_1)^{2t}} + \alpha^2 \frac{1}{1 + (1 - q_2)^T(1 - q_2)^{2t}} + 2\alpha(1 - \alpha) \frac{1 - (1 - q_1)^T(1 - q_2)^t}{1 - (1 - q_1)T(1 - q_2)T} \right]
$$

Finally, to average over $t$ ranging over the block, we use

$$
\frac{1}{T} \sum_{t=1}^{T} (1 - q_i)^{2t} = \frac{(1 - q_i)^2 1 - (1 - q_i)^{2T}}{T 1 - (1 - q_i)^2} = \frac{1 - (1 - q_i)^{2T}}{T [(1 - q_i)^{-2} - 1]}
$$

$$
\frac{1}{T} \sum_{t=1}^{T} (1 - Q)^t = \frac{1 - Q 1 - (1 - Q)^{T}}{T 1 - (1 - Q)} = \frac{1 - (1 - Q)^T}{T [(1 - Q)^{-1} - 1]}
$$

to obtain

$$
[b_i^2] = \left[\delta p^2\right] \left[(1 - \alpha)^2 \frac{1 - (1 - q_1)^T}{T \left[(1 - q_1)^{-2} - 1\right]} + \alpha^2 \frac{1}{T \left[(1 - q_2)^{-2} - 1\right]} + 2\alpha(1 - \alpha) \frac{2 - (1 - q_1)^T - (1 - q_2)^T}{T \left[(1 - Q)^{-1} - 1\right]} \right].
$$
APPENDIX A. PROBABILITY ESTIMATION ON MULTIPLE TIMESCALES

In the regime where \( q_2 \ll T^{-1} \ll q_1 \ll 1 \) (or \( \tau_2 \gg T \gg \tau_1 \gg 1 \)), we have approximately \((1 - q_1)^T \approx 0\), \((1 - q_2)^T \approx 1 - q_2 T\) and \([1 - q_i]^{-2} - 1 \approx 2q_i\)
and \(([1 - Q]^{-1} - 1 \approx q_1 + q_2\), yielding

\[
\bar{b}_t^2 = [\delta p^2]_\pi \left[ (1 - \alpha)^2 \frac{T_1}{T} + \alpha^2 + 2\alpha(1 - \alpha) \left( \frac{T_1}{T} + \frac{T_1}{T_2} \right) \right],
\]

(A.54)

and in the limit \( \tau_2, T \to \infty \):

\[
\bar{b}_t^2 = [\delta p^2]_\pi \alpha^2.
\]

(A.55)

A.4 Average squared error, optimal \( \alpha \), and undermatching

The long-run average squared error is the sum of average variance and average bias squared and thus from Eqs. (A.17) and (A.53) is given by

\[
\bar{(v_t - p_0)^2} = \bar{(\delta v_t)^2} + \bar{b}_t^2 = C_1 (1 - \alpha)^2 + C_2 \alpha^2 + 2C_3 \alpha(1 - \alpha) \quad \text{(A.56)}
\]

where we defined

\[
C_1 = \frac{[p(1 - p)]_\pi}{2\tau_1 - 1} + [\delta p^2]_\pi \frac{1 - (1 - q_1)^T}{T} \frac{1}{[(1 - q_1)^{-2} - 1]}, \quad \text{(A.57)}
\]

\[
C_2 = \frac{[p(1 - p)]_\pi}{2\tau_2 - 1} + [\delta p^2]_\pi \frac{1 - (1 - q_2)^T}{T} \frac{1}{[(1 - q_2)^{-2} - 1]}, \quad \text{(A.58)}
\]

\[
C_3 = \frac{[p(1 - p)]_\pi}{\tau_1 + \tau_2 - 1} + [\delta p^2]_\pi \frac{2}{T} \frac{(1 - q_1)^T - (1 - q_2)^T}{[(1 - q_1)^{-1}(1 - q_2)^{-1} - 1]} \quad \text{(A.59)}
\]
Here, in each line the first term is the contribution of the variance and the second is the contribution of the average squared bias. Note that in general

\[
[p(1 - p)]_\pi = [p]_\pi (1 - [p]_\pi) - [\delta p^2]_\pi. \quad (A.60)
\]

In particular, for the case where \(\pi(\cdot)\) is the uniform distribution on \([0, 1]\), we have

\[
[\delta p^2]_\pi = \left[(p - \frac{1}{2})^2\right]_\pi = \frac{1}{12} \quad (A.61)
\]

\[
[p(1 - p)]_\pi = \frac{1}{4} - [\delta p^2]_\pi = \frac{1}{6}. \quad (A.62)
\]

To find the optimal \(\alpha\) we have to set the derivative of Eq. (A.56) w.r.t. \(\alpha\) to zero. The latter is proportional to

\[
C_1(\alpha - 1) + C_2\alpha + C_3(1 - 2\alpha), \quad (A.63)
\]

and setting it equal to zero yields

\[
\alpha_* = \frac{C_1 - C_3}{C_1 + C_2 - 2C_3}. \quad (A.64)
\]

We can use Eq. (A.54), to simplify Eq. (A.57) in the regime \(q_2 \ll T^{-1} \ll q_1 \ll 1\) (or \(\tau_2 \gg T \gg \tau_1 \gg 1\)), obtaining\(^1\)

\[
C_1 \approx \frac{[p(1 - p)]_\pi}{2\tau_1 - 1} + [\delta p^2]_\pi \frac{\tau_1}{T} \quad (A.65)
\]

\[
C_2 \approx \frac{[p(1 - p)]_\pi}{2\tau_2 - 1} + [\delta p^2]_\pi \quad (A.66)
\]

\[
C_3 \approx \frac{[p(1 - p)]_\pi}{\tau_1 + \tau_2 - 1} + [\delta p^2]_\pi \left(\frac{\tau_1}{T} + \frac{\tau_1}{\tau_2}\right). \quad (A.67)
\]

\(^1\)To be really consistent in the approximations, the first terms on the rights sides of Eq. (A.65) must also be expanded.
We see that the largest contribution to average error, which is $O(1)$, comes from the bias squared contributed by the slow time scale (the second term in $C_2$). After that we have the contribution of the fast time scale to variance (first term in $C_1$) which is $O(\tau_1^{-1})$ and smaller. For this reason, for realistic underlying time-scales, the optimal $\alpha$’s will turn out to mainly optimize the squared bias, and hence will be small.

It is much easier to derive these results in the extreme limit $\tau_2, T \to \infty$ (keeping $\tau_2 \gg T$). Firstly, in this case, given that $v_2$ is a very long-term average of $s_t$, its value is always very close to the long term average of $p$, i.e. $[p]_\pi$, with small fluctuations, $\delta v_2$, of the order of $1/\sqrt{\tau_2}$. Thus we can ignore the latter and safely write

$$v_{2,t} \approx [p]_\pi.$$  \hspace{1cm} (A.68)

In particular, it is only $v_{1,t}$ which contributes to the variance:

$$\langle \delta v_t^2 \rangle \approx (1 - \alpha)^2 \langle \delta v_{1,t}^2 \rangle.$$  \hspace{1cm} (A.69)

Furthermore, given that $\tau_1 \ll T$, the main contribution to the averages of $\langle v_{1,t} \rangle$ or $\langle \delta v_{1,t}^2 \rangle$ over $t$ running from 1 : $T$ comes from $t$’s within the current block that are much larger than $\tau_1$ (i.e., we can ignore the transient behavior of $v_{1,t}$ at the beginning of the block and only consider its steady-state behavior). This means that in Eqs. (A.12) and (A.27), we can safely replace $p_{k(t-n)}$ with $p_0$, the head probability in the current block. The geometric series thus become
APPENDIX A. PROBABILITY ESTIMATION ON MULTIPLE TIMESCALES

infinite and we obtain

\[ \langle \delta v^2_{1,t} \rangle \approx q_1^2 \sum_{n=0}^{\infty} (1 - q_1)^{2n} p_0 (1 - p_0) = \frac{p_0 (1 - p_0)}{2 \tau_1 - 1}. \]  \hspace{1cm} (A.70)

\[ \langle v_{1,t} \rangle \approx q_1 \sum_{n=0}^{\infty} (1 - q_1)^n p_0 = p_0 \]  \hspace{1cm} (A.71)

Averaging Eq. (A.70) over \( \pi \), and using Eq. (A.69) we obtain

\[ \left[ \langle \delta v^2_{1} \rangle \right]_{\pi} \approx (1 - \alpha)^2 \left[ \frac{p(1 - p)}{2 \tau_1 - 1} \right]_{\pi}. \]  \hspace{1cm} (A.72)

For the full bias we have \( b_t = \langle v_t \rangle - p_0 = (1 - \alpha) \langle v_{1,t} \rangle + \alpha \langle v_{2,t} \rangle - p_0 \), which by Eq. (A.68) and (A.71), yields \( b_t = \alpha[p]_{\pi} - p_0 = -\alpha \delta p_0 \) (this yields \( 1 - \alpha \) for the undermatching slope, as an approximation to Eq. (A.80)). Thus

\[ \left[ b^2_t \right]_{\pi} \approx \alpha^2 \left[ \delta p^2 \right]_{\pi}. \]  \hspace{1cm} (A.73)

Finally for the average square error we obtain Eq. (A.56) with

\[ C_1 \approx \frac{\left[ p(1 - p) \right]_{\pi}}{2 \tau_1 - 1} \] \hspace{1cm} (A.74)

\[ C_2 \approx \left[ \delta p^2 \right]_{\pi} \] \hspace{1cm} (A.75)

\[ C_3 \approx 0. \] \hspace{1cm} (A.76)

A.4.1 Time-dependent undermatching slope

Going back to Eq. (A.36) for the bias, since the second term in Eq. (A.36) vanishes after averaging over \( p_{k+1} \), for the transient of bias conditional on \( p_0 \) but averaged over \( p_k \) in past blocks we obtain

\[ [v_t - p_0|p_0]_{\pi} = -(p_0 - [p]_{\pi}) \left[ (1 - \alpha)(1 - q_1)^t + \alpha(1 - q_2)^t \right]. \]  \hspace{1cm} (A.77)
which we can also write in a form corresponding to the slope of the matching law plot

\[
\frac{[v_t|p_0]_\pi - [p]_\pi}{p_0 - [p]_\pi} = 1 - (1 - \alpha)(1 - q_1)^t - \alpha(1 - q_2)^t. \tag{A.78}
\]

(assuming a “symmetric” distribution \(\pi(\cdot), [p]_\pi = \frac{1}{2}\)). In particular, when \(\tau_2 \gg T\) (or \(q_2 T \ll 1\)) \((1 - q_2)^t\) remains approximately equal to unity even for \(t = T\) (at the end of the block). Thus we have

\[
\frac{[v_t|p_0]_\pi - [p]_\pi}{p_0 - [p]_\pi} = (1 - \alpha) \left[1 - (1 - q_1)^t\right], \quad (\tau_2 \gg T). \tag{A.79}
\]

This shows that there is more undermatching at the beginning of the block, than at the end (where \((1 - q_1)^t \ll 1\), if \(\tau_1 \ll T\)). If we average this over the whole block we obtain for the block-averaged undermatching slope

\[
\frac{[v_t|p_0]_\pi - [p]_\pi}{p_0 - [p]_\pi} = (1 - \alpha) \left[1 - \frac{1 - (1 - q_1)^T}{Tq_1}\right], \quad (\tau_2 \gg T) \tag{A.80}
\]

\[
\approx (1 - \alpha) \left[1 - \frac{\tau_1}{T}\right]. \quad (\tau_2 \gg T \gg \tau_1) \tag{A.81}
\]