Learning with Surprise: Theory and Applications

THÈSE N^{O} 7418 (2016)

PRÉSENTÉE LE 9 DÉCEMBRE 2016
À LA FACULTÉ INFORMATIQUE ET COMMUNICATIONS
LABORATOIRE DE CALCUL NEUROMIMÉTIQUE (IC/SV)
PROGRAMME DOCTORAL EN INFORMATIQUE ET COMMUNICATIONS

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Mohammadjavad FARAJI

acceptée sur proposition du jury:

Prof. R. Urbanke, président du jury
Prof. W. Gerstner, directeur de thèse
Prof. W. Senn, rapporteur
Prof. K. E. Stephan, rapporteur
Prof. M. Gastpar, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE
Suisse
2016

Dedicated to my parents and my wife
for their everlasting love and support

Acknowledgements

I would like to express my sincere gratitude to my advisor Prof. Wulfram Gerstner for forging my scientific personality and for providing me an opportunity to grow. I am very grateful to him for the freedom he gave to me in identifying my research topics and for his continuous support, trust, and wise guidance throughout my PhD study.

I am indebted to Prof. Kerstin Preuschoff for her encouragement and enthusiasm about my work. She has taught me innumerable lessons and insights on the workings of academic research. Her technical and editorial advice was essential to the completion of this dissertation.

I would also like to thank my thesis committee members : Prof. Klaas Enno Stephan, Prof. Walter Senn, Prof. Michael Gastpar, and Prof. Rudiger Urbanke, for accepting to assess my thesis and for their insightful comments, which incented me to widen my research from various perspectives.

I am very thankful to all the LCN lab members, past and present, from Richard Naud (the first one who left the group after my arrival) to Bernd Illing (the last one who arrived in the group before I left) for their friendship, stimulating discussions, and for all the fun we have had in the last five years.

My foremost gratitude and respect go to my parents, for their unconditional love and dedication. I would like to deeply appreciate them for all the sacrifices they made for me, and for providing me an everlasting support for making progress. They receive my deepest gratitude and love for their dedication. I have no doubt that I am truly indebted to them for all the success I have achieved in my life.

I also appreciate my family, especially my brothers and sisters, for their friendship and support, and for being beside my parents, while their youngest son was far away. I am really grateful to all of them and I wish them all the best.

Especially, I express my deepest appreciation to my beloved wife Simin, for standing beside me in every single step towards this achievement. I cannot express my feelings about having such a wonderful partner in my life. She has always been my inspiration and motivation for making progress. I truly appreciate the love of my life, for believing in me, and for the encouragement she has always given to me for a brighter future.

Abstract

Everybody knows what it feels to be surprised. Surprise raises our attention and is crucial for learning. It is a ubiquitous concept whose traces have been found in both neuroscience and machine learning. However, a comprehensive theory has not yet been developed that addresses fundamental problems about surprise : (1) surprise is difficult to quantify. How should we measure the level of surprise when we encounter an unexpected event ? What is the link between surprise and startle responses in behavioral biology ? (2) the key role of surprise in learning is somewhat unclear. We believe that surprise drives attention and modifies learning ; but, how should surprise be incorporated, in general paradigms of learning ? and (3) can we develop a biologically plausible theory that explains how surprise can be neurally calculated and implemented in the brain ?

I propose a theoretical framework to address the above issues about surprise. There are three components to this framework : (1) a subjective confidence-adjusted measure of surprise, that can be used for quantification purposes, (2) a surprise-minimization learning rule that models the role of surprise in learning by balancing the relative contribution of new and old data for inference about the world, and (3) a surprise-modulated Hebbian plasticity rule that can be implemented in both artificial and spiking neural networks. The proposed online rule links surprise to the activity of the neuromodulatory system in the brain, and belongs to the class of neo-Hebbian plasticity rules.

My work on the foundations of surprise provides a suitable framework for future studies on learning with surprise. Reinforcement learning methods can be enhanced by incorporating the proposed theory of surprise. The theory could ultimately become interesting for the analysis of fMRI and EEG data. It may also inspire new synaptic plasticity rules that are under the simultaneous control of reward and surprise. Moreover, the proposed theory can be used to make testable predictions about the time course of the neural substrate of surprise (e.g., noradrenaline), and suggests behavioral experiments that can be performed on real animals for studying surprise-related neural activity.

Key words : Surprise, Multi-factor learning rules, Synaptic plasticity, Neuromodulation, Noradrenaline, Locus coeruleus, Spiking networks, Free energy, Maze exploration, Hidden Markov model, Expectation maximization, Gradient descent, Information theory, Bayesian inference.

Résumé

Tout le monde sait ce que cela fait d'être surpris. La surprise attire notre attention et est cruciale pour l'apprentissage. C'est un concept omniprésent dont les origines ont été fondées au travers des neurosciences et de l'apprentissage automatique. Cependant, aucune théorie globale n'a encore été développée abordant les problèmes fondamentaux concernant la surprise car : (1) la surprise est difficile à quantifier. Comment mesurer le niveau de surprise quand nous rencontrons un événement inattendu ? Quel est le lien entre la surprise et le sursaut en biologie comportementale ? (2) le rôle clé de la surprise dans l'apprentissage reste un peu flou. Nous croyons que la surprise attire l'attention et modifie l'apprentissage ; mais, comment la surprise doit-elle être incorporée dans les paradigmes d'apprentissage ? et (3) pouvons-nous développer une théorie biologiquement plausible qui explique comment la surprise peut être calculée et implémentée dans le cerveau ?

Je propose un cadre théorique pour aborder les questions mentionnées ci-dessus à propos de la surprise. Il existe trois composantes dans ce cadre théorique : (1) une mesure de surprise subjective ajustée en fonction de la confiance, qui peut être utilisée à des fins de quantification ; (2) une règle d'apprentissage de minimisation de la surprise qui modélise le rôle de la surprise dans l'apprentissage en équilibrant la contribution relative de données nouvelles et anciennes inférant le monde extérieur, et (3) une règle de plasticité Hebbienne modulée par la surprise qui peut être implémentée dans des réseaux de neurones artificiels et à impulsions. La règle proposée ici relie la surprise à l'activité du système de neuromodulation dans le cerveau, et appartient à la classe des règles de plasticité néo-Hebbienne.

Mon travail sur les fondements de la surprise fournit un cadre approprié pour les futures études sur l'apprentissage avec la surprise. Les méthodes d'apprentissage par renforcement peuvent être améliorées en incorporant la théorie proposée de la surprise. Cette théorie pourrait finalement être intéressante pour l'analyse des données d'IRMf et d'EEG. Il peut également inspirer de nouvelles règles de plasticité synaptique qui sont sous le contrôle simultané de la récompense et de la surprise. De plus, la théorie proposée peut être utilisée pour établir des prédictions pouvant être testées sur la dynamique des neuromodulateurs de la surprise (par exemple, la noradrénaline) et suggère des expériences comportementales pouvant être réalisées sur des animaux réels pour étudier l'activité neuronale liée à la surprise.

Mots clefs : Surprise, Règles d'apprentissage multi-facteurs, Plasticité synaptique, Neuromodulation, Noradrenaline, Locus coeruleus, Neurones à impulsions, Énergie libre, Exploration labyrinthe, Modèle de Markov caché, Espérance-maximisation, Algorithme du gradient, Théorie de l'information, Inférence Bayésienne.

Contents

Acknowledgements **i**

Abstract (English/Français) **iii**

List of figures **xi**

1 Introduction **1**
- 1.1 Surprise in neuroscience 1
  - 1.1.1 Neural correlate of surprise 2
- 1.2 Surprise in machine learning 3
  - 1.2.1 Implications of surprise 4
- 1.3 Multi-factor learning rules for synaptic plasticity 5
- 1.4 Structure of the thesis 6

2 A Mathematical Description of Surprise **9**
- 2.1 Probability-based surprise measures 10
  - 2.1.1 Shannon surprise 10
  - 2.1.2 Bayesian surprise 12
  - 2.1.3 Confidence-corrected surprise 12
- 2.2 Discussion 15
  - 2.2.1 Absolute z-score 15
  - 2.2.2 Binary or graded surprise? 15
  - 2.2.3 Subjective perception affects surprise 16
- 2.3 Materials and Methods 17
  - 2.3.1 The scaled likelihood is the posterior belief under the flat prior . 17
  - 2.3.2 The raw surprise increases with the Shannon surprise and the Bayesian surprise 17
  - 2.3.3 Absolute z-score is linked to probability-based surprise measures in a Gaussian setting 18

3 Balancing New Against Old Information: The Role of Surprise in Learning **21**
- 3.1 Introduction 21

3.2 Results 22
3.2.1 Surprise minimization: the SMiLe-rule 22
3.2.2 Surprise-modulated belief update 26
3.2.3 Simulations 27
3.3 Discussion 37
3.4 Materials and Methods 42

4 A Biologically Plausible 3-Factor Learning Rule from Gradient Descent Optimization 51
4.1 Introduction 52
4.2 Results 52
4.2.1 Functional gradient rule: Theory 52
4.2.2 Functional gradient rule: Applications 55
4.3 Neural implementation of the SMiLe rule 60
4.3.1 Neural network model 60
4.3.2 Simulation results 67
4.4 Discussion 70
4.5 Materials and Methods 73

5 Future Works 79
5.1 Dissociation between negative and positive surprise 79
5.2 Surprise in language of neurons 80
5.3 Relation between neuromodulators and the global factors 81
5.4 Interplay between surprise and different forms of uncertainty 81

A Palm Theory of Surprise 83
A.1 Description and repertoire 84

B Generalized Expected Utility Maximization for Decision Making 89
B.1 Derivation of the soft-max rule from expected utility maximization . . 91

C Learning Associations with a Neurally Computed Global Novelty Signal 93
C.1 A neural network model 93
C.2 Our measure of novelty 95
C.3 Simulation 96
C.4 Discussion 96

D What Hinders Bayesian Optimal Critic to Distinguish Similar Tasks? 101
D.1 Argument 1: using statistical decision theory 101
D.2 Argument 2: using Bayesian reasoning 103
D.3 Calculation of the risk functions 104

Contributions **111**

Bibliography **123**

Curriculum Vitae **125**

List of Figures

1.1 Hebbian vs. 3-factor plasticity rules 6

2.1 Confidence-corrected surprise 11

3.1 Constraint surprise minimization. 25
3.2 Gaussian mean estimation task 28
3.3 Maze exploration task 31
3.4 Time-series of relevant signals in the surprise-modulated belief update 33
3.5 True and estimated state transition probabilities in the maze exploration task 35
3.6 Comparison of surprise-modulated belief update with an online EM algorithm for the hierarchical Bayesian model 36
3.7 The estimation error in the maze exploration task 37

4.1 Neural network model for the maze-exploration task. 61
4.2 Inputs and outputs of the recognition network 63
4.3 Local and global traces required for plasticity 67
4.4 The estimation error and the modulatory signal in spiking network during maze-exploration task 69
4.5 The effect of number of samples in the recognition network and the prediction network on the learning performance 70

A.1 Description functions in Palm theory of surprise 85

C.1 Neural network model for novelty-based clustering 97

D.1 Mixture of two Gaussian distributions 106
D.2 Decision risks for optimal Bayesian critic 107
D.3 Bayesian decision for identification of two similar tasks 108
D.4 Bayesian decision for identification of two different tasks 109
D.5 Estimation uncertainty for different mean values 110

1 Introduction

A strong mathematical framework is required to elucidate how the brain *learns* to make decisions, to form new memories, and to organize behaviors. One of the crucial yet largely undetermined factors that affects learning and plasticity in the brain is *Surprise*. It is a widely used concept describing a range of phenomena from unexpected events to behavioral responses. Surprise can play a role in learning that is similar to that of reward, but surprise is obviously much less well understood than reward. Although traces of surprise and surprise-related concepts (such as novelty) have been found in both neuroscience and machine learning, a well-established theory that links conceptual as well as computational considerations of surprise to behavioral research and neuroscience is still missing.

Surprise is difficult to quantify. Neither is there a comprehensive theory that quantitatively links surprise to the startle responses in biology, nor is there an agreement on how surprise should affect learning speed or other parameters in statistical learning algorithms. Moreover, much less is known about how surprise is *neurally* calculated and how it affects plasticity in a biologically plausible way. The research described in this thesis aims in providing a *theory* of quantification, utilization, and neural implementation of surprise. The theory could ultimately become interesting as a correlator for fMRI or EEG, but could also influence learning theory.

1.1 Surprise in neuroscience

Behaviorally, surprise can be identified through startle responses [Kalat, 2012], which are vital for humans and animals. It manifests itself as physiological responses such as pupil dilation [Hess and Polt, 1960, Nassar et al., 2012] and tension in the muscles [Kalat, 2012]. Surprise is usually followed by emotions such as joy or confusion. Neurally, the P300 component of the event-related potential [Pineda et al., 1997, Missonnier et al., 1999] measured by electroencephalography is associated with the violation of expectation [Jaskowski et al., 1994, Kolossa et al., 2015]. Surprising events have been shown to influence the development of the sensory cortex [Fairhall et al., 2001], and to drive attention [Itti and Baldi, 2009]. It triggers associative learning [Schultz and Dickinson, 2000, Fletcher et al., 2001] and affects memory formation [Ranganath and Rainer, 2003, Hasselmo, 1999, Wallenstein et al., 1998].

Novel stimuli are often very surprising. Although surprise and novelty differ in some of their conceptual aspects, for many researchers they are considered as interchangeable concepts. Therefore, many interesting neuroscientific facts that have been discovered about novelty can also be associated to surprise and its importance in controlling learning and attention. For instance, neural responses have been reported to be larger for novel (i.e., surprising) stimuli than repeated and prolonged stimuli in cortical areas [Müller et al., 1999, Ulanovsky et al., 2003, Nelken et al., 1999] as well as subcortical structures [Solomon et al., 2004, Kennedy et al., 2003].

Memory retention, in both humans and animals, is enhanced when something novel happens [Takeuchi et al., 2016]. Recent findings, using optogenetic methods, show that neuronal firing in Locus coeruleus (LC, a brainstem neuromodulatory nuclei that releases Noradrenaline (NE) throughout the brain) is sensitive to environmental novelty, and its dopaminergic projection to the hippocampus modulates memory formation [Takeuchi et al., 2016]. Neurophysiological evidence suggests that Hippocampus plays a key role in novelty (and surprise) detection [Knight et al., 1996, Stern et al., 1996, Li et al., 2003]. Moreover, it has been hypothesized that seeking novel and surprising inputs during exploration of unknown environments is associated with a dopamine receptor gene [Ebstein et al., 1996, Lusher et al., 2001].

1.1.1 Neural correlate of surprise

There is ample evidence for a neural substrate of surprise. Existing measures of expectation violations such as absolute and variance-scaled reward prediction errors [Schultz et al., 1997, Schultz, 2016, Schultz, 2015], unexpected uncertainty [Yu and Dayan, 2005] and risk prediction errors [Preuschoff et al., 2008] have been linked to different neuromodulatory systems. Among those, the *noradrenergic system* has emerged as a prime candidate for signaling unexpected uncertainty and surprise: noradrenergic neurons respond to unexpected changes such as the presence of a novel stimulus, unexpected pairing of a stimulus with a reinforcement signal during conditioning, and reversal of the contingencies [Sara and Segal, 1991, Sara et al., 1994, Vankov et al., 1995, Aston-Jones et al., 1997]. The P300 component of the event-related potential [Pineda et al., 1997, Missonnier et al., 1999] which is associated with novelty [Donchin et al., 1978] and surprise [Jaskowski et al., 1994] is modulated by Noradrenaline (NE). It also modulates pupil size [Costa and Rudebeck, 2016, Preuschoff et al., 2011] as a physiological response to surprise.

The dynamics of the noradrenergic system are fast enough to quickly respond to unexpected events [Rajkowski et al., 1994, Clayton et al., 2004, Bouret and Sara, 2004], a functional requirement for surprise to control learning; see [Sara, 2009, Bouret and Sara, 2005, Aston-Jones and Cohen, 2005] for a review. Locus coeruleus (LC) and anterior cingulate cortex (ACC, a part of the brain's limbic system) have been frequently reported as neural structures included in the surprise-related circuitry of the brain [Hayden et al., 2011, Bouret and Sara, 2005]. For instance, the neural responses to rewards in macaque dorsal anterior cingulate cortex (ACC) enhances when the outcome is surprising [Hayden et al., 2011]. It has been shown that locus coeruleus-norepinephrine (LC-NE) system regulates cognitive performance in cortical areas [Usher et al., 1999] and influences neuronal responses via gain modulation [Aston-Jones and Cohen, 2005]. ACC, whose functional role in conflict/error detection and performance monitoring has been frequently reported [Carter et al., 1998, Bush et al., 2000], sends prominent inputs to LC neurons [Aston-Jones and Cohen, 2005], consistent with aforementioned hypothesis that links surprise to LC and noradrenaline.

1.2 Surprise in machine learning

Quantities related to surprise have been previously used in machine learning. Planning to be surprised so as to maximize information gain has been suggested as an optimal exploration technique in dynamic environments [Sun et al., 2011, Little and Sommer, 2011, Kolter and Ng, 2009]. Signatures of surprise have been observed in artificial models of curiosity and interestingness both of which drive active exploration for learning unknown environments, in the absence of external reward [Frank et al., 2013]. Surprise is also linked to Bayesian experimental design [Chaloner and Verdinelli, 1995] and active learning [Settles, 2010]. Furthermore, a surprise measure defined as a prediction error has been optimized in the context of free energy minimization [Friston, 2010, Friston and Kiebel, 2009, Rezende and Gerstner, 2014, Brea et al., 2013].

Being able to detect novel and surprising stimuli is necessary for efficiently learning new memories without altering past useful memories [Carpenter and Grossberg, 1988]. Furthermore, it is critical to consider novelty information (related to the structure of the environment) in addition to reward information in models of reinforcement learning (RL) [Sutton and Barto, 1998b]. Surprise enables an agent to generate (trigger) new states (in the context of RL), an essential feature for learning new environments. As such, surprise is a crucial factor in both learning and memory formation [Ranganath and Rainer, 2003], and should be incorporated in existing machine learning tools to empower them for learning a wide range of learning problems.

1.2.1 Implications of surprise

Surprise has wide-reaching implications: surprise can not only *modulate* learning and drive attention, but can be used as a *trigger* signal for an algorithm that needs to choose between several uncertain states or actions as is the case in change point detection [Nassar et al., 2010,Wilson et al., 2013,Rüter et al., 2012], memory and cluster formation [Gershman and Niv, 2015], exploration/exploitation tradeoff [Cohen et al., 2007,Jepma and Nieuwenhuis, 2011], novelty detection [Knight et al., 1996,Bishop, 1994], and network reset [Bouret and Sara, 2005].

Surprise can be used for learning unknown environmental *states* that a learning agent encounters while interacting with an environment. This empowers the existing reinforcement learning algorithms for coping with learning paradigms in which the environment is not completely known for the agent. The exploration-exploitation tradeoff can also be addressed by surprise. Surprise appears whenever large mismatches between bottom-up sensory information and top-down expectations occur. For instance if a learning agent expects a reward at a certain location, and that location is manually changed, a surprise may be required for increasing the exploration to quickly adapt to the new environment [Cohen et al., 2007,Jepma and Nieuwenhuis, 2011]. The idea is related to the vigilance signal in adaptive resonance theory (ART) [Carpenter and Grossberg, 1988].

ART is based on the insight that learning in artificial and biological neural systems requires a balance between *plasticity* and *stability* (the challenge of stability-plasticity dilemma [Carpenter and Grossberg, 1988]). While plasticity is necessary for being able to integrate new knowledge, stability is needed to prevent forgetting previously learned memories. The role of surprise for this purpose is crucial, and it is highlighted in ART [Carpenter and Grossberg, 1988], that has been proposed to address this challenge by using a surprise-like “vigilance” signal.

Humans and animals learn invariant properties of the environment by building a set of critical features or prototypes that provide compressed representation of the input data space. These prototypes efficiently represent different classes of environmental inputs that the learner encountered beforehand. Such a coding is required for perception and cognition, because it is neither possible nor efficient to store the full sensory information in a limited memory system. Once a data sample arrives, our brain decides whether it should be associated to one of the existing prototypes or whether it belongs to a new prototype which has never been learned (or has been forgotten in case of earlier occurrence). The answer to this question is determined after the brain evaluates how novel (surprising) the data sample is. If the newly acquired sample does not belong to any of the existing prototypes, then surprise could trigger mechanisms for the formation of new memories by opening a gate for a set of synapses that do not yet represent environmental features.

1.3 Multi-factor learning rules for synaptic plasticity

Learning in neural system has been postulated to be linked to synaptic plasticity. This refers to experimental findings in biology which suggest that the strength of a connection between neurons in neural circuitry of the brain is not fixed, but subject to change. The change in strength of neural connections can persist from a few seconds to days and even years, serving as a neural substrate of memory. Theoretical descriptions of synaptic plasticity have been dominated by Hebb's rule [Hebb, 2002] which is based on two major insights: *locality* and *coactivity*. According to Hebb's rule, both pre- and post-synaptic neurons have to be active to make their connection stronger. We call these, the two *local* factors necessary for Hebbian learning. Hebbian rules are simple yet invaluable tools for modeling developmental plasticity such as development of receptive fields. However, Hebbian rules are limited to the realm of unsupervised learning paradigms. Unsupervised paradigms, however, are often incapable of coping with motor learning tasks or conditioning.

Empirical studies, on the other hand, show the existence of additional *global* factors that can influence synaptic plasticity [Reynolds and Wickens, 2002]. These global factors correspond to the activity of neuromodulators that are broadly distributed via multiple ramifications of axons. Deficits in activity of the neuromodulatory system (corresponding to global factors) in humans and animals leave many tasks unlearnable [Decker and McGaugh, 1991]. For instance, Dopamine (DA) as a neuromodulator is used in signaling reward prediction error that takes part in temporal difference (TD) learning algorithms such as Q-learning and SARSA [Schultz et al., 1997, Sutton and Barto, 1998b, Steinberg et al., 2013]. Acetylcholine (ACh) is another candidate neuromodulator used in signaling alertness [Posner and Fan, 2004] and modulating spike-timing dependent plasticity (STDP) [Seol et al., 2007, Couey et al., 2007]. Noradrenaline (NA) influences STDP [Lin et al., 2003, Seol et al., 2007] and modulates neural responses [Aston-Jones and Cohen, 2005, Usher et al., 1999].

It is thus of interest to expand on Hebbian learning rules and formulate general new synaptic plasticity rules that combine the two local Hebbian activity factors with one or multiple global factors. The simplest 3-factor learning rule, including two Hebbian terms modulated by a third factor, is sketched in Fig. 1.1. We believe that surprise-based learning falls into the category of multi-factor learning rules. A surprise signal, just like a reward or success signal in RL, can be theoretically interpreted as a global factor, and biologically explained by the release of a non-specific neuromodulator (e.g., norepinephrine (NE) released from locus coeruleus neurons). Therefore, it can modulate the strength of plasticity that depends, as before, on local Hebbian factors (i.e., the coactivity of pre- and post-synaptic neurons).

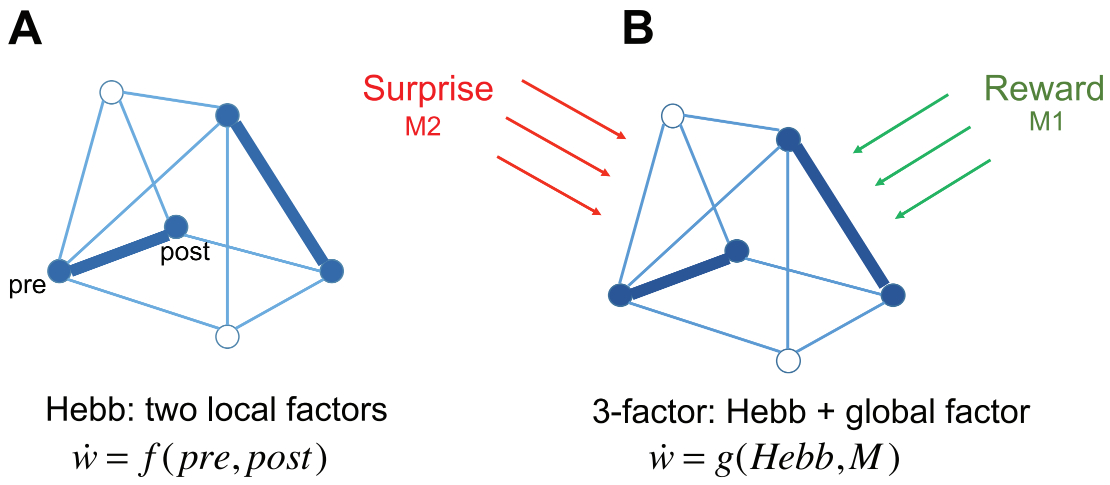

Figure 1.1 – **Hebbian versus 3-factor plasticity. A.** Standard Hebbian plasticity rule is described by the co-activity of pre and post synaptic neurons. All the information that is needed for change in the strength of a connection is locally available for the neurons that are connected by that synapse. **B.** In 3-factor learning rules a global factor (such as reward or surprise) is required to modulate the strength of plasticity in addition to the two local Hebbian factors.

1.4 Structure of the thesis

In Chapter 2, we first conceptually discuss the requirements that should be taken into account for quantification of surprise. We review existing methods and explain their shortcomings. Inspired by information theoretic and Bayesian approaches, we then propose a *confidence-corrected surprise* measure to incorporate subjectivity and uncertainty, two conceptually different aspects of surprise.

We use our proposed measure of surprise in Chapter 3 to develop a new framework for surprise-driven learning. We formulate the principle of surprise minimization as a learning strategy and derive a class of learning rules which obey that principle. A surprise-minimization learning rule, or SMiLe-rule, is derived by a constraint optimization problem in that framework. The SMiLe rule can be used for learning within changing environments by dynamically adjusting the balance between new and old information for inference about the world, without making prior assumptions about the temporal statistics of the environment.

We apply our proposed method to a dynamic decision making task in a Gaussian environment, as well as an exploration task in a maze-like environment. We demonstrate how surprise and uncertainty interact with each other to make learning in changing environments possible. The proposed algorithm benefits from a reduced computational complexity and simpler implementation compared to an explicit solution of a hierarchical Bayesian model. The proposed surprise-modulated belief update algorithm provides a framework to study and model the behavior of humans and animals encountering surprising events. It captures a wide range of behaviors in realistic experimental environments. Moreover, it makes testable prediction about the time course of Noradrenaline [Sara, 2009] or LC-dopamine [Takeuchi et al., 2016] as a neuronal surprise signal, and suggests behavioral experiments that can be performed on real animals.

Chapter 4 links our theory of surprise to the multi-factor learning rules introduced above, and demonstrates how surprise could play the role of a global factor for affecting synaptic plasticity. We first propose a general framework for approximating the exact optimal solution (i.e., the maximizer) of a functional (objective function) that is expressed as an average over many data samples, using an online stochastic learning rule. The proposed learning rule has the form of a 3-factor learning rule, if a neural network is used for parametrization. We then apply our proposed technique to the objective function of the SMiLe rule. We show that the obtained online rule can be interpreted as a covariance learning rule. We implement the online rule in a spiking neural network to demonstrate that our proposed SMiLe rule can also be neurally implemented.

In summary, our work on the foundations of surprise in this thesis provides a framework for future studies on learning with surprise. These include computational studies, such as understanding the interplay between surprise and reward, and neurobiological studies, such as unraveling the interaction between different neural circuits that are functionally involved in learning under surprise. This helps us in addressing unresolved issues about understanding the neural basis of learning.

My contribution to each of these chapters is detailed in the last chapter of the thesis.

2 A Mathematical Description of Surprise

The Webster dictionary defines surprise as “an unexpected event, piece of information” and “the feeling caused by something that is unexpected or unusual” [merriam-webster.com]. Therefore, surprise is *unexpectedness* and represents the gap between what happens and what was expected to happen.

Surprise occurs whenever there is *uncertainty*, be it in the world or in the model that we build of the world. While the former corresponds to the probabilistic nature of the world, the latter is caused by an imperfect internal model of the outside world. We emphasize that surprise is *subjective*: events that are surprising to me may not be surprising to you, although we may both have used the same data to build our models of the world. Subjectivity may arise from different methods for building our internal models, or different prior beliefs about the world [Baldi and Itti, 2010, Palm, 2012]. Individuals may also differ in the way they *perceive* a same event. Therefore they may differently be surprised by the same data because of that subjective perception.

Model uncertainty differs from subjectivity in that the former refers to uncertainty in parameter estimation, given the available data, that remains even in the “best” model (e.g., Bayes-optimal). The latter incorporates individual differences in the construction of a (potentially suboptimal) model given identical data.

Inspired by information theoretic and Bayesian approaches we propose a *confidence-corrected surprise* measure to incorporate subjectivity and uncertainty, two conceptually different aspects of surprise. Our proposed measure of surprise inherits the advantages of Shannon surprise [Shannon, 1948, Tribus, 1961] (a data-driven measure of unexpectedness) and Bayesian surprise [Baldi and Itti, 2010, Itti and Baldi, 2005] (a model-driven approach for quantifying surprise), and overcome their shortcomings. We also emphasize that the confidence-corrected surprise is defined for a *single* data sample, such that an organism can respond to a single event. In contrast, information theoretic quantities, such as data entropy and mutual information, are usually defined as *average* quantities, and thus are not suitable for quantifying surprise in one-shot paradigms.

2.1 Probability-based surprise measures

In order to quantify surprise we need to measure “how much *wow*" [Baldi and Itti, 2010] we get when we encounter an event that deviates from our expectation. Throughout this section, we use a (Bayesian) statistical framework to formulate probability-based surprise measures.

2.1.1 Shannon surprise

We assume that the world is governed by a set of parameters θ^* chosen by nature. If θ^* is known, the *information content* $-\ln p(X|\theta^*)$ for a specific outcome $X \in \mathscr{X}$ is a measure of surprise [Tribus, 1961, Shannon, 1948, Palm, 2012] which says that the occurrence of a rare (i.e., unlikely to occur) data sample X is surprising.

As the information content relates to the *true* probabilities $p(X|\theta^*)$ of samples in the *real* world, it is an *objective*, model-independent, measure of unexpectedness. However, the true set of parameters θ^*, and so the true probability $p(X|\theta^*)$, is rarely known to the observer, such that it is difficult to evaluate the exact information content of a data sample X.

We can bypass this issue by modeling the world as a joint distribution $p(X,\theta) = p(X|\theta)\pi_0(\theta)$ which specifies how data X is generated if the model parameter is $\theta \in \Theta$. The distribution $\pi_0(\theta)$ represents the current belief of the observer about the model parameters θ before X is observed. In what follows we may call the distribution $\pi_0(\theta)$ the prior belief, but it always reflect the most recent belief before data sample X is observed.

The *marginal likelihood* $Z(X) = \int_\Theta p(X,\theta)d\theta$ is a subjective interpretation of the true likelihood $p(X|\theta^*)$, where we integrate out the model parameters θ from the joint distribution. Therefore, we can replace the exact information content $-\ln p(X|\theta^*)$ with the *Shannon surprise*

$$S_{Sh}(X;\pi_0) = -\ln z(X) = -\ln\left[\int_\Theta p(X|\theta)\pi_0(\theta)d\theta\right], \tag{2.1}$$

which depends on the marginal likelihood $Z(X)$ and can be considered as a subjective version of the information content.

Although Shannon surprise [Eq. (2.1)] captures both uncertainty and subjectivity, it may not be the best measure for quantification of surprise. One of the shortcomings of the Shannon surprise is that the calculation of the marginal likelihood $Z(X)$ (also known as the evidence function) could be very intractable [MacKay, 2003,Barber, 2012]. Moreover, Shannon surprise does not incorporate the effect of model uncertainty, that a subject has about his expectation of the world, in surprise evaluation (see Fig. 2.1).

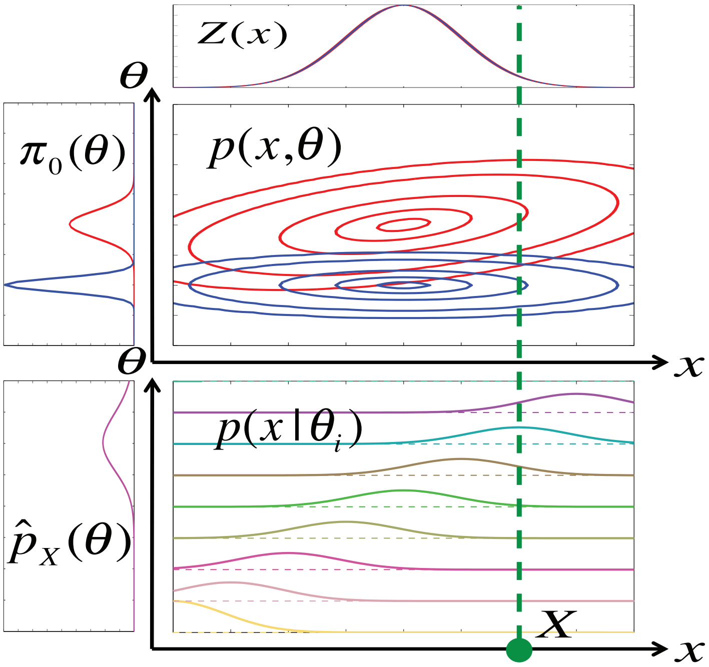

Figure 2.1 – **Confidence-corrected surprise.** The impact of confidence on surprise. Top: Two distinct internal models (red and blue), described by joint distributions $p(x,\theta)$ (contour plots) over observable data x and model parameters θ, may have the same marginal distribution $Z(x) = \int_\theta p(x,\theta)d\theta$ (distributions along the x-axis coincide) but differ in the marginal distribution $\pi_0(\theta) = \int_x p(x,\theta)dx$ (distributions along the θ-axis). Surprise measures that are computed with respect to $Z(x)$ neglect the uncertainty as measured by the entropy $\mathscr{H}(\pi_0)$. Therefore, a given data sample X (green dot) may be equally surprising in terms of the Shannon surprise $S_{Sh}(X)$ [Eq (2.1)] but results in higher confidence-corrected surprise $S_{corr}(X)$ [Eq (2.8)] for the blue as compared to the red model, because π_0 in the red model is wider (corresponding to a larger entropy) than in the blue model. Bottom: The scaled likelihood $\hat{p}_X(\theta)$ (magenta) for the "red" internal model is calculated by evaluating the conditional probability distribution functions $p(x|\theta_i)$ (specified by different color for each θ_i) at $x = X$ (intersection of dashed green line with colored curves). The confidence-corrected surprise $S_{corr}(X)$ is the KL divergence between $\hat{p}_X(\theta)$ (bottom, magenta) and $\pi_0(\theta)$ (top, red).

2.1.2 Bayesian surprise

Once a data sample X is observed, a subject can modify his current belief $\pi_0(\theta)$ about the model parameters using Bayes' rule:

$$\pi(\theta|X) = \frac{p(X|\theta)\pi_0(\theta)}{Z(X)}. \tag{2.2}$$

Bayesian surprise [Itti and Baldi, 2005, Baldi and Itti, 2010] is defined as the Kullback-Leibler (KL) divergence [Kullback and Leibler, 1951] between the prior $\pi_0(\theta)$ and the posterior $\pi(\theta|X)$ either in the form

$$S_{Bayes}(X; \pi_0) = D_{KL}[\pi_0(\theta)||\pi(\theta|X)] \tag{2.3}$$

or in the mirror form $D_{KL}[\pi(\theta|X)||\pi_0(\theta)]$. Bayesian surprise measures the discrepancy or dissimilarity between the prior $\pi(\theta)$ and the posterior $\pi(\theta|X)$ believes about the model parameters θ. According to this measure, a data sample X is more surprising than a data sample X' if it causes a larger change in our belief.

One of the shortcomings of the Bayesian surprise is that it is computed only *after* learning (i.e., once we have changed our belief from prior to posterior). However, behavioral and neural responses indicate that surprise is concurrent with the unexpected event. Our working hypothesis is that the brain evaluates surprise even before recognition, inference or learning occurs. We thus need to evaluate surprise *before* we update our belief so that surprise may control learning rather than emerge from it.

We believe that Bayesian surprise [Eq. (2.3)] resembles, by construction, a measure that quantifies the *effectiveness* of a data sample X on belief update, rather than quantifying how surprising that data sample is perceived. But since surprising samples result in larger change in our belief than non-surprising samples (it will be discussed later in Chapter 3), we may be deceived by Eq. (2.3) as a measure of surprise, while it should not be the case.

2.1.3 Confidence-corrected surprise

Shannon surprise Eq. (2.1) and Bayesian surprise Eq. (2.3) are two distinct yet complementary approaches for calculating surprise. Shannon surprise is about *data* as it captures the inherent unexpectedness of a piece of data given a model. However, it suffers from not covering the influence of model uncertainty on surprise. Bayesian surprise is about a *model* as it measures the change in belief about model parameters.

However, it is computed only after learning, which is inconsistent with behavioral and neural data that suggest an instantaneous response to surprise. Our definition of *confidence-corrected surprise* (what follows) combines these two measures to use their complementary benefits and overcome their shortcomings.

Our confidence-corrected surprise measure is derived in three steps. First we replace the exact information content $-\ln p(X|\theta^*)$ of the observed data sample X with a weighted average over all possible model parameters. It gives us the *raw* surprise of a data sample X:

$$S_{raw}(X;\pi_0) := -\int_\Theta \pi_0(\theta)\ \ln p(X|\theta)\ d\theta. \tag{2.4}$$

In the raw surprise [Eq. (2.4)], we calculate the information content $-\ln p(X|\theta)$ of a data sample X if the true model parameter is θ, and then we average this quantity over all model parameters with a weight that is determined by the current belief $\pi_0(\theta)$. Interestingly, the raw surprise in Eq (2.11) can be expressed as a sum of the Shannon surprise and the Bayesian surprise (see **Materials and Methods** for the proof):

$$S_{raw}(X;\pi_0) = S_{Sh}(X;\pi_0) + S_{Bayes}(X;\pi_0). \tag{2.5}$$

As such, it combines both data-driven approach (Shannon surprise) and the model-driven approach (Bayesian surprise) for measuring surprise.

In addition to the raw surprise [Eq (2.4)] being subjective, we would also like to capture the impact of a subject's *confidence* in her belief. Intuitively, if we are uncertain about what to expect (because we have not yet learned the structure of the world), receiving a data sample that occurs with low probability under the present model is less surprising than a low-probability sample in a situation when we are almost certain about the world.

Our confidence about the current model of the world is represented by the *entropy* $\mathscr{H}(\pi_0) = -\int_\Theta \pi_0(\theta)\ln\pi_0(\theta)d\theta$ of our current belief about the model parameters. To arrive at the *confidence-corrected surprise*, we subtract the entropy $\mathscr{H}(\pi_0)$ of our current belief from the raw surprise, i.e.,

$$S_{raw}(X;\pi_0) - \mathscr{H}(\pi_0) = \int_\Theta \pi_0(\theta)\ln\frac{\pi_0(\theta)}{p(X|\theta)}\ d\theta. \tag{2.6}$$

While the right-hand side of Eq (2.6) is reminiscent of a KL divergence (between $\pi_0(\theta)$ and $p(X|\theta)$ as a function of θ), the likelihood function $p(X|\theta)$ is *not* a normalized probability distribution function with respect to θ. To rewrite Eq (2.6) as a KL divergence, we divide the likelihood $p(X|\theta)$ by a scaling factor $||p_X|| = \int_\Theta p(X|\theta')\, d\theta'$. The *scaled likelihood*

$$\hat{p}_X(\theta) = \frac{p(X|\theta)}{||p_X||} = \frac{p(X|\theta)}{\int_\Theta p(X|\theta')\, d\theta'}, \tag{2.7}$$

can be considered as a probability distribution function over θ, just like the prior $\pi_0(\theta)$. Therefore, the scaled likelihood $\hat{p}_X(\theta)$ and the prior belief $\pi_0(\theta)$ both belong to the space of well-defined probability density functions for the model parameters θ. Note that we may need to discretize the space of the model parameters θ to ensure that everything is well-defined and easy to be calculated.

The KL divergence $D_{KL}[\pi_0(\theta)||\hat{p}_X(\theta)]$ between the two distributions is the confidence-corrected surprise:

$$S_{corr}(X;\pi_0) = D_{KL}[\pi_0(\theta)||\hat{p}_X(\theta)] = \int_\Theta \pi_0(\theta) \ln \frac{\pi_0(\theta)}{\hat{p}_X(\theta)}\, d\theta. \tag{2.8}$$

The confidence-corrected surprise measure [Eq. (2.8)] represents the difference between what we expected to happen (as indicated by the current belief $\pi_0(\theta)$) and what actually happened in the world, where the relevance of a new data point X is indicated by the (scaled) likelihood $\hat{p}_X(\theta)$ [Eq. (2.7)]. Therefore it meets our requirements: a subjective, confidence-adjusted measure of the difference between expectation and realization.

We can alternatively interpret the confidence-corrected surprise in the following way. The scaled likelihood $\hat{p}(\theta)$ is in fact the posterior belief that is achieved under a *flat* prior, i.e., where we have no prior knowledge about the world (see **Materials and Methods**). The prior $\pi_0(\theta)$, on the other hand, can be interpreted as a posterior belief that is achieved without taking into account the newly acquired data sample X. The confidence-corrected surprise [Eq. (2.8)] quantifies how different these two posteriors are, using the KL divergence.

Note that our proposed confidence-corrected surprise measure $S_{corr}(X;\pi_0)$ in Eq (2.8) inherits the property of the raw surprise $S_{raw}(X;\pi_0)$ in Eq (2.4), that can be expressed as the sum of Shannon surprise and Bayesian surprise (see Eq. (2.5)). As such, it also combines the benefits of both data-driven and model-driven approaches for measuring surprise.

In principal one could have a fixed model of the world, with no ability to further adapt it, and one can be surprised many times under this model. However, in order for surprise to be (behaviorally) meaningful, i.e., carry valuable information, there needs to be consequences to surprise. These include interruption of an ongoing action, attentional shifts, change in choice, or learning. In Chapter 3 we will more precisely discuss the implications of surprise in learning.

2.2 Discussion

2.2.1 Absolute z-score

Apart from probability-based surprise measures, model-free approaches can also be used to quantify surprise in simplified experimental paradigms. In a simple paradigm, model-free approaches might even be preferred to probability-based surprise measures in fMRI and behavioral studies. Here we would like to emphasize that the confidence-corrected surprise can be simplified when they are put in a given context. Here is an example:

In the context of reward-based learning, the *prediction error* signal $\delta = r - \mathbb{E}(r)$ quantifies the difference between the actual reward r and the expected reward $\mathbb{E}(r)$, and has been frequently used in error-driven algorithms such as temporal difference (TD) learning [Sutton and Barto, 1998a]. Larger prediction errors in noisy and *volatile* environments (where the standard deviation σ of random reward r is high) are less surprising than in *stable* environments (with small σ). As such the scaled absolute prediction error (also known as z-score)

$$S(r) = \frac{|r - \mathbb{E}(r)|}{\sigma}, \tag{2.9}$$

may be considered as one of the simplest model-free approaches for quantifying surprise.

All probability-based surprise measures, we introduced in this chapter, can be simplified to the absolute z-score Eq. (2.9), in case of random reward delivery that is modeled in a Gaussian setting (see **Materials and Methods**).

2.2.2 Binary or graded surprise?

A key question in the quantification of surprise is whether we should think about surprise in a binary way (surprised or not surprised) or in a graded way (different levels of surprise). Behavioral sciences almost suggest the former (i.e., in case of a surprising event, I either interrupt what I'm doing or I don't), but from the view point of neural implementations, surprise should be graded.

All the surprise measures introduced here can be bounded in the range of $[0,1]$ by using a sigmoid function (e.g., hyperbolic tangent). This formulation also accommodates binary theories of surprise in which an organism is either surprised or not, without specifying a level of surprise. Moreover, a further incorporation of subjective parameters to determine the shape of the function can be beneficial for fitting data to behavior. An example in which the propensity of a subject in changing his belief is modeled by a subjective parameter will be provided in Chapter 3.

2.2.3 Subjective perception affects surprise

Surprise is not directly related to the observation. It rather depends on how an outcome is *perceived* in the subject's mind. Different individuals may differently perceive a same outcome, even if the world is completely known and even if they do not need to build an internal model for that. In other words, subjectivity may not only arise from different priors or different methods for model construction, but the subjects may also have different perception approaches. For clarification of the statement above, we provide an example, studied in Palm theory of surprise (see **Appendix A** for a review, [Palm, 2012]).

In the context of game of lotto 6 numbers are drawn from a set $L = \{1, 2, ..., 49\}$ of numbers without replacement. Without loss of generality, we display these 6 random numbers ordered by their size $x_1 < x_2 < ... < x_6$ with $x_i \in L$. In what follows, we would like to discuss how surprising an observed data batch $X = \{x_1, x_2, ..., x_6\}$ is perceived in different subjects' minds. Note that in this example, there is *no learning* as the model of the world is completely known (i.e., subject knows that all data batches X are *equally* probable with probability $p(X) = \frac{(49-6)!}{49!}$. In case of no learning, all surprise measures we discussed earlier would be equivalent to the information content $-\ln p(x|\theta^*)$, except the Bayesian surprise which is not well-defined in this context (because it depends on model change while we have no change in the model). The information content, however, results in an *equal* amount of surprise for all outcomes X, which is not the case in real behavioral experiment.

For instance, the occurrence of $X = \{11, 12, 13, 14, 15, 16\}$ seems to be perceived as much more surprising than the occurrence of $Y = \{7, 16, 23, 35, 40, 48\}$, in reality. This is because we "perceive" X and Y differently. X is perceived as a set of subsequent numbers which is very unlikely to occur again. However, data batch Y is interpreted as an ordinary set of numbers with no significant relation between its entries. We can simply model this finding by replacing $p(X)$ with $p(c[X])$ where $c[X]$ is the *class* of all data batches (including X) that are similarly perceived as X. In other words, to quantify how surprising the outcome X above is perceived, we need to evaluate how surprising is to have a data batch with six subsequent numbers. All data batches that fulfill this description belongs to the class $c[X]$. In fact, the surprise of X should be calculated by $-\ln p(c[X])$ and not $-\ln p(X)$.

Perception is a subjective concept and is affected by subjects' background. To clarify, we ask a question. Among outcome $Y = \{7,16,23,35,40,48\}$ and $Z = \{2,7,19,23,37,43\}$, which one is more surprising? The answer depends on the subjective perception. For those who recognize that all the numbers in Z are prime numbers, this configuration is probably more surprising than Y. This is because the probability of observing six prime numbers as an outcome is less than an appearance of an ordinary set of numbers. But for those who could not easily recognize such a relation between the entries, both configurations Y and Z might be equally surprising.

2.3 Materials and Methods

2.3.1 The scaled likelihood is the posterior belief under the flat prior

Let us assume that all model parameters θ must stay in some bounded convex interval of volume A. Given a data sample X, the posterior belief $p^{flat}(\theta|X)$ about the model parameters θ (derived by the Bayes rule) under the assumption of a flat prior $\hat{\pi}_0(\theta) = 1/A$ is:

$$p^{flat}(\theta|X) = \frac{p(X|\theta)\hat{\pi}_0(\theta)}{\int_\Theta p(X|\theta)\hat{\pi}_0(\theta)\, d\theta} = \frac{p(X|\theta)}{\int_\Theta p(X|\theta)\, d\theta} = \frac{p(X|\theta)}{||p_X||} = \hat{p}_X(\theta). \tag{2.10}$$

2.3.2 The raw surprise increases with the Shannon surprise and the Bayesian surprise

An interesting feature of the raw surprise Eq. (2.4) is that it incorporates both data-driven (Shannon surprise) and model-driven (Bayesian surprise) approaches for calculating surprise. This is because the raw surprise can be expressed as the sum of the Shannon surprise and Bayesian surprise:

$$\begin{aligned} S_{raw}(X;\pi_0) \quad & \overset{(2.4)}{=} \quad -\int_\Theta \pi_0(\theta)\ln p(X|\theta)\ d\theta \\ & \overset{(2.2)}{=} \quad -\int_\Theta \pi_0(\theta)\ln\big[\frac{\pi(\theta|X)\left(\int_\Theta p(X|\theta)\pi_0(\theta)\ d\theta\right)}{\pi_0(\theta)}\big]\ d\theta \\ & = \quad D_{KL}[\pi_0(\theta)||\pi(\theta|X)] - \ln\big[\int_\Theta p(X|\theta)\pi_0(\theta)\ d\theta\big] \\ & \overset{(2.3),(2.1)}{=} \quad S_{Bayes}(X;\pi_0) + S_{Sh}(X;\pi_0). \end{aligned} \tag{2.11}$$

Therefore, the raw surprise is always lower bounded by the Shannon surprise, i.e.,

$$S_{raw}(X;\pi_0) \geq S_{Sh}(X,\pi_0), \tag{2.12}$$

because the difference between these two surprise measures is equal to the Bayesian surprise, according to Eq. (2.11), which is expressed as a KL divergence (a non-negative quantity). The fact that the raw surprise is always bigger than or equal to the Shannon surprise, can also be explained by the Jensen's inequality: If θ is a random variable and $\phi(\theta) = -\ln p(X|\theta)$ is a convex function, then $\mathbb{E}[\phi(\theta)] \geq \phi(\mathbb{E}[\theta])$, i.e.,

$$S_{raw}(X;\pi_0) \overset{(2.4)}{=} \mathbb{E}_{\pi_0}[-\ln p(X|\theta)] \ \geq\ -\ln\mathbb{E}_{\pi_0}[p(X|\theta)] \overset{(2.1)}{=} S_{Sh}(X;\pi_0). \tag{2.13}$$

2.3.3 Absolute z-score is linked to probability-based surprise measures in a Gaussian setting

In the following we provide a simple example in a Gaussian setting, and analytically calculate all the probability-based surprise measures introduced in this chapter, for a given data sample X. We show that all these surprise measures can be linked to the z-score $\mathcal{Z}(X)$ using a quadratic mapping:

$$S(X) = m\mathcal{Z}(X)^2 + n, \tag{2.14}$$

for some constants $m, n \in \mathbb{R}$.

Suppose that the world generates samples that are normally distributed around the (unknown) mean μ^*. We assume that the variance σ_x^2 of the distribution from which samples X are drawn is known. Therefore, the only parameter that we may be uncertain about (in our internal model of the external world) is the true underlying mean. The uncertainty about the true mean μ^* is modeled by the current belief

$\pi_0(\mu) \sim \mathcal{N}(\mu^*, \sigma_\mu^2)$ which is a normal distribution with mean μ^* and variance σ_μ^2. Larger σ_μ^2 implies higher uncertainty about the mean. Note that the probability of observing X given the mean μ is also normal, i.e., $p(X|\mu) \sim \mathcal{N}(\mu, \sigma_x^2)$. We first calculate all the probability-based surprise measures, introduced in this chapter, for a given data sample X:

The information content (as an objective, model-independent measure of surprise) is equal to

$$-\ln p(X|\mu^*) = \frac{1}{2}\ln[2\pi\sigma_x^2] + \frac{(X-\mu^*)^2}{2\sigma_x^2} \tag{2.15}$$

Since both likelihood $p(X|\mu)$ and the prior $\pi_0(\mu)$ are normal, the marginal likelihood $Z(X) = \int_\mu p(X|\mu)\pi_0(\mu)d\mu$ is also expressed by a normal distribution $\mathcal{N}(\mu^*, \sigma_x^2 + \sigma_\mu^2)$. Therefore, the Shannon surprise $S_{Sh}(X;\pi_0)$ [Eq. (2.1)] is equal to

$$S_{Sh}(X;\pi_0) = \frac{1}{2}\ln[2\pi(\sigma_x^2 + \sigma_\mu^2)] + \frac{(X-\mu^*)^2}{2(\sigma_x^2 + \sigma_\mu^2)}. \tag{2.16}$$

The posterior belief $\pi(\mu|X)$ that is obtained by the Bayes' rule in Eq. (2.2) has also a normal distribution $\mathcal{N}(\alpha X + (1-\alpha)\mu^*, \sigma_x^2 \oplus \sigma_\mu^2)$, where $\alpha = \frac{\sigma_\mu^2}{\sigma_x^2+\sigma_\mu^2}$ and $\sigma_x^2 \oplus \sigma_\mu^2 = \frac{\sigma_x^2\sigma_\mu^2}{\sigma_x^2+\sigma_\mu^2}$. The Bayesian surprise $S_{Bayes}(X;\pi_0)$ [Eq. (2.3)] is therefore equal to:

$$S_{Bayes}(X;\pi_0) = \frac{\sigma_\mu^2}{2\sigma_x^2}\left(\frac{(X-\mu^*)^2}{\sigma_\mu^2 + \sigma_x^2}\right) + \frac{1}{2}\left(\frac{\sigma_\mu^2}{\sigma_x^2} - \ln\left[1 + \frac{\sigma_\mu^2}{\sigma_x^2}\right]\right), \tag{2.17}$$

where we used the following formula for the KL divergence of the normal distributions:

$$D_{KL}[\mathcal{N}(a_1, b_1^2)||\mathcal{N}(a_2, b_2^2)] = \frac{(a_1 - a_2)^2}{2b_2^2} + \frac{1}{2}\left(\frac{b_1^2}{b_2^2} - 1 - \ln\frac{b_1^2}{b_2^2}\right). \tag{2.18}$$

The raw surprise $S_{raw}(X;\pi_0)$ [Eq. (2.4)] is equal to

$$\begin{aligned}
S_{raw}(X;\pi_0) &\overset{(2.4)}{=} \int_\mu -\ln p(X|\mu)\, \mathcal{N}(\mu|\mu^*,\sigma_\mu^2)\, d\mu \\
&\overset{(2.15)}{=} \int_{-\infty}^{\infty} \left(\frac{1}{2}\ln[2\pi\sigma_x^2] + \frac{(X-\mu)^2}{2\sigma_x^2}\right) \mathcal{N}(\mu|\mu^*,\sigma_\mu^2)\, d\mu \\
&= \frac{1}{2}\ln[2\pi\sigma_x^2] + \frac{1}{2\sigma_x^2}\int_{-\infty}^{\infty}(X^2-2X\mu+\mu^2)\, \mathcal{N}(\mu|\mu^*,\sigma_\mu^2)\, d\mu \\
&= \frac{1}{2}\ln[2\pi\sigma_x^2] + \frac{1}{2\sigma_x^2}\left[X^2-2X\mu^*+(\sigma_\mu^2+\mu^{*2})\right] \\
&= \frac{1}{2}\ln[2\pi\sigma_x^2] + \frac{(X-\mu^*)^2}{2\sigma_x^2} + \frac{\sigma_\mu^2}{2\sigma_x^2}. \qquad (2.19)
\end{aligned}$$

The confidence-corrected surprise [Eq. (2.8)] is equal to

$$S_{corr}(X;\pi_0) = \frac{(X-\mu^*)^2}{2\sigma_x^2} + \frac{1}{2}\left(\frac{\sigma_\mu^2}{\sigma_x^2} - 1 - \ln\frac{\sigma_\mu^2}{\sigma_x^2}\right) \qquad (2.20)$$

where we used Eq. (2.18) for calculating the KL divergence between $\hat{p}_X(\theta) \sim \mathcal{N}(X,\sigma_x^2)$ and $\pi_0(\mu) \sim \mathcal{N}(\mu^*,\sigma_\mu^2)$. The confidence-corrected surprise [Eq. (2.8)], in this example, can also be derived by subtracting the entropy form the raw surprise:

$$\begin{aligned}
S_{corr}(X;\pi_0) &= S_{raw}(X;\pi_0) - \mathcal{H}(\pi_0) \\
&\overset{(2.19)}{=} \frac{1}{2}\ln[2\pi\sigma_x^2] + \frac{(X-\mu^*)^2}{2\sigma_x^2} + \frac{\sigma_\mu^2}{2\sigma_x^2} - \frac{1}{2}\ln[2\pi e\sigma_\mu^2] \\
&= \frac{(X-\mu^*)^2}{2\sigma_x^2} + \frac{1}{2}\left(\frac{\sigma_\mu^2}{\sigma_x^2} - 1 - \ln\frac{\sigma_\mu^2}{\sigma_x^2}\right), \qquad (2.21)
\end{aligned}$$

where we used $\mathcal{H}(\pi_0) = \frac{1}{2}\ln[2\pi e\sigma_\mu^2]$ in the second line of derivation above. Note that in case of Gaussian likelihood, the scaled likelihood is the same as the likelihood (because $||p_X|| = 1, \forall X$). Therefore, the confidence-corrected surprise can be expressed as Eq. (2.6).

To see why all probability-based surprise measures above can be linked to the absolute z-score $\mathcal{Z}(X) = \frac{X-\mu^*}{\sigma_x^2}$ via Eq. (2.14), assume $\sigma_x^2 = \sigma_\mu^2$ and rewrite Eqs.(2.15-2.20).

3 Balancing New Against Old Information: The Role of Surprise in Learning

Encountering unexpected (surprising) events is part of our daily experience. How humans and animals can rapidly detect unexpected events and quickly adapt to changing environments is an open question. Our hypothesis is that humans and animals use a surprise signal to define the moments when learning should be most effective. In the present study, a new framework for surprise-driven learning is proposed which consists of two components: (i) a confidence-adjusted surprise measure to capture environmental statistics as well as subjective beliefs (see Chapter 2), and (ii) a surprise-minimization learning rule, or SMiLe-rule, which dynamically adjust the balance between new and old information for inference about the world, without making prior assumptions about the temporal statistics of the environment.

We apply our framework to a dynamic decision-making task and a maze exploration task to demonstrate that it is suitable for learning in complex environments that undergo gradual or sudden changes. The proposed algorithm benefits from a reduced computational complexity and simpler implementation compared to an explicit solution of a hierarchical Bayesian model. The proposed surprise-modulated belief update algorithm is able to capture a wide range of behaviors in realistic experimental environments. It provides a framework to study the behavior of humans and animals encountering surprising events. Moreover, it makes testable prediction about the time course of Noradrenaline as a neuronal surprise signal.

3.1 Introduction

Humans and animals rely on previously learned knowledge to guide their behavior. A crucial challenge when collecting new data in uncertain environments is the balance between new and old information. How much should we trust what we have learned in the past and how much should we adjust our model of the world based on newly acquired data? In noisy environments, individual data samples are not reliable and a model needs to average over the past data. However, when a structural change occurs in the environment, the most recent data samples are the most informative ones and we would like to quickly forget what was learned in the past.

Both humans and animals adaptively adjust the relative contribution of old and newly acquired data on learning [Behrens et al., 2007, Nassar et al., 2012, Krugel et al., 2009, Pearce and Hall, 1980] and rapidly adapt to changing environments [Pearce and Hall, 1980, Wilson et al., 1992, Holland, 1997]. To capture this behavior, existing models detect and respond to sudden changes using (absolute) reward prediction errors [Hayden et al., 2011, Pearce and Hall, 1980], risk prediction errors [Preuschoff and Bossaerts, 2007, Preuschoff et al., 2008], uncertainty-based jump detection [Nassar et al., 2010, Payzan-LeNestour and Bossaerts, 2011] and hierarchical modeling [Behrens et al., 2007, Adams and MacKay, 2007]. The nature of the environmental change determines which of these models works best. Here we aim to generalize these approaches by using surprise as a trigger for shifting the balance between old and new information.

We formulate the principle of surprise minimization as a learning strategy and derive a class of learning rules which obey that principle. We then propose a *surprise-modulated* belief update rule that can be used for learning within changing environments. We apply our proposed method to a dynamic decision making task in a Gaussian environment, as well as an exploration task in a maze-like environment. We demonstrate how surprise and uncertainty interact with each other to make learning in changing environments possible. Finally, we discuss implications of surprise in reinforcement learning, and link surprise and its role in learning/plasticity to existing neurophysiological evidence and behavioral data.

3.2 Results

3.2.1 Surprise minimization: the SMiLe-rule

Successful learning implies an adaptation to the environment such that an event occurring for a second time is perceived as less surprising than the first time. In the following *surprise minimization* refers to a learning strategy which modifies the internal model of the external world such that the unexpected observation becomes less surprising if it happens again in the near future. Surprise minimization is akin to – though more general than – reward prediction error learning. Reward based learning modifies the reward expectation such that a recurring reward results in a smaller reward prediction error. Similarly, surprise-minimization learning results in a smaller surprise for recurring events.

To mathematically formulate learning through surprise minimization, we define a *learning rule* $L(X, \pi_0)$ as a mapping from a prior belief $\pi_0(\theta)$ to a posterior belief $q(\theta)$ after receiving data sample X, i.e., $q = L(X, \pi_0)$. Moreover, we define a *belief update* as the learning step after a single data sample.

We define the class $\mathscr{L}$ of *plausible* learning rules as the set of those learning rules L for which the surprise $\mathscr{S}(X; q)$ of *any* data sample X under the posterior belief $q(\theta)$ is *at most as surprising as* the surprise $\mathscr{S}(X; \pi_0)$ of that data sample under the prior belief $\pi_0(\theta)$, i.e.,

$$\mathscr{L} = \{L : \mathscr{S}(X; q) \leq \mathscr{S}(X; \pi_0),\ q = L(X, \pi_0), \forall X \in \mathscr{X}\}. \tag{3.1}$$

In other words, if the same data sample X occurs a second time right after a belief update, it is perceived as less surprising than the first time.

After the belief update we can measure how much the new data X has impacted the internal model by comparing the surprise of data sample X under the posterior belief to its surprise under the prior belief:

$$\Delta\mathscr{S}(X; L) = \mathscr{S}(X; \pi_0) - \mathscr{S}(X; q). \tag{3.2}$$

Given a learning rule L, a data sample X is considered more effective for a belief update than X', if $\Delta\mathscr{S}(X; L) > \Delta\mathscr{S}(X'; L)$. Note that definitions in Eqs (3.1) and (3.2) do not depend on our specific choice of surprise measure $\mathscr{S}$. In the following we choose $\mathscr{S}$ to be the confidence-corrected surprise S_{corr} [Eq (2.8)].

The *impact function* $\Delta S_{corr}(X; L)$ [Eq (3.2)], for a given data sample X, is maximized by the learning rule that maps the prior belief $\pi_0(\theta)$ to the scaled likelihood $\hat{p}_X(\theta)$. However, as this posterior distribution $q = \hat{p}_X$ does not depend on the prior belief π_0, it discards all previously learned information. Therefore, it amounts to a valid though meaningless solution.

To avoid overfitting to the last data sample, we need to limit our search to posteriors q that are not too different from the prior π_0. This limited set can be expressed as the set of posteriors q that fulfill the constraint $D_{KL}[q||\pi_0] \leq B$, for some non-negative upper bound $B \geq 0$. The parameter B determines how much we allow our belief to change after receiving a data sample X. Maximizing the impact function $\Delta S_{corr}(X; L)$ under such a constraint, is equivalent to the following constraint optimization problem:

$$\min_{q:D_{KL}[q||\pi_0]\leq B} S_{corr}(X;q). \tag{3.3}$$

Using the method of Lagrange multipliers we find the solution of problem in Eq (3.3) to be

$$q_\gamma(\theta) = \frac{p(X|\theta)^\gamma \pi_0(\theta)^{1-\gamma}}{Z(X;\gamma)}, \tag{3.4}$$

where $Z(X;\gamma) = \int_\Theta p(X|\theta)^\gamma \pi_0(\theta)^{1-\gamma}\, d\theta$ is a normalizing factor and $0 \leq \gamma \leq 1$ is uniquely determined by the bound B (see **Materials and Methods** for the proof). The unique relationship between γ and B means that once B has been chosen, γ is no longer a free parameter and vice versa.

We call the learning rule of Eq (3.4) *surprise minimization learning (SMiLe)* rule. It is reminiscent of Bayes' rule except for the parameter γ which modulates the relative contribution of the likelihood $p(X|\theta)$ and the prior $\pi_0(\theta)$ to the posterior $q(\theta)$. Note that the SMiLe rule belongs to the class $\mathscr{L}$ of plausible learning rules, for all $0 \leq \gamma \leq 1$.

Choosing γ in the range $0 \leq \gamma \leq 1$ is equivalent to choosing a bound $B \geq 0$. To understand how the optimal solution in Eq (3.4), and thus γ, relates to the boundary B, we illustrate its limiting cases (see Fig 3.1): (i) $B = 0$ yields $\gamma = 0$ and the posterior q is identical to the prior π_0. In other words, the new information is discarded. (ii) For $B \geq B_{max} = D_{KL}[\hat{p}_X||\pi_0]$, the solution is always the scaled likelihood $\hat{p}_X$ (corresponding to $\gamma = 1$) because $q = \hat{p}_X$ fulfills the constraint $D_{KL}[q||\pi_0] \leq B$ for any $B \geq B_{max}$ and minimizes $S_{corr}(X;q)$ among all posteriors q. This is equivalent to the unconstrained case, and implies that all previous information is discarded. (iii) For $0 < B < B_{max}$ the optimal solution is the posterior q_γ [Eq (3.4)] with $0 < \gamma < 1$ satisfying $D_{KL}[q_\gamma||\pi_0] = B$. Moreover, $B > B'$ implies $\gamma > \gamma'$ (see Fig 3.1, and **Materials and Methods** for the proof).

While the SMiLe rule [Eq (3.4)] depends on a parameter γ which is uniquely determined by the bound B, we have yet to indicate how to choose B. Highly surprising data should result in larger belief shifts. As such, the bound B should increase with the level of surprise S_{corr}.

The definition of an optimal (nonlinear) mapping from S_{corr} to B (and thus to γ) would require further assumptions about how surprise is related to the bound and we will therefore not search for optimality. However, it is instructive to study a few examples. For instance, if the nonlinear mapping were a step function, the system

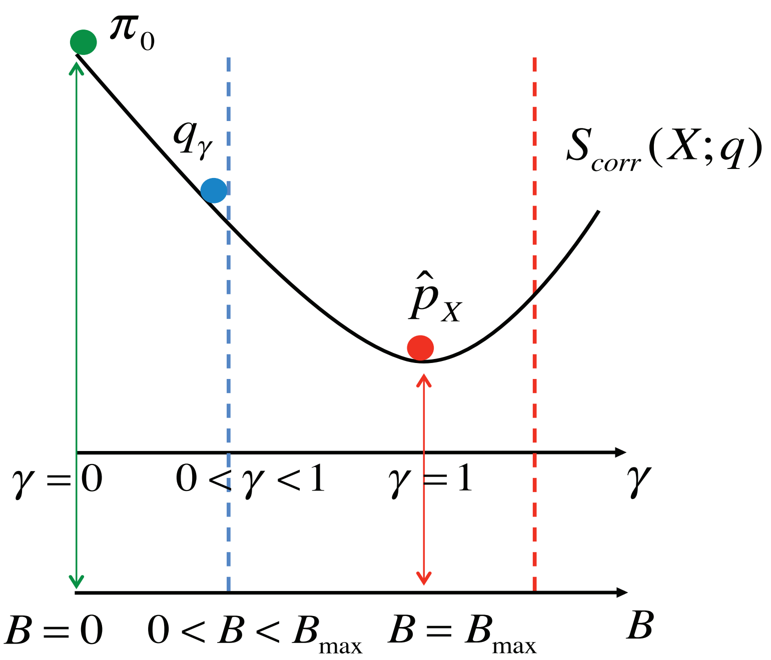

Figure 3.1 – **Constraint surprise minimization.** Solutions to the (constraint) optimization problem in Eq (3.3). The objective function, i.e. the posterior surprise $S_{corr}(X;q)$ (black) for a given data sample X, is a parabolic landscape over γ where each γ corresponds to a unique posterior q_γ. Its global minimum is at $\gamma = 1$ (corresponding to $q_1 = \hat{p}_X$) which is equivalent to discarding all previously observed samples. The boundary B constrains the range of γ and thus the set of admissible posteriors. At $B = 0$ no change is allowed resulting in $\gamma = 0$ with a posterior equals to the prior π_0 (green). $B \geq B_{max} = D_{KL}[\hat{p}_X||\pi_0]$ (red dashed line) implies that we allow posteriors that are further away from the prior than the sample itself so the optimal solution is the scaled likelihood $\hat{p}_X$ or $\gamma = 1$ as for the unconstrained problem. For $0 < B < B_{max}$ (blue dashed line) the objective function is minimized by q_γ in Eq (3.4) that fulfills the constraint $D_{KL}[q_\gamma||\pi_0] = B$ with $0 < \gamma < 1$.

would make a binary choice between either keeping the old belief or relying on the last new data point. On the other hand, an extremely slow increase would amount to largely ignoring the surprise and sticking to the same old belief. Therefore, the sharpness of the transition in the mapping function matters. The *exact* link between the bound and surprise is, however, not crucial as long as B is monotonic in surprise in a *reasonable* way.

In the following, we choose a simple monotonic function to link the bound to the surprise. For each data sample X, we take

$$B(X) = \frac{mS_{corr}(X;\pi_0)}{1 + mS_{corr}(X;\pi_0)} B_{max}(X), \tag{3.5}$$

where $B_{max}(X) = D_{KL}[\hat{p}_X||\pi_0]$. Here, the monotonic function depends on a subject-specific parameter m that describes an organism's propensity toward changing its belief. Note that in Eq (3.5), $m = 0$ indicates that the subject will never change her belief. As m increases so does a subject's willingness to change her belief. Thus, differences in m from one subject to the next will eventually allow us to capture heterogeneity in belief update strategies. Although m is inserted in Eq (3.5) to model subjective behaviors, one could also search for the best m algorithmically in a given simulated environment or other computational setting.

Note that biological correlates of surprise such as pupil dilation or the activity of a neuromodulator will normally saturate at some maximal value, consistent with our choice of a saturating function in Eq (3.5).

3.2.2 Surprise-modulated belief update

The surprise-modulated belief update combines the confidence-corrected surprise [Eq (2.8)] and the SMiLe rule [Eq (3.4)] to dynamically update our belief: after receiving a new data point X, we evaluate the surprise $S_{corr}(X;\pi_0)$ which sets the bound B [Eq (3.5)] for our update and allows us to solve for γ. We then update the belief, using the SMiLe rule [Eq (3.4)] with parameter γ (see Algorithm 1).

Algorithm 1 Pseudo algorithm for surprise-modulated belief update (SMiLe)

```
$N$ ← number of data samples
Belief ← $\pi_0$ (the prior belief)
$m$ ← 0.1 (subject-dependent)
for $n$: 1 to $N$ do
    $X_n$ ← a new data sample
    (i) evaluate the surprise $S_{corr}(X_n;\text{Belief})$, Eq (2.8)
    (ii-a) calculate $B_{max}(X_n) = D_{KL}[\hat{p}_{X_n}||\text{Belief}]$
    (ii-b) choose the bound $B(X_n) = \frac{mS_{corr}(X_n;\text{Belief})}{1+mS_{corr}(X_n;\text{Belief})} B_{max}(X_n)$
    (iii) find $\gamma$ by solving $D_{KL}[q_\gamma||\text{Belief}] = B(X_n)$
    (iv) update using SMiLe, Eq (3.4): Belief($\theta$) ← $\frac{p(X_n|\theta)^\gamma \text{Belief}(\theta)^{1-\gamma}}{\int_\Theta p(X_n|\theta)^\gamma \text{Belief}(\theta)^{1-\gamma}\, d\theta}$
Return Belief;
```

Note 1: In each iteration, we first calculate the surprise, step (i), before the model is updated in step (iv).
Note 2: The steps (ii-a), (ii-b), and (iii) can be merged and approximated by $\gamma = f(S_{corr}(X_n;\text{Belief}))$ where $f(.)$ is a subjective function that increases with surprise.

The parameter γ in the SMiLe rule controls the *impact* of a data sample X on belief update such that a bigger γ causes a larger impact. More precisely, the impact function $\Delta S_{corr}(X; L)$ in Eq (3.2), where L is replaced by the SMiLe rule [Eq (3.4)], is an increasing function of γ (see **Materials and Methods** for the proof).

We note that in classical models of perception and attention [Itti and Baldi, 2009, Baldi and Itti, 2010], surprise has been defined as a measure of belief change (such as $D_{KL}[q_\gamma||\pi_0]$ or its mirror form $D_{KL}[\pi_0||q_\gamma]$). We emphasize that our model of surprise is "fast" in the sense that it can be evaluated *before* the beliefs are changed. Interestingly, the impact function is linked to the measure of *belief change* by the following equation (see **Materials and Methods** for derivation),

$$\Delta S_{corr}(X; L) = \frac{1}{\gamma} D_{KL}[\pi_0||q_\gamma] + \left(\frac{1}{\gamma} - 1\right) D_{KL}[q_\gamma||\pi_0] \geq 0. \tag{3.6}$$

Therefore *a larger reduction in the surprise implies a bigger change in belief.*

3.2.3 Simulations

In the following we will look at two examples to illustrate the functionality of our proposed surprise-modulated belief update Algorithm 1. The first is a simple, one-dimensional dynamic decision-making task which has been used in behavioral studies [Nassar et al., 2012, Behrens et al., 2007] of learning under uncertainty. While somewhat artificial as a task, it is appealing as it nicely isolates different forms of uncertainty. This allows us: (i) to demonstrate the basic quantities and properties of our algorithm, and (ii) to show how its flexibility allows it to capture a wide range of behaviors. The second example is a multi-dimensional maze-exploration task which we will use to demonstrate how our algorithm extends to and performs in more complex and realistic experimental environments.

Gaussian estimation

Task. In the one-dimensional dynamic decision-making task, subjects are asked to estimate the mean of a distribution based on consecutively and independently drawn samples. At each time step n, a data sample X_n is drawn from a normal distribution $\mathcal{N}(\mu_n, \sigma_x^2)$ and the subject is asked to provide her current estimate $\hat{\mu}_n$ of the mean of the distribution. Throughout the experiment, the mean may change without warning (Fig 3.2A). Changes occur with a *hazard rate* of $H = 0.066$. The variance σ_x^2 remains fixed.

Model. We model the subject's belief *before* the n-th sample X_n is observed, as the

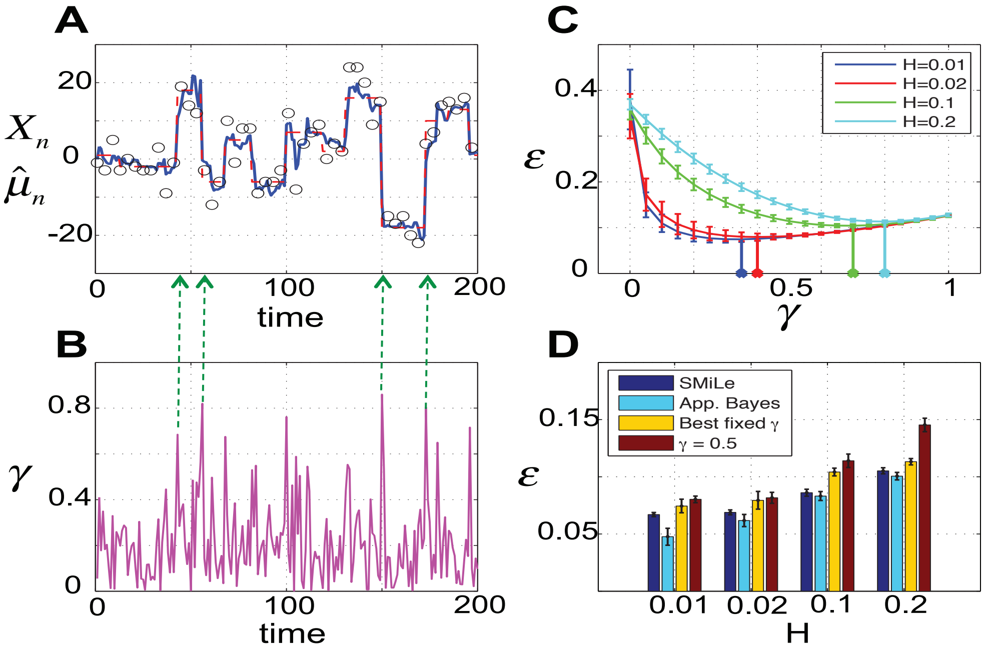

Figure 3.2 – **Gaussian mean estimation task.** At each time step, a data sample X_n is independently drawn from a normal distribution whose underlying mean may change within the interval $[-20, 20]$ at unpredictable change points. On average, the underlying mean remains unchanged for 15 time steps corresponding to a hazard rate $H = 0.066$. The standard deviation of the distribution is fixed to 4 and is assumed to be known to the subject. **A.** Using a surprise-modulated belief update (Algorithm 1), the estimated mean (blue) quickly approaches the true mean (dashed red) given observed samples (black circles). A few selected change points are indicated by green arrows. **B.** The weight factor γ in Eq (3.8) (magenta) increases at the change points, resulting in higher influence of newly acquired data samples on the posterior mean. **C.** The estimation error ϵ per time step versus the weight factor $0 \leq \gamma \leq 1$ in the delta-rule method with constant γ for four different hazard rates. The minimum estimation error (for best fixed γ) is achieved by a γ (points on the horizontal axis) that decreases with the hazard rate, indicating that a bigger γ is preferred in volatile environments. Error bars indicate standard deviation over all trials and 50 episodes. **D.** For all models, the average estimation error ϵ increases with the hazard rate. Moreover, surprise-modulated belief update (SMile, dark blue) outperforms the delta-rule with the *best* fixed γ (Best fixed γ, yellow). The best fixed γ for each hazard rate corresponds to the learning rate that has minimal estimation error (indicated by points on the horizontal axis in sub-figure **C**). Although the surprise-modulated SMile rule performs worse than the approximate Bayesian delta-rule [Nassar et al., 2010] (App. Bayes, light blue), the difference in the performance is not significant, except for the very small hazard rate of 0.01.

normal distribution $\mathcal{N}(\hat{\mu}_{n-1}, \sigma^2_{n-1})$ where $\hat{\mu}_{n-1}$ is the estimated mean and σ^2_{n-1} determines how uncertain the subject is about her estimation. In order to keep the scenario as simple as possible, we assume $\sigma^2_0 = \sigma^2_x$. The posterior mean $\hat{\mu}_n$ resulting from the surprise-modulated belief update (Algorithm 1) is a *weighted average* of the prior mean $\hat{\mu}_{n-1}$ and the new sample X_n (see **Materials and Methods** for derivation),

$$\hat{\mu}_n = \gamma X_n + (1-\gamma)\hat{\mu}_{n-1}. \tag{3.7}$$

The weight factor, that determines to what extent a new sample X_n affects the posterior mean $\hat{\mu}_n$, is determined by γ which increases with the surprise $S_{corr}(X_n)$ of that sample (Fig 3.2B), i.e.,

$$\gamma = \sqrt{\frac{mS_{corr}(X_n)}{1 + mS_{corr}(X_n)}}, \quad S_{corr}(X_n) = \frac{(X_n - \hat{\mu}_{n-1})^2}{2\sigma^2_x}. \tag{3.8}$$

Note that in this example, the confidence-corrected surprise measure is related to the *normalized unsigned prediction error* $|X_n - \hat{\mu}_{n-1}|/\sigma_x$. This outcome of our SMiLe-update is consistent with recent approaches in reward learning that suggest to use reward prediction errors scaled by standard deviation or variance [Preuschoff and Bossaerts, 2007].

Results. The confidence-corrected surprise increases suddenly in response to the samples immediately after the change points, as they are unexpected under the current prior. As a consequence, surprising samples increase the influence of a new data sample on the posterior mean (Fig 3.2B). We can compare our surprise modulated belief update [Eqs (3.7) and (3.8)] with a delta-rule [Eq (3.7)] with *constant* weighting factor γ. To enable a fair comparison we consider two situations: (i) we arbitrarily fix γ at 0.5 or (ii) for a given hazard rate H, we first search for the optimal value of fixed γ so as to minimize the estimation error (Fig 3.2C). We find that our surprise-modulated belief update outperforms the delta-rule with *any* constant learning rate (Fig 3.2D). This clearly shows that an adaptive learning rate is preferable to a fixed learning rate.

We also compared our proposed algorithm with a delta-rule that approximates the optimal Bayesian solution [Nassar et al., 2010]. In the optimal model, the subject knows a-priori that the mean will change at unknown points in time, i.e., the subject makes use of a hierarchical statistical model of the world. The algorithm proposed in [Nassar et al., 2010] provides an efficient approximate solution to estimate the parameters of the hierarchical model. In this algorithm, the subject increases the learning rate as a function of the probability of encountering a change point at a given time step. This probability requires knowledge or online estimation of the hazard rate, which indicates how frequently change points occur. Although our surprise-modulated belief update does not outperform the approximate Bayesian delta-rule, the difference in performance is, in most cases, not significant (see Fig 3.2D). In other words, our method, which does not require any information about the hazard rate, can nearly reach the quality of the optimal Bayesian solution, with significantly reduced computational complexity. Note that the SMiLe rule is not designed for (almost) stationary environments where no fundamental change in context occurs. Therefore, in the case where the true mean is constant (low hazard rate), the SMiLe rule results in increased volatility in estimation. This is why the difference in performance of SMiLe and the optimized Bayesian delta-rule becomes more evident for smaller hazard rates than bigger ones (see Fig 3.2D).

Maze exploration

Task. The maze exploration task is similar to tasks used in behavioral neuroscience and robotics [Morris, 1984, Gillner and Mallot, 1998, Nelson et al., 2004, Rezende and Gerstner, 2014]. There are two environments $\mathscr{A}$ and $\mathscr{B}$, each composed of the same uniquely labeled (e.g., by colors or cue cards) rooms. $\mathscr{A}$ and $\mathscr{B}$ only differ in the topology / spatial arrangement of rooms (see Fig 3.3). Neighboring rooms are connected and accessible through doors. Initially, the agent is placed into either $\mathscr{A}$ or $\mathscr{B}$. At each time step, a door of the current room opens and the agent moves into the adjacent room, thus exploring the environment. After a random exploration time the environment is switched. Once it is changed, the agent must quickly adapt to the new environment. Note that this task differs from a reinforcement learning task because the task at hand just consists of the *exploration* phase. In particular, there is no reward involved in learning.

Model. We model the knowledge of the environment by a learning agent that updates a set of parameters $\alpha(s,\check{s}) \geq 1$ used for describing its belief about *state transitions* from $s \in \{1,2,\ldots,16\}$ to $\check{s} \in \{1,2,\ldots,16\}\backslash s$, where 16 is the number of rooms. More precisely, an agent's belief about how likely it is to visit $\check{s}$, given the current state s, is modeled by a *Dirichlet distribution* parametrized by a *vector* of parameters $\vec{\alpha}(s) \in \mathbb{R}^{15}$. The components of the vector $\vec{\alpha}(s)$ are denoted as $\alpha(s,\check{s})$.

The surprise-modulated belief update (Algorithm 1), with the Dirichlet distribution inserted, yields Algorithm 2 for the maze exploration task (see **Materials and Methods** for derivation). Immediately after a transition from the current state s to the next state s', the posterior belief q_γ obtained by the SMiLe rule [Eq (3.4)] is a Dirichlet distribution $\vec{\alpha}_{new}(s)$ with components $\alpha_{new}(s,\check{s}) = \gamma(1+[\check{s}=s']) + (1-\gamma)\alpha_{old}(s,\check{s})$,

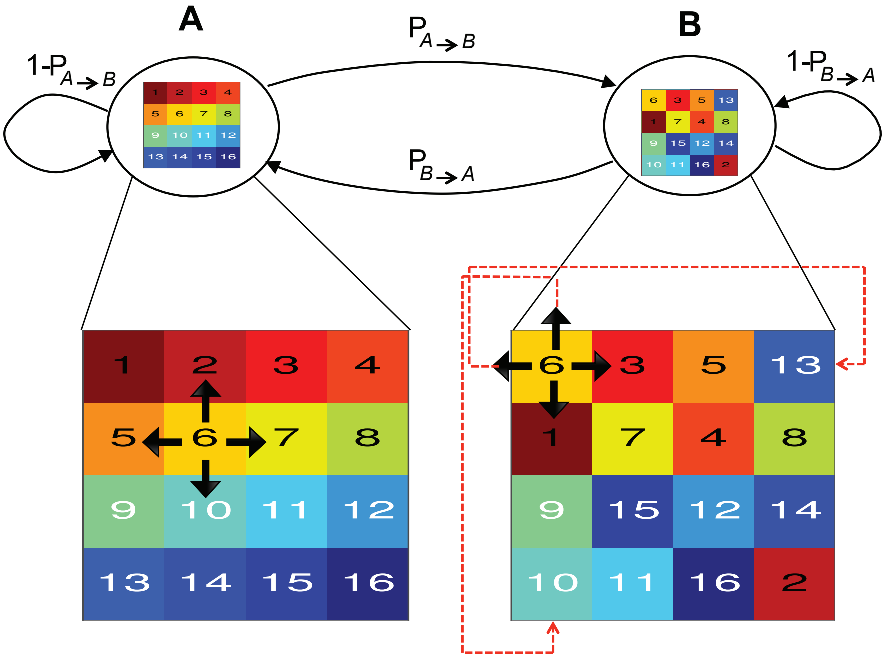

Figure 3.3 – **Maze exploration task.** Environments $\mathscr{A}$ (left) and $\mathscr{B}$ (right) both consist of 16 rooms, but differ in topology. At each time step, one of the four available doors (up, down, right, left) in the current room (e.g. $s = 6$) is randomly opened (with probability 0.25). While the learning agent is in environment $\mathscr{A}$, the environment may change to $\mathscr{B}$ with probability $P_{\mathscr{A}\to\mathscr{B}} \leq 0.1$ in the next time step of duration Δt. Similarly, $P_{\mathscr{B}\to\mathscr{A}}$ indicates the environment switches from $\mathscr{B}$ to $\mathscr{A}$. Therefore, as the agent starts moving out of state $s = 6$, depending on the current environment and switch probabilities $P_{\mathscr{A}\to\mathscr{B}}$ and $P_{\mathscr{B}\to\mathscr{A}}$, it will end up in environment $\mathscr{A}$ (i.e., $s' \in \{2, 10, 7, 5\}$) or $\mathscr{B}$ (i.e., $s' \in \{10, 1, 3, 13\}$). The duration of a stay in environment $\mathscr{A}$ is therefore exponentially distributed with mean $\tau_{\mathscr{A}} = \Delta t / P_{\mathscr{A}\to\mathscr{B}}$, where the parameter $\tau_{\mathscr{A}}$ determines the *time scale of stability* in environment $\mathscr{A}$, i.e., for larger $\tau_{\mathscr{A}}$ an agent has more time for adapting to $\mathscr{A}$ after a change point. The *expected fraction of time spent in total* within environment $\mathscr{A}$ is equal to $\psi_{\mathscr{A}} = P_{\mathscr{B}\to\mathscr{A}} / (P_{\mathscr{B}\to\mathscr{A}} + P_{\mathscr{A}\to\mathscr{B}})$. Note that $\tau_{\mathscr{A}}$ and $\psi_{\mathscr{A}}$ are two free parameters that we can change to study how the agent performs in different circumstances (e.g., see Fig 3.7).

that can be written as a *weighted average* of the parameters of the prior belief π_0 (i.e., $\alpha_{old}(s, \check{s})$) and those of the scaled likelihood $\hat{p}_X$ (i.e., $1 + [\check{s} = s']$). Here, $[\check{s} = s']$ indicates a number that is 1 if the condition in square brackets is satisfied, and 0 otherwise.

In order to see how well our proposed surprise-modulated belief update algorithm performs in this task, we compare it with a naive Bayesian learner and an online expectation-maximization (EM) algorithm [Mongillo and Deneve, 2008]. While in the former the agent assumes that there is only a single stable, but stochastic environment, the latter benefits from knowing the true hidden Markov model (HMM) of the task and approximates the optimal hierarchical Bayesian solution (see **Materials and Methods**).

Algorithm 2 Surprise-modulated belief update for the maze exploration task

```
$N \leftarrow$ number of data samples
$\alpha(s,\check{s}) = 1, \quad \forall s \in \{1,2,...,16\}, \check{s} \in \{1,2,...,16\} \backslash \{s\}$ (a uniform prior belief)
$m \leftarrow 0.1$ (subject-dependent)
Start in state $s$
for $n$: 1 to $N$ do
    # at this time step we only update the parameters that describe state transitions from the current state $s$ to all possible next states $\check{s} \in \{1,2,...,16\} \backslash \{s\}$. The prior belief, for the state $s$, is $\pi_0 \sim Dir(\mathbf{a})$, $\mathbf{a} \in \mathbb{R}^{15}$, $\mathbf{a}(\check{s}) = \alpha(s,\check{s})$.
    $X_n : s \rightarrow s'$ (a new transition is observed)
    # the scaled likelihood is $\hat{p}_X \sim Dir(\mathbf{b}), \mathbf{b} \in \mathbb{R}^{15}$, $\mathbf{b}(\check{s}) = 1 + [\check{s} = s']$
    (i) $S_{corr}(X_n;\pi_0) = D_{KL}[Dir(\mathbf{a})||Dir(\mathbf{b})]$
    (ii-a) $B_{max}(X_n) = D_{KL}[Dir(\mathbf{b})||Dir(\mathbf{a})]$
    (ii-b) $B(X_n) = \frac{mS_{corr}(X_n;\pi_0)}{1+mS_{corr}(X_n;\pi_0)} B_{max}(X_n)$
    (iii) find $\gamma$ by solving $D_{KL}[Dir(\gamma\mathbf{b} + (1-\gamma)\mathbf{a})||Dir(\mathbf{a})] = B(X_n)$
    (iv) $\alpha(s,\check{s}) \leftarrow (1-\gamma)\alpha(s,\check{s}) + \gamma(1 + [\check{s} = s'])$
Return $\alpha(s,\check{s}), \forall s, \check{s}$;
```

Note 1: $D_{KL}[Dir(\mathbf{m})||Dir(\mathbf{n})] = \ln\Gamma(\sum_{\check{s}}\mathbf{m}(\check{s})) - \ln\Gamma(\sum_{\check{s}}\mathbf{n}(\check{s})) - \sum_{\check{s}}\ln\Gamma(\mathbf{m}(\check{s})) + \sum_{\check{s}}\ln\Gamma(\mathbf{n}(\check{s})) + \sum_{\check{s}}(\mathbf{m}(\check{s}) - \mathbf{n}(\check{s}))(\Psi(\mathbf{m}(\check{s})) - \Psi(\sum_{\check{s}}\mathbf{m}(\check{s})))$.
Note 2: $\Gamma(.)$ and $\Psi(.)$ denote the *gamma* and *digamma* functions, respectively. $[\check{s} = s']$ denotes the Iverson bracket, a number that is 1 if the condition in square brackets is satisfied, and 0 otherwise.

Results. Similar to the Gaussian mean estimation task, surprise is initially high and slowly decreases as the agent learns the topology of the environment (Fig 3.4A). When the environment is switched, the sudden increase in the surprise signal (Fig 3.4A) causes the parameter γ to increase (Fig 3.4B). This is equivalent to discounting previously learned information and results in a quick adaptation to the new environment. To quantify the adaptation to the new environment, we compare the state transition probabilities of the current model with the true transition probabilities of the two environments. We find that the estimation error of the state transition probabilities in the new environment is quickly reduced after the switch points (Fig 3.4C). Following a change point, the model uncertainty, measured as the entropy of the current belief about the state transition probabilities, increases indicating that the current model of the topology is inaccurate (Fig 3.4D). A few time steps later the uncertainty slowly decreases, indicating increased confidence in what is learned in the new environment.

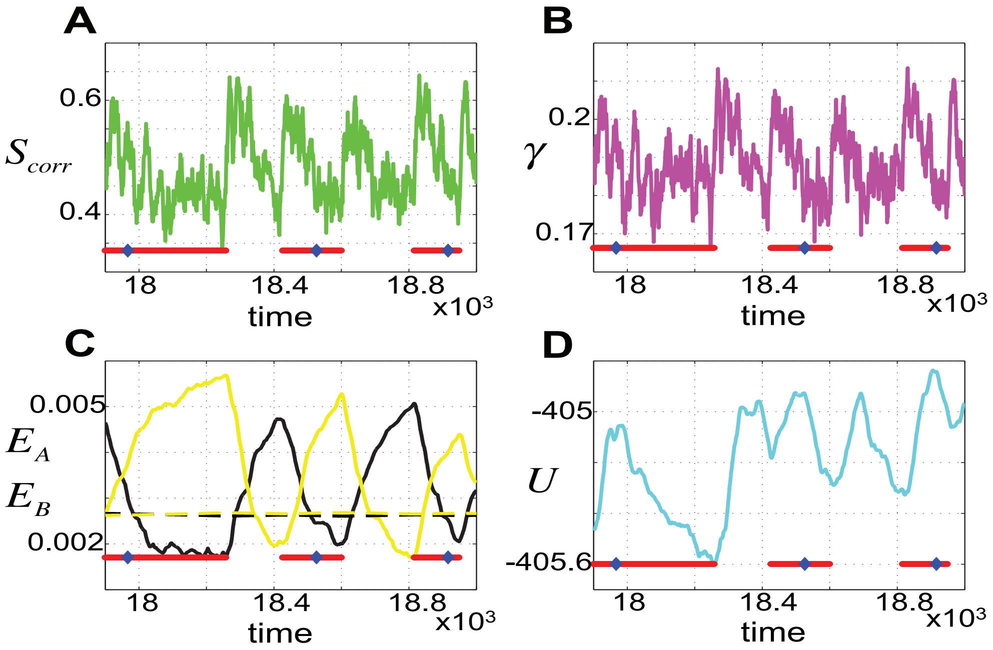

Figure 3.4 – **Time-series of relevant signals in the surprise-modulated belief update (Algorithm 2) applied to the maze exploration task.** All the curves have been smoothed with an exponential moving average (EMA) with a decay constant 0.1. The plots are shown for 1100 time steps (horizontal axis) toward the end of a simulation with 20000 time steps. The agent visits environments $\mathscr{A}$ and $\mathscr{B}$ *equally often* and spends *on average* 200 time steps in each environment before a switch occurs. Red bars indicate the time that the agent explores environment $\mathscr{A}$. Blue diamonds indicate 100 time steps after a change point from $\mathscr{B}$ to $\mathscr{A}$. **A.** Confidence-corrected surprise S_{corr} [Eq (2.8)] (green) increases at switch points and decreases (with fluctuations) till the next change point. **B.** The parameter γ (magenta) increases with the surprise at the change points and causes the next data samples to be more effective on belief update than the samples before the change point. **C.** The estimation errors for the transition matrix $\hat{T}$, $E_{\mathscr{A}}[t] = ||\hat{T}[t] - T_{\mathscr{A}}||_2 = 256^{-1}\sum_{s,s'}[\hat{T}[t](s,s') - T_{\mathscr{A}}(s,s')]^2$ (solid black) and $E_{\mathscr{B}}[t] = ||\hat{T}[t] - T_{\mathscr{B}}||_2$ (solid yellow) while in environment $\mathscr{A}$ and $\mathscr{B}$, respectively, indicate a rapid adaptation to the new environment after the change points. The dashed black and yellow lines correspond to the estimation errors $E_{\mathscr{A}}$ and $E_{\mathscr{B}}$, respectively, when the naive Bayes rule (as a control experiment) is used for belief update. The naive Bayes rule converges to a stationary solution (no significant change in the estimation error after a switch of environment). **D.** The model uncertainty (light blue) increases for a few time steps following a change in the environment, an alert that the current model might be wrong. It then starts decreasing as the agent becomes more certain in the new environment.

If we look more closely at the model parameters, we find that the surprise-modulated belief update (Algorithm 2) enables the agent to adjust the estimated state transition probabilities. In Fig 3.5 we compare the estimated and the true transition probabilities 100 time steps after a switch. Given that the environment is characterized by 64 different transitions (in a space of $16 \times 15 = 240$ potential transitions), 100 time steps allow an agent to explore only a fraction of the potential transitions. Nevertheless, 100 time steps after a switch, the matrix of transition probabilities already resembles that of the present environment (Figs 3.5C and 3.5D).

The surprise-modulated belief update is a method of quick learning. How well does our SMiLe update rule perform relative to other existing models? We compared it with two well-known models. First, we compared to a naive Bayesian learner which tries to estimate the 240 state transition probabilities using Bayes rule. Note that, by construction, the naive Bayesian learner is not aware of the switches between the environments. Second, we compared to a hierarchical statistical model that reflects the architecture of the *true world* as in Fig 3.3. The task is to estimate the 2×240 state transitions in the two environments as well as transition probabilities between the environments $p_{\mathscr{A}\to\mathscr{B}}$ and $p_{\mathscr{B}\to\mathscr{A}}$ by an online EM algorithm.

For the naive Bayesian learner, we find that its behavior indicates a steady increase in certainty, regardless of how surprising the samples are. In other words, it is incapable of changing its belief after it has sufficiently explored the environments (Fig 3.4C). The state transition probabilities are estimated by averaging over the true parameters of both environments, where the weight of averaging is determined by the fraction of time spent in the corresponding environment (Figs 3.5E and 3.5F).

The comparison of our surprise-modulated belief update with the online EM algorithm for the hierarchical Bayesian model associated with the changing environments provides several insights (see Fig 3.6). First, already after less than 1000 time steps, the estimation error for environment $\mathscr{A}$ during short episodes in environment $\mathscr{A}$ drops below $E_{\mathscr{A}} = 0.002$. Only after 10000 time steps, the online EM algorithm achieves the same level of accuracy. While the solution of the SMiLe rule in the long run is not as good, our algorithm benefits from a reduced computational complexity and simpler implementation.

To further investigate the ability of an agent to adapt to the new environment after a switch, we analyzed performance as a function of two free parameters that control the setting of the task: (i) the fraction of time spent in environment $\mathscr{A}$, and (ii) the average time spent in environment $\mathscr{A}$ before a switch to $\mathscr{B}$ occurs. To do so, we calculate the average estimation error in state transition probabilities 64 time steps after a switch occurs. We consider only those switches after which the agent stays in that environment for *at least* 64 time steps. Note that 64 is the minimum number of time steps that is required to ensure that all possible transitions from 16 room to their 4

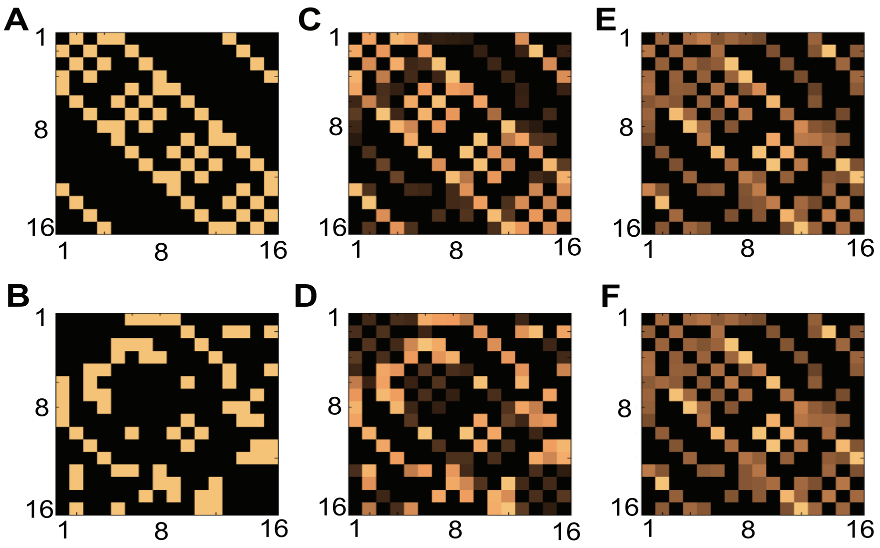

Figure 3.5 – **True and estimated state transition probabilities in the maze exploration task.** The color intensity for each entry (s, s') represents the probability of transition from a current state s (row) to a next state s' (column). **A.** The *true* state transition probability matrix $T_{\mathscr{A}}(s, s')$ in environment $\mathscr{A}$. Each row $T_{\mathscr{A}}(s, :)$ has only four non-zero entries (small squares with the light brown color) whose position indicate the neighboring rooms of state s in environment $\mathscr{A}$. Note that $\sum_{\check{s}} T_{\mathscr{A}}(s, \check{s}) = 1\ , \forall s$. **B.** The true state transition probability matrix $T_{\mathscr{B}}(s, s')$ for the environment $\mathscr{B}$ which has a different topology compared to $\mathscr{A}$. **C.** The *estimated* state transition probability matrix $\hat{\mathscr{T}}_{\mathscr{A}}$ when the surprise-modulated Algorithm 2 is used for belief update. $\hat{\mathscr{T}}_{\mathscr{A}} = K^{-1} \sum_{k=1}^{K} \hat{T}[t^k_{\mathscr{B}\to\mathscr{A}} + 100]$ is calculated by averaging the estimated transition matrix $\hat{T}[t]$ at 100 time steps after each of K change points $t^k_{\mathscr{B}\to\mathscr{A}}$. Here, $t^k_{\mathscr{B}\to\mathscr{A}}$ denotes the k-th time that the environment is changed from $\mathscr{B}$ to $\mathscr{A}$ and *has remained unchanged* for at least the next 100 time steps (relevant time points are indicated by blue diamonds in Fig 3.4). The similarity between $\hat{\mathscr{T}}_{\mathscr{A}}$ and $T_{\mathscr{A}}$ indicates that Algorithm 2 enables the agent to quickly adapt to environment $\mathscr{A}$ once a switch from $\mathscr{B}$ to $\mathscr{A}$ occurs. **D.** The estimated transition matrix $\hat{\mathscr{T}}_{\mathscr{B}}$ (similarly defined as $\hat{\mathscr{T}}_{\mathscr{A}}$ but for environment $\mathscr{B}$) when Algorithm 2 is used for belief update. Note its similarity to the true matrix $T_{\mathscr{B}}$. **E-F.** The estimated state transition probability matrices $\hat{\mathscr{T}}_{\mathscr{A}}$ (top) and $\hat{\mathscr{T}}_{\mathscr{B}}$ (bottom) when the naive Bayesian method (as a control experiment) is used for belief update. A Bayesian agent does not adapt well to the new environment after a switch occurs, because it learns a weighted average of true transition matrices $T_{\mathscr{A}}$ and $T_{\mathscr{B}}$, where the weight is proportional to the fraction of time spent in each environment. Since both environments are visited equally in this experiment, the estimated quantities approach $(T_{\mathscr{A}} + T_{\mathscr{B}})/2$.

neighbors *could* occur. A smaller estimation error for a given pair of free parameters indicates a faster adaptation to the new environment for that setting.

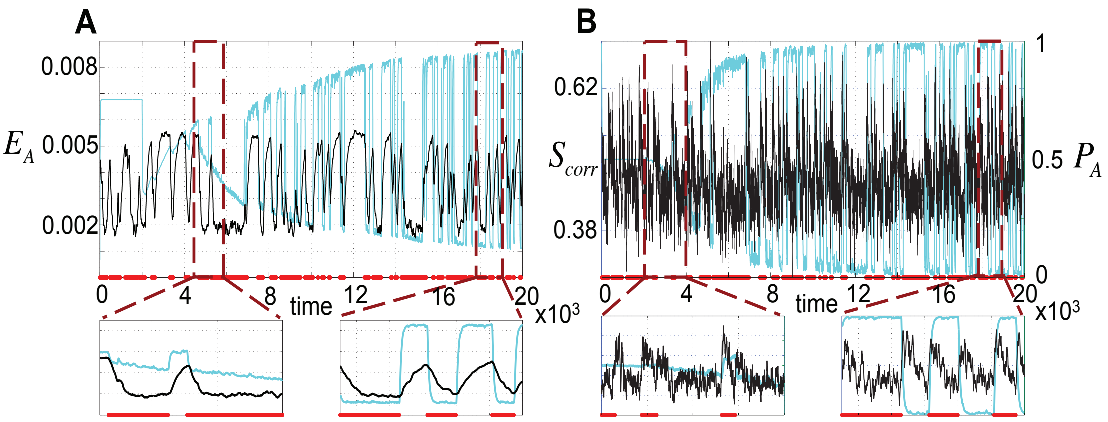

Figure 3.6 – **Comparison of surprise-modulated belief update with an online EM algorithm for the hierarchical Bayesian model. A.** The estimation error $E_{\mathscr{A}}$ (vertical axis) of state transition probabilities within environment $\mathscr{A}$ versus time (horizontal axis), for surprise-modulated belief update (black) and online EM learner (blue). Bottom plots depict zooms during the early (left) and late (right) phases of a simulation of 20000 time steps. In the early phase of learning (bottom left), the surprise-modulated belief update enables the agent to quickly learn model parameters after a switch to environment $\mathscr{A}$ (indicated by red bars). In the late phase of learning (right), however, the online EM algorithm adapts to the new environment faster and more accurately than the surprise-modulated belief update. **B.** The inferred probability $P_{\mathscr{A}}$ of being in environment $\mathscr{A}$ (blue, right vertical axis) used in the online EM algorithm, and the confidence-corrected surprise S_{corr} (black, left vertical axis) used in the surprise-modulated belief update.

We found that the surprise-modulated belief enables an agent to quickly readjust its estimation of model parameters, even if the fraction of time spent in an environment is relatively short. In that sense, it behaves similarly to the approximate hierarchical Bayesian approach (online EM algorithm). This is not, however, the case for a naive Bayesian learner whose estimation error in each environment depends on the fraction of time spent in the corresponding environment (see Fig 3.7).

The naive Bayesian learner suffers from low accuracy in estimation and cannot adapt to environmental changes. A full hierarchical Bayesian model, however, requires prior information about the task and is computationally demanding. For example, the computational load of the hierarchical Bayesian model increases with the number N of environments between which switching occurs. The surprise-modulated belief update, however, balances accuracy and computational complexity: computational complexity remains, by construction, independent of the number of switched environments.

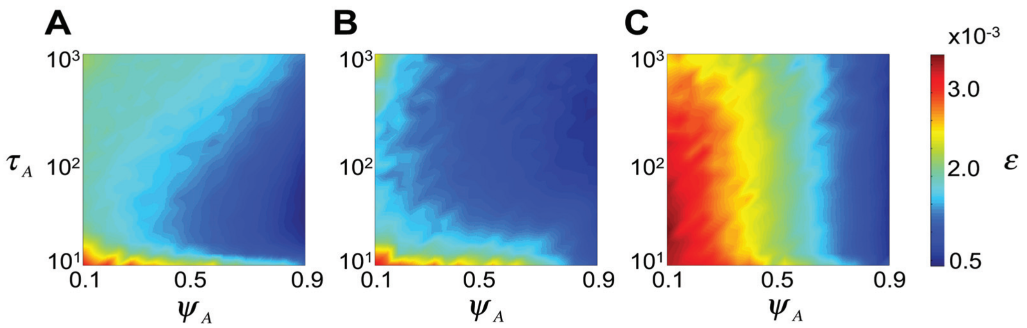

Figure 3.7 – **The estimation error ϵ in the maze exploration task,** as a function of (1) the average time spent in environment $\mathscr{A}$ before a switch to environment $\mathscr{B}$ ($\tau_{\mathscr{A}} = \Delta t / p_{\mathscr{A}\to\mathscr{B}}$, vertical axis) and (2) the fraction of time spent in environment $\mathscr{A}$ ($\psi_{\mathscr{A}} = P_{\mathscr{B}\to\mathscr{A}} / (P_{\mathscr{B}\to\mathscr{A}} + P_{\mathscr{A}\to\mathscr{B}})$, horizontal axis). **A.** The average estimation error (of state transition probabilities), 64 time steps after a switch from $\mathscr{B}$ to $\mathscr{A}$, when surprise-modulated belief update (Algorithm 2) is used for learning. The spread of blue color (lower estimation error) illustrates that the surprise-modulated belief update enables an agent to quickly adapt to the environment visited after a switch. For each pair $(\tau_{\mathscr{A}}, \psi_{\mathscr{A}})$, the simulation is repeated for 20 episodes, each consisting of 20000 time steps. In each episode a different rearrangement of rooms for building environment $\mathscr{B}$ is used to make sure that the result is not biased by a specific choice of this environment. **B.** The average estimation error when the online EM algorithm is used for learning the hierarchical statistical model. **C.** The average estimation error when the naive Bayesian learner is used for belief update. The estimation error for this model is mainly determined by the fraction of time spent in environment $\mathscr{A}$ (i.e., $\psi_{\mathscr{A}}$). The estimation error decreases with the time spent in environment $\mathscr{A}$, regardless of the time scale of stability determined by $\tau_{\mathscr{A}}$.

3.3 Discussion

We proposed a new framework for surprise-driven learning. There are two components to this framework: (i) a confidence-adjusted surprise measure to capture environmental statistics as well as subjective beliefs, and (ii) the surprise-minimization learning rule, or SMiLe-rule, which dynamically adjusts the balance between new and old information without prior assumptions about the temporal statistics in the environment. Within this framework, surprise is a single subject-specific variable that determines a subject's propensity to modify existing beliefs. This algorithm is suitable for learning in complex environments that are either stable or undergo gradual or sudden changes. The latter are signalled by high surprise and result in placing more weight on new information. The significance of the proposed method is that it neither requires knowledge of the full Bayesian model of the environment nor any prior assumption about the temporal statistics in the environment. Moreover, it provides a simple framework that could potentially be implemented in a neurally plausible way using probabilistic population codes [Ma et al., 2006, Beck et al., 2008].

New versus old information

The proposed algorithm's performance is primarily driven by two features: (i) the algorithm adaptively increases the influence of new data on the belief update as a function of how surprising the data was; and (ii) the algorithm increases model uncertainty in the face of surprising data thus increasing the influence of new data on current *and* future belief updates. The importance of the first point has been recognized and incorporated previously [Nassar et al., 2012, Pearce and Hall, 1980]. The second point is particularly worth noting: a surprising sample not only signals a potential change, it also signals that our current model may be wrong, so that we should be *less* certain about its accuracy. This increase in model uncertainty implies discounting the influence of past information in current and future belief updates.

Both humans and animals adaptively adjust the relative contribution of old and newly acquired data on learning [Behrens et al., 2007, Nassar et al., 2012, Krugel et al., 2009, Pearce and Hall, 1980] and rapidly adapt to changing environments [Pearce and Hall, 1980, Wilson et al., 1992, Holland, 1997]. Standard Bayesian and reinforcement learning models in humans [Tenenbaum and Griffiths, 2001] or animals [Dayan et al., 2000, Kakade and Dayan, 2002] assume a stable environment and are slow to adapt to sudden changes in the environment. To quickly learn in dynamic environments, models need to include a way to detect and respond to sudden changes.

A full (hierarchical) Bayesian approach works only if the subject is aware of the correct model of the task, (e.g., the time scale of change in the environment or the number of environments between which switches occur). Calculating the probability of a change point in a Gaussian estimation task [Nassar et al., 2010], estimating the volatility of the environment in a reversal learning task [Behrens et al., 2007], and dynamically forgetting the past information with a controlled time constant [Rüter et al., 2012] are all examples of addressing learning in changing environments without explicit knowledge of the full Bayesian model.

In changing environments, hierarchical Bayesian models outperform the standard delta-rule with a fixed learning rate. However, hierarchical models either make assumptions about how fast the world is changing on average or about the underlying data generating process, in order to accurately detect a change in the environment. While our proposed surprise-based algorithm may not be theoretically optimal, it approximates the optimal (hierarchical) Bayesian solution without making any such assumption.

Model uncertainty

The ability of our proposed method to increase model uncertainty solves a common problem in standard Bayesian learning, namely, a model uncertainty or a learning rate approaching zero when the number of data samples increases. This is particularly prominent in Bayes' rule which is derived under the assumption of *stationarity* and which thus reduces posterior uncertainty in each step no matter how surprising a sample is. The SMiLe rule [Eq (3.4)] guarantees that a small model uncertainty remains even after a long stationary period. This remaining uncertainty ensures that an organism can still detect a change even after having spent an extensive amount of time in a given environment (see Fig 3.4C). One might argue, that reducing the learning rate to zero after extensive training is desirable under certain conditions as it corresponds to the well-documented phenomenon of overtraining whereby an organism no longer responds to changes in goal value. We would argue that this insensitivity is a consequence of behavioral control being handed over to the habitual system and thus to a different neural substrate [Balleine and O'Doherty, 2010, Balleine and Dickinson, 1998, Redgrave et al., 2010].

Potential applications

Surprise minimization is a more general approach to learning than learning by reward prediction error. Recent approaches in reward learning suggest using a scaled reward prediction error [Preuschoff and Bossaerts, 2007]. A recurring problem in reward-based learning is the observation that subjects use different learning rates on a trial-by-trial basis even in stable environments. Researchers typically assume an average learning rate for fitting data. Note that in our approach, the learning rate varies naturally as a function of the last data point (as it should) while keeping the subject-specific parameter m constant.

Note that both confidence-corrected surprise and the SMiLe rule have wide-reaching implications outside the framework presented here. On the one hand, our surprise measure can not only *modulate* learning, but can be used as a *trigger* signal for an algorithm that needs to choose between several uncertain states or actions as is the case in change point detection [Nassar et al., 2010, Wilson et al., 2013, Rüter et al., 2012], memory and cluster formation [Gershman and Niv, 2015], exploration/exploitation tradeoff [Cohen et al., 2007, Jepma and Nieuwenhuis, 2011], novelty detection [Knight et al., 1996, Bishop, 1994], and network reset [Bouret and Sara, 2005]. On the other hand, the SMiLe-rule could add flexibility in learning and replace existing learning algorithms in scenarios where dynamically balancing old and new information is desired. This includes fitting γ to behavioral data without computing surprise or controlling γ by something other than surprise. Replacing the full Bayesian model of a learning task in changing environment with the SMile rule simplifies calculations, which makes the SMiLe-framework suitable for fitting relevant parameters to behavioral data.

Relation to the free energy principle and variational methods

Although the free energy principle [Friston, 2010] is a contemporary theory of surprise minimization, the idea behind it differs from our proposed surprise modulated belief update. In the free energy principle (or in variational Bayesian methods in general) the aim is to "approximate" the Bayesian posterior $q(\theta) \propto p(X|\theta)\pi_0(\theta)$ that is difficult or intractable to be directly derived. In fact, given a candidate posterior distribution $q(\theta)$ a variational bound $J(X;q) = D_{KL}[q||\pi_0] - \langle \ln p(X|\theta)\rangle_q$ on the Shannon surprise $S_{Sh}(X;q)$ [Eq. (2.1)] is introduced such that $S_{Sh}(X;q) \leq J(X;q)$. Then the aim is to minimize the bound $J(X;q)$ with respect to q. The minimum of the variational bound $J(X;q)$ with respect to q simply recovers the posterior Bayes $q(\theta) \propto p(X|\theta)\pi_0(\theta)$.

In the surprise-modulated belief update, however, the aim is *not* to approximate the Bayesian posterior. Instead, we would like to have a belief-update rule that enables us to dynamically adjust the balance between the influence of the prior and the likelihood for deriving the posterior, using a surprise-related signal. Although our approach for deriving the SMiLe rule Eq. (3.4) was different from the variational method, we can rederive an equation that is somewhat equivalent to the SMiLe rule but using the variational method. The constraint $D_{KL}[q||\pi_0] \leq B$ that we introduced for deriving the SMiLe posterior $q(\theta) \propto p(X|\theta)^{\gamma}\pi_0(\theta)^{1-\gamma}$ (see Eqs. (3.3), (3.4)) can also be imposed on the variational bound $J'(X;q) = \beta D_{KL}[q||\pi_0] - \langle \ln p(X|\theta)\rangle_q$ with the help of a Lagrange multiplier β. By minimizing the variational bound $J'(X;q)$ we can interpolate between the prior ($\beta \to \infty$) and the scaled likelihood ($\beta \to 0$).

Experimental predictions

The *noradrenergic system* has emerged as a prime candidate for signaling unexpected uncertainty and surprise. We predict that in experiments with changing environments, the activity of NE should exhibit a high correlation with the confidence-corrected surprise signal. Note that Acetylcholine (ACh), on the other hand, is a candidate neuromodulator for encoding expected uncertainty [Yu and Dayan, 2005] and thus is linked to the model uncertainty (although it might also be linked to other forms of uncertainty such as environmental stochasticity).

A variety of experimental findings are consistent with and can be explained by our definition of confidence-corrected surprise and the SMiLe rule. It has been suggested that ACh and NE interact in a complex manner [Yu and Dayan, 2005]. For instance, the effectiveness of NE in controlling learning and detecting contextual changes is gated by ACh [Yu and Dayan, 2005]. This is consistent with our hypothesis that if an agent is uncertain about its current model of the world, unexpected events are perceived as less surprising than when the agent is almost certain about its model (the idea behind the confidence-corrected surprise). The impairment of adaptation to contextual changes due to NE depletion [Sara, 1998] can be explained by the incapability of subjects to respond to surprising events signaled by NE. The absence/suppression of ACh (low model uncertainty) implies little or no variability of the environment so that even small prediction error signals are perceived as surprising [Jones and Higgins, 1995], consistent with the excessive activation of NE system in such situations.

Moreover, there is empirical evidence that NE and ACh both affect synaptic plasticity in the cortex and the hippocampus [Gu, 2002, Bear and Singer, 1986], suppress cortical processing [Kimura et al., 1999, Kobayashi et al., 2000], and facilitate information processing from thalamus to the cerebral cortex [Gil et al., 1997, Hasselmo et al., 1996, Hsieh et al., 2000]. This is consistent with our theory that surprise balances the influence of newly acquired data (thalamocortical pathway) and old information (corticocortical pathway) during belief update.

In summary, we proposed a measure of surprise and a surprise-modulated belief update algorithm that can be used for modeling how humans and animals learn in changing environments. Our results suggest that the proposed method can approximate an optimal hierarchical Bayesian learner, but with significantly reduced computational complexity. Our model provides a framework for future studies on learning with surprise. These include computational studies, such as how the proposed model can be neurally implemented, and neurobiological studies, such as unraveling the interaction between different neural circuits that are functionally involved in learning under surprise.

3.4 Materials and Methods

Derivation of the SMiLe rule.

We note that the KL divergence $D_{KL}[a||b]$ is convex with respect to the first argument a. Therefore, both the objective function $S_{corr}(X;q)$ in Eq (2.8) and the constraint $D_{KL}[q||\pi_0] \leq B$ in the optimization problem in Eq (3.3) are convex with respect to q, which ensures the existence of the optimal solution. In the following, small numbers above an equality sign refer to equations in the main text.

We solve the constraint optimization by introducing a non-negative Lagrange multiplier $\lambda^{-1} \geq 0$ and a Lagrangian

$$\begin{aligned} \mathbb{L}(q,\lambda) &= S_{corr}(X;q) - \frac{1}{\lambda}(B - D_{KL}[q||\pi_0]) \\ &\overset{(2.6)}{=} \langle -\ln p(X|\theta) + \ln q(\theta) + \frac{1}{\lambda}\ln\frac{q(\theta)}{\pi_0(\theta)}\rangle_q - \frac{B}{\lambda} + \ln||p||, \end{aligned} \tag{3.9}$$

where $\langle . \rangle_q$ denotes the average with respect to q. Similar to the standard approach that is used in support vector machines [Schölkopf and Smola, 2002], the Lagrangian $\mathbb{L}$ defined in Eq (3.9) must be minimized with respect to the primal variable q and maximized with respect to the dual variable λ (i.e., a saddle point must be found). Therefore the constraint problem in Eq (3.3) can be expressed as

$$\arg\min_q \max_{\lambda\geq 0} \mathbb{L}(q,\lambda). \tag{3.10}$$

By taking the derivative of $\mathbb{L}$ with respect to q and setting it equal to zero,

$$\frac{\partial \mathbb{L}}{\partial q} = -\ln p(X|\theta) + \left[1 + \ln q(\theta)\right] + \frac{1}{\lambda}\left[1 + \ln\frac{q(\theta)}{\pi_0(\theta)}\right] = 0, \tag{3.11}$$

we find that the Lagrangian in Eq (3.9) is minimized by the SMiLe rule [Eq (3.4)], i.e., $q(\theta) \propto p(X|\theta)^{\gamma}\pi_0(\theta)^{1-\gamma}$, where γ is determined by the Lagrange multiplier λ:

$$0 \leq \gamma = \frac{\lambda}{\lambda+1} \leq 1. \tag{3.12}$$

Note that the constant $Z(X;\gamma)$ in Eq (3.4) follows from straight normalization of $q(\theta)$ to integral one.

A larger bound $B > B'$ on belief change implies a bigger $\gamma > \gamma'$ in the SMiLe rule.

For $0 < B < B_{max}$ the solution of optimization problem in Eq (3.3) is the posterior q_γ [Eq (3.4)] with $0 < \gamma < 1$ satisfying $D_{KL}[q_\gamma || \pi_0] = B$. In order to prove that $B > B'$ implies $\gamma > \gamma'$, we just need to show that $D_{KL}[q_\gamma || \pi_0]$ is an increasing function of γ and thus its first derivative with respect to γ is always non-negative.

For this purpose, first we need to evaluate the derivative of $q_\gamma(\theta)$, [Eq (3.4)], with respect to γ:

$$\begin{aligned}
\frac{\partial}{\partial\gamma} q_\gamma(\theta) &= \frac{1}{Z(X;\gamma)} \frac{\partial}{\partial\gamma}\big[p(X|\theta)^\gamma \pi_0(\theta)^{1-\gamma}\big] + p(X|\theta)^\gamma \pi_0(\theta)^{1-\gamma} \frac{\partial}{\partial\gamma}\big[\frac{1}{Z(X;\gamma)}\big] \\
&= \frac{1}{Z(X;\gamma)}\big[p(X|\theta)^\gamma \pi_0(\theta)^{1-\gamma} \ln\frac{p(X|\theta)}{\pi_0(\theta)}\big] - \frac{p(X|\theta)^\gamma \pi_0(\theta)^{1-\gamma}}{Z(X;\gamma)^2} \frac{\partial}{\partial\gamma}\big[Z(X;\gamma)\big] \\
&= q_\gamma(\theta) \ln\frac{p(X|\theta)}{\pi_0(\theta)} - q_\gamma(\theta)\frac{1}{Z(X;\gamma)}\big[\int_\Theta \ln\frac{p(X|\theta)}{\pi_0(\theta)}\, p(X|\theta)^\gamma \pi_0(\theta)^{1-\gamma}\, d\theta\big] \\
&= q_\gamma(\theta)\left(\ln\frac{p(X|\theta)}{\pi_0(\theta)} - \langle \ln\frac{p(X|\theta)}{\pi_0(\theta)}\rangle_{q_\gamma}\right).
\end{aligned} \tag{3.13}$$

Note also that

$$\int_\Theta \frac{\partial}{\partial\gamma} q_\gamma(\theta)\, d\theta \overset{(3.13)}{=} \int_\Theta q_\gamma(\theta)\left(\ln\frac{p(X|\theta)}{\pi_0(\theta)} - \langle \ln\frac{p(X|\theta)}{\pi_0(\theta)}\rangle_{q_\gamma}\right) d\theta = 0. \tag{3.14}$$

Then we calculate the derivative of $D_{KL}[q_\gamma||\pi_0]$ with respect to γ:

$$
\begin{aligned}
\frac{\partial}{\partial\gamma} D_{KL}[q_\gamma||\pi_0] \quad &= \quad \int_\Theta \frac{\partial}{\partial\gamma}\big[q_\gamma \ln\frac{q_\gamma(\theta)}{\pi_0(\theta)}\big]\, d\theta \\
&= \quad \int_\Theta \left(\ln\frac{q_\gamma(\theta)}{\pi_0(\theta)}\, \frac{\partial}{\partial\gamma}\big[q_\gamma(\theta)\big] + q_\gamma(\theta)\frac{\partial}{\partial\gamma}\big[\ln\frac{q_\gamma(\theta)}{\pi_0(\theta)}\big]\right)\, d\theta \\
&= \quad \int_\Theta \left(\ln\frac{q_\gamma(\theta)}{\pi_0(\theta)} + 1\right)\, \frac{\partial}{\partial\gamma}\big[q_\gamma(\theta)\big]\, d\theta \\
&\overset{(3.4)}{=} \quad \int_\Theta \left(\gamma \ln\frac{p(X|\theta)}{\pi_0(\theta)} - \ln Z(X;\gamma) + 1\right)\, \frac{\partial}{\partial\gamma}\big[q_\gamma(\theta)\big]\, d\theta \\
&\overset{(3.14)}{=} \quad \gamma\int_\Theta \left(\ln\frac{p(X|\theta)}{\pi_0(\theta)}\right)\, \frac{\partial}{\partial\gamma}\big[q_\gamma(\theta)\big]\, d\theta \\
&\overset{(3.13)}{=} \quad \gamma\int_\Theta \left(\ln\frac{p(X|\theta)}{\pi_0(\theta)}\right)\, \left(\ln\frac{p(X|\theta)}{\pi_0(\theta)} - \langle\ln\frac{p(X|\theta)}{\pi_0(\theta)}\rangle_{q_\gamma}\right) q_\gamma(\theta)\, d\theta \\
&= \quad \gamma\int_\Theta \left(\ln\frac{p(X|\theta)}{\pi_0(\theta)}\right)^2 q_\gamma(\theta)\, d\theta \\
&\qquad - \gamma\,\langle\ln\frac{p(X|\theta)}{\pi_0(\theta)}\rangle_{q_\gamma} \int_\Theta \left(\ln\frac{p(X|\theta)}{\pi_0(\theta)}\right) q_\gamma(\theta)\, d\theta \\
&= \quad \gamma\left(\langle\left(\ln\frac{p(X|\theta)}{q_\gamma(\theta)}\right)^2\rangle_{q_\gamma} - \left(\langle\ln\frac{p(X|\theta)}{\pi_0(\theta)}\rangle_{q_\gamma}\right)^2\right) \\
&= \quad \gamma\ var[\ln\frac{p(X|\theta)}{\pi_0(\theta)}] \geq 0. \qquad (3.15)
\end{aligned}
$$

The impact of a data sample X on belief update increases with γ in the SMiLe rule.

To prove the statement above we need to show that the impact function $\Delta S_{corr}(X;L)$ in Eq (3.2), where L is replaced by the SMiLe rule in Eq (3.4), increases with the parameter γ. In the following we show that the first derivative of the impact function

$\Delta S_{corr}(X;L(\gamma))$ with respect to γ is always non-negative.

$$
\begin{aligned}
\frac{\partial}{\partial\gamma}\Delta S_{corr}(X;L(\gamma)) \quad &= \quad -\frac{\partial}{\partial\gamma}S_{corr}(X;q_\gamma) \overset{(2.8)}{=} \int_\Theta \frac{\partial}{\partial\gamma}\left[q_\gamma \ln\frac{p(X|\theta)}{q_\gamma(\theta)}\right] d\theta \\
&= \quad \int_\Theta \left(\ln\frac{p(X|\theta)}{q_\gamma(\theta)}\,\frac{\partial}{\partial\gamma}\left[q_\gamma(\theta)\right] + q_\gamma(\theta)\frac{\partial}{\partial\gamma}\left[\ln\frac{p(X|\theta)}{q_\gamma(\theta)}\right]\right) d\theta \\
&= \quad \int_\Theta \left(\ln\frac{p(X|\theta)}{q_\gamma(\theta)} - 1\right) \frac{\partial}{\partial\gamma}\left[q_\gamma(\theta)\right] d\theta \\
&\overset{(3.4)}{=} \quad \int_\Theta \left((1-\gamma)\ln\frac{p(X|\theta)}{\pi_0(\theta)} + \ln Z(X;\gamma) - 1\right) \frac{\partial}{\partial\gamma}\left[q_\gamma(\theta)\right] d\theta \\
&\overset{(3.14)}{=} \quad (1-\gamma)\int_\Theta \left(\ln\frac{p(X|\theta)}{\pi_0(\theta)}\right) \frac{\partial}{\partial\gamma}\left[q_\gamma(\theta)\right] d\theta \\
&\overset{(3.13)}{=} \quad (1-\gamma)\int_\Theta \left(\ln\frac{p(X|\theta)}{\pi_0(\theta)}\right)\left(\ln\frac{p(X|\theta)}{\pi_0(\theta)} - \langle\ln\frac{p(X|\theta)}{\pi_0(\theta)}\rangle_{q_\gamma}\right) q_\gamma(\theta)\; d\theta \\
&= \quad (1-\gamma)\int_\Theta \left(\ln\frac{p(X|\theta)}{\pi_0(\theta)}\right)^2 q_\gamma(\theta)\; d\theta \\
&\qquad -\; (1-\gamma)\,\langle\ln\frac{p(X|\theta)}{\pi_0(\theta)}\rangle_{q_\gamma} \int_\Theta \left(\ln\frac{p(X|\theta)}{\pi_0(\theta)}\right) q_\gamma(\theta)\; d\theta \\
&= \quad (1-\gamma)\left(\langle\left(\ln\frac{p(X|\theta)}{q_\gamma(\theta)}\right)^2\rangle_{q_\gamma} - \left(\langle\ln\frac{p(X|\theta)}{\pi_0(\theta)}\rangle_{q_\gamma}\right)^2\right) \\
&= \quad (1-\gamma)\; var[\ln\frac{p(X|\theta)}{\pi_0(\theta)}] \geq 0. \qquad (3.16)
\end{aligned}
$$

A larger reduction in the surprise implies a bigger change in belief.

The minimal value of the Lagrangian $\mathbb{L}(q,\lambda)$ in Eq (3.9) that is achieved by the posterior q_γ in Eq (3.4), obtained by the SMiLe rule, is equal to

$$
\begin{aligned}
\mathbb{L}(q_\gamma,\lambda) &\overset{(3.9)}{=} \langle -\ln p(X|\theta) + \ln q_\gamma(\theta) + \frac{1}{\lambda}\ln\frac{q_\gamma(\theta)}{\pi_0(\theta)}\rangle_{q_\gamma} \overbrace{-\frac{B}{\lambda} + \ln||p||}^{=C} \\
&= \langle -\ln p(X|\theta) + \ln\frac{p(X|\theta)^\gamma \pi_0(\theta)^{1-\gamma}}{Z(X;\gamma)} + \frac{1}{\lambda}\ln\frac{p(X|\theta)^\gamma \pi_0(\theta)^{1-\gamma}}{Z(X;\gamma)\pi_0(\theta)}\rangle_{q_\gamma} + C \\
&= \langle (-1+\gamma+\frac{\gamma}{\lambda})\ln p(X|\theta) + (1-\gamma-\frac{\gamma}{\lambda})\ln\pi_0 - (1+\frac{1}{\lambda})\ln Z(X;\gamma)\rangle_{q_\gamma} + C \\
&= \langle \left(-1+\gamma(1+\frac{1}{\lambda})\right)\ln\frac{p(X|\theta)}{\pi_0(\theta)} - (1+\frac{1}{\lambda})\ln Z(X;\gamma)\rangle_{q_\gamma} + C \\
&= -\frac{1}{\gamma}\ln Z(X;\gamma) + C. \qquad (3.17)
\end{aligned}
$$

Note that we used the equality $\frac{1}{\gamma} = 1 + \frac{1}{\lambda}$, from Eq (3.12), in the last line of Eq (3.17). If the minimizer q_γ is approximated by any other distribution function q, then its corresponding functional value $\mathbb{L}(q,\lambda)$ differs from its minimal value $\mathbb{L}(q_\gamma,\lambda)$ in proportion to the KL divergence $D_{KL}[q||q_\gamma]$. This is because,

$$
\begin{aligned}
\mathbb{L}(q,\lambda) - \mathbb{L}(q_\gamma,\lambda) &\overset{(3.9),(3.17)}{=} \langle -\ln p(X|\theta) + \ln q(\theta) + \frac{1}{\lambda}\ln\frac{q(\theta)}{\pi_0(\theta)}\rangle_q + \frac{1}{\gamma}\ln Z(X;\gamma) \\
&= \frac{1}{\gamma}\langle -\ln p(X|\theta)^\gamma + \ln q(\theta)^\gamma + \ln\left(\frac{q(\theta)}{\pi_0(\theta)}\right)^{\frac{\gamma}{\lambda}} + \ln Z(X;\gamma)\rangle_q \\
&= \frac{1}{\gamma}\langle \ln\frac{q(\theta)^{\gamma(1+\frac{1}{\lambda})} Z(X;\gamma)}{p(X|\theta)^\gamma \pi_0(\theta)^{\frac{\gamma}{\lambda}}}\rangle_q = \frac{1}{\gamma}\langle \ln\frac{q(\theta) Z(X;\gamma)}{p(X|\theta)^\gamma \pi_0(\theta)^{1-\gamma}}\rangle_q \\
&= \frac{1}{\gamma} D_{KL}[q||q_\gamma]. \qquad (3.18)
\end{aligned}
$$

Replacing q with π_0 in Eq (3.18) follows the impact function $\Delta S_{corr}(X;L)$ in Eq (3.2) to be,

$$\begin{aligned}
\Delta S_{corr}(X;L(\gamma)) &= S_{corr}(X;\pi_0) - S_{corr}(X;q_\gamma) \\
&\overset{(3.9)}{=} \mathbb{L}(\pi_0,\lambda) + \frac{1}{\lambda}B - \mathbb{L}(q_\gamma,\lambda) - \frac{1}{\lambda}(B - D_{KL}[q_\gamma||\pi_0]) \\
&= \mathbb{L}(\pi_0,\lambda) - \mathbb{L}(q_\gamma,\lambda) + \frac{1}{\lambda}D_{KL}[q_\gamma||\pi_0] \\
&\overset{(3.18)}{=} \frac{1}{\gamma}D_{KL}[\pi_0||q_\gamma] + \frac{1}{\lambda}D_{KL}[q_\gamma||\pi_0] \\
&\overset{(3.12)}{=} \frac{1}{\gamma}D_{KL}[\pi_0||q_\gamma] + \left(\frac{1}{\gamma} - 1\right)D_{KL}[q_\gamma||\pi_0] \geq 0.
\end{aligned} \tag{3.19}$$

Therefore, the reduction in the posterior surprise is related to the belief changes $D_{KL}[\pi_0||q_\gamma]$ and $D_{KL}[q_\gamma||\pi_0]$ via Eq (3.19). Note that the equality in Eq (3.19) holds if and only if there is no change in the prior belief, i.e., if $q_\gamma = \pi_0$. This happens only if $\gamma = 0$ which is equivalent to the full neglect of the new data point in deriving the posterior belief.

The SMiLe rule for beliefs described by Gaussian distribution.

Suppose we have drawn $n-1$ samples $X_1, ..., X_{n-1}$ from a Gaussian distribution of known variance σ_x^2, but unknown mean. The empirical mean after $n-1$ samples is $\hat{\mu}_{n-1}$.

Assume that the current belief about the mean μ is a normal distribution, i.e., $\pi_0(\mu) \sim \mathcal{N}(\hat{\mu}_{n-1}, \sigma_{n-1}^2)$. Since the likelihood of receiving a new sample X_n is also normal, i.e., $p(X_n|\mu) \sim \mathcal{N}(\mu, \sigma_x^2)$, the posterior belief obtained by the SMiLe rule [Eq (3.4)] is

$$\begin{aligned}
q_\gamma(\mu) &\propto \left(exp\left(-\frac{(X_n-\mu)^2}{2\sigma_x^2}\right)\right)^\gamma \left(exp\left(-\frac{(\mu-\hat{\mu}_{n-1})^2}{2\sigma_{n-1}^2}\right)\right)^{1-\gamma} \\
&\propto exp\left(-\frac{(X_n-\mu)^2}{2(\sigma_x')^2}\right) exp\left(-\frac{(\mu-\hat{\mu}_{n-1})^2}{2(\sigma_{n-1}')^2}\right),
\end{aligned} \tag{3.20}$$

where $(\sigma_x')^2 = \sigma_x^2/\gamma$ and $(\sigma_{n-1}')^2 = \sigma_{n-1}^2/(1-\gamma)$. Because the product of two Gaussians is a Gaussian, we arrive at a posterior distribution $q_\gamma \sim \mathcal{N}(\hat{\mu}_n, \sigma_n^2)$ with the mean $\hat{\mu}_n = w_n X_n + (1-w_n)\hat{\mu}_{n-1}$ (with $w_n = \frac{(\sigma_{n-1}')^2}{(\sigma_x')^2 + (\sigma_{n-1}')^2}$), and the variance $\sigma_n^2 = \left(\frac{1}{(\sigma_x')^2} + \frac{1}{(\sigma_{n-1}')^2}\right)^{-1}$; see [MacKay, 2003] for the exact derivation. Assuming $\sigma_{n-1}^2 = \sigma_x^2$, then $w_n = \gamma$. Moreover, we can evaluate the confidence-corrected surprise to be

$$S_{corr}(X_n;\pi_0) = D_{KL}[\mathcal{N}(\hat{\mu}_{n-1},\sigma^2_{n-1})||\mathcal{N}(X_n,\sigma^2_x)] = \frac{(X_n - \hat{\mu}_{n-1})^2}{2\sigma^2_x}. \qquad (3.21)$$

Note that in Eq (3.21), we used the following equality in Eq (3.22) (assuming $\sigma^2_x = \sigma^2_{n-1}$),

$$D_{KL}[\mathcal{N}(a_1,b_1^2)||\mathcal{N}(a_2,b_2^2)] = \frac{(a_1-a_2)^2}{2b_2^2} + \frac{1}{2}\left(\frac{b_1^2}{b_2^2} - 1 - \ln\frac{b_1^2}{b_2^2}\right). \qquad (3.22)$$

The SMiLe rule for beliefs described by a Dirichlet distribution.

Assume that the current belief about the probability of transition from state $s \in \{1,2,...,D\}$ to all $D-1$ possible next states $\check{s} \in \{1,2,...,D\}\backslash s$ is described by a Dirichlet distribution $\pi_0(\theta_s) \propto \Pi_{\check{s}}\, \theta(s,\check{s})^{\alpha(s,\check{s})-1}$ parametrized by $\alpha_s = \alpha(s,:)$. Here, $\theta_s = \theta(s,:)$ denotes a *vector* of random variable $\theta(s,\check{s})$ that determines the probability of transition from s to $\check{s}$, i.e., $0 \le \theta(s,\check{s}) \le 1$ and $\sum_{\check{s}}\theta(s,\check{s}) = 1$. The likelihood function for an occurred transition $X: s \to s'$ is $p(X|\theta_s) = \theta(s,s') = \Pi_{\check{s}}\, \theta(s,\check{s})^{[\check{s}=s']}$, where [.] denotes the Iverson bracket. Therefore, the posterior belief $q_\gamma(\theta_s)$ obtained by the SMiLe rule [Eq (3.4)],

$$q_\gamma(\theta_s) \propto \left(\Pi_{\check{s}}\, \theta(s,\check{s})^{[\check{s}=s']}\right)^{\gamma} . \left(\Pi_{\check{s}}\, \theta(s,\check{s})^{\alpha(s,\check{s})-1}\right)^{1-\gamma} \propto \Pi_{\check{s}}\, \theta(s,\check{s})^{\beta(s,\check{s})-1}, \qquad (3.23)$$

is again a Dirichlet distribution parametrized by $\beta(s,\check{s}) = (1-\gamma)\alpha(s,\check{s}) + \gamma(1+[\check{s}=s'])$.

The probability $\hat{T}[t](s,s')$ of transition from s to s' at time step t is estimated by $\hat{T}[t](s,s') = \frac{\alpha[t](s,s')-1+\epsilon}{\sum_{\check{s}}(\alpha[t](s,\check{s})-1+\epsilon)}$, where $\alpha[t](s,\check{s})$ denotes the updated model parameter at time step t. Here, $\epsilon > 0$ is a very small number which prevents the denominator from being zero.

The online EM algorithm for the maze-exploration task.

The online EM algorithm, presented in [Mongillo and Deneve, 2008], is an estimation algorithm for the unknown parameters of a hidden Markov model (HMM). For the maze-exploration task we adapted the method presented in [Mongillo and Deneve, 2008] such that the transition probability to a new room also depends on the previously visited room (and not just the current environment). The HMM of the maze-exploration task consists of two sets of unknown parameters: (i) a set $\mathbf{P} = [P_{ij}]_{2\times 2}$ of (unknown) switch probabilities from environment i to j (where we use 1 for environment $\mathscr{A}$ and 2 for environment $\mathscr{B}$), and (ii) a set $\mathbf{T} = [T_{jss'}]_{2\times16\times16}$ of state transition probabilities, where $T_{jss'}$ denotes the probability of transition from state s to state s' within environment j. The set of all unknown parameters is denoted by $\Theta \equiv (\mathbf{P},\mathbf{T})$.

At each time step t, we estimate the probability $q_l^t = P(E_t = l|s_{0\to t})$ of being in environment $E_t = l \in \{1,2\}$, given all previous state transitions $s_{0\to t} = \{s_0, s_1, ..., s_t\}$. The probability q_l^t can be recursively calculated by

$$\hat{q}_l^t = \sum_m \hat{q}_m^{t-1}\gamma_{ml}^t, \tag{3.24}$$

where $\gamma_{ml}^t = \frac{P(s'=s_t|s=s_{t-1},E_t=l)P(E_t=l|E_{t-1}=m)}{P(s'=s_t|s_{0\to(t-1)})}$ belongs to a set of auxiliary variables $\Gamma = [\gamma_{lh}]_{2\times2}$ that are calculated by the last estimate $\hat{\Theta}^{t-1}$ of the model parameters:

$$\gamma_{lh}^t = \frac{\hat{P}_{lh}^{t-1}\hat{T}_{hs_{t-1}s_t}^{t-1}}{\sum_{m,n}\hat{q}_m^{t-1}\hat{P}_{mn}^{t-1}\hat{T}_{ns_{t-1}s_t}^{t-1}}. \tag{3.25}$$

Then, using these auxiliary variables γ_{lh}, a set $\Phi = [\hat{\phi}_{i,j,s,s',h}]_{2\times2\times16\times16\times2}$ of parameters is recursively updated:

$$\hat{\phi}_{i,j,s,s',h}^t = \sum_l \gamma_{lh}^t\left[(1-\eta)\hat{\phi}_{i,j,s,s',l}^{t-1} + \eta\hat{q}_l^{t-1}\Delta_{ijss'}^{lhs_{t-1}s_t}\right], \tag{3.26}$$

where $\Delta_{ijss'}^{lhs_{t-1}s_t} = \delta(i-l)\delta(j-h)\delta(s-s_{t-1})\delta(s'-s_t)$, $\delta(.)$ is the Kronecker delta (i.e., 1 when its argument is zero and 0 otherwise), and η is the learning rate.

Finally, the model parameters are updated by

$$\hat{P}_{ij}^t = \frac{\sum_{s,s',h}\hat{\phi}_{ijss'h}^t}{\sum_{j,s,s',h}\hat{\phi}_{ijss'h}^t}; \quad \hat{T}_{jss'}^t = \frac{\sum_{i,h}\hat{\phi}_{ijss'h}^t}{\sum_{i,s',h}\hat{\phi}_{ijss'h}^t}. \tag{3.27}$$

We emphasize that in order for the online EM algorithm to work properly, some technical considerations must be respected. For instance, in the beginning of learning, only online estimation of Φ must be updated (without updating the model parameters Θ), so that the estimation error for the first 2000 time steps of our simulation (Fig 3.6A, blue) remains fixed. Moreover, we found that the online EM algorithm works well only if it is correctly initialized. To make our comparison fair, we assumed the agent "believes in" frequent transitions between environments by initializing the probabilities $\hat{P}_{ij}^0$ that describe the switch between environment $\mathscr{A}$ and $\mathscr{B}$ to be very close to true ones. Without such an assumption, the online EM takes even more time than what we reported here to learn the maze-exploration task. The actual initialization values were $\hat{P}^0_{12} = \hat{P}^0_{21} = 0.1$ while the true values were $P_{12} = P_{21} = 0.005$.

4 A Biologically Plausible 3-Factor Learning Rule from Gradient Descent Optimization

One of the most frequent problems in both decision making and reinforcement learning (RL) is maximizing an expected quantity involving functionals such as reward or utility. Generally, these problems consist of computing the optimal solution of a density function. Instead of trying to find this exact solution, a common approach is to approximate it iteratively through a learning process.

In this work we propose a functional gradient rule for the maximization of a general form of density-dependent functionals using a stochastic gradient ascent algorithm. If a neural network is used for parametrization of the desired density function, the proposed learning rule can be viewed as a modulated Hebbian rule. Such a learning rule is biologically plausible, because it consists of both local and global factors corresponding to the coactivity of pre/post-synaptic neurons and the effect of neuromodulation, respectively.

We first apply our technique to standard reward maximization in RL and a variational learning problem to show that reward and surprise signals can be interpreted as third factors in this framework. We then use our functional gradient method to derive an online rule for the approximation of the SMiLe rule, introduced in Chapter 3. We implement the aforementioned maze-exploration task in a spiking neural network using our proposed online rule. We show that the proposed online rule is a covariance learning rule, where the strength of the connections between the neurons changes as a function of covariance between the activity of post-synaptic neurons and an estimate of confidence-corrected surprise.

4.1 Introduction

Maximizing an expected quantity, is one of the most frequently encountered problems in both decision making [Janis and Mann, 1977] and reinforcement learning [Sutton and Barto, 1998b]. It usually implies computing the optimal solution of a density function. The density might represent a learning agent's policy in RL, or the likelihood of selecting different choices in a decision making process. We introduce a functional gradient rule for the maximization of a general form of density-dependent functionals, such as reward or utility, using a stochastic gradient ascent algorithm. The resulting learning rule approaches the optimal solution through an iterative process. This learning rule benefits from biological plausibility if a neural network is used for parametrization of the desired density function. As we will see below, it is consistent with a modulated Hebbian learning rule (i.e., 3-factor learning rule) in which both global and local factors influence the synaptic connections among the neurons.

We first apply our technique to standard reward maximization in RL. As expected, this yields the standard policy gradient rule [Baxter et al., 2001, Florian, 2007, Peters and Schaal, 2006, Pfister et al., 2006, Williams, 1992, Xie and Seung, 2004, Sutton et al., 1999], in which parameters of the model are updated proportional to the amount of reward. Next, we use variational free energy as a functional and find that the estimated change in parameters is modulated by a measure of surprise (the subjective Shannon surprise Eq. (2.1)). We then apply our technique to the constraint surprise minimization problem, introduced in Chapter 3, to approximate the SMiLe rule. It yields an online rule, in which the estimated change in the parameters is determined by the covariance of surprise and the activity of the post-synaptic neuron. These examples demonstrate that reward and surprise can both play the role of global third factors in the general framework of three-factor learning rules.

4.2 Results

4.2.1 Functional gradient rule: Theory

We apply stochastic gradient ascent to approximate the optimal density function that maximizes a functional

$$\mathbb{F}[P] = \langle \mathscr{F}[P] \rangle_P \,, \tag{4.1}$$

where $\langle . \rangle_P$ denotes the average with respect to the probability density $P(x)$ of a random variable x. The term $\mathscr{F}[P]$ might be considered as a general form of reward, utility, or surprise function which may itself depend on the density function P. The general form of the online gradient rule is given in the following theorem.

Theorem 1 (functional gradient rule): The stochastic gradient ascent algorithm for maximizing the functional $\mathbb{F}[P]$ in Eq. (4.1) over all possible distributions P parametrized by $\theta \in \mathbb{R}^n$ yields the online learning rule,

$$\Delta\theta \propto \tilde{\mathscr{F}} \nabla_\theta \ln P, \tag{4.2}$$

where the multiplicative factor $\tilde{\mathscr{F}}$ is defined as

$$\tilde{\mathscr{F}} = \frac{\partial}{\partial P}\left(P\mathscr{F}[P]\right) = \mathscr{F}[P] + \frac{\partial \mathscr{F}[P]}{\partial \ln P}, \tag{4.3}$$

evaluated at a sample $x = X^*$.

Proof: In order to have an online learning rule for $\theta \in \mathbb{R}^n$, we first need to calculate the gradient of $\mathscr{F}[P]$ with respect to θ. The exact gradient of the functional in Eq. (4.1) is

$$\begin{aligned}
\nabla_\theta \langle \mathscr{F}[P] \rangle_P &= \int dx \left[P\nabla_\theta \mathscr{F}[P] + \mathscr{F}[P]\nabla_\theta P \right] \\
&= \int dx \left[P\left(\frac{\partial \mathscr{F}[P]}{\partial P} \nabla_\theta P \right) + \mathscr{F}[P]\left(P\nabla_\theta \ln P\right) \right] \\
&= \langle \frac{\partial \mathscr{F}[P]}{\partial P} \nabla_\theta P + \mathscr{F}[P]\nabla_\theta \ln P \rangle_P = \langle \frac{\partial \mathscr{F}[P]}{\partial P}\left(P\nabla_\theta \ln P\right) + \mathscr{F}[P]\nabla_\theta \ln P \rangle_P \\
&= \langle \left(\frac{\partial \mathscr{F}[P]}{\partial P} P + \mathscr{F}[P] \right) \nabla_\theta \ln P \rangle_P = \langle \frac{\partial \left(P\mathscr{F}[P]\right)}{\partial P} \nabla_\theta \ln P \rangle_P ,
\end{aligned} \tag{4.4}$$

where we used the equality $P\nabla_\theta \ln P = \nabla_\theta P$ in the first two lines of derivation above. In gradient ascent algorithm, the amount of change in the model parameter is in proportion to the exact gradient. However it might be difficult or intractable to exactly calculate the gradient. For instance, the exact gradient in Eq. (4.4) is expressed as an average quantity $\langle . \rangle_P$ which may be difficult to be calculated at each time step. Therefore, we may replace the exact gradient with an estimate using one or a few samples drawn from the distribution P. This replacement is the idea behind the stochastic gradient method. For deriving the online learning rule Eq. (4.2), we replace the exact gradient with a point-estimate of that quantity to change the model parameters (at each time step) by a $\Delta\theta$ that fulfills the equation $\langle \Delta\theta \rangle_P = \nabla_\theta \mathbb{F}[P]$.

Corollary 1: The multiplicative factor $\tilde{\mathscr{F}}$ in the learning rule (4.2) can be replaced by

$\tilde{\mathscr{F}} + c$ where c is a constant or a variable that does not depend on x, because

$$\langle\big(\tilde{\mathscr{F}} + c\big)\,\nabla_\theta \ln P\rangle_P = \langle\tilde{\mathscr{F}}\nabla_\theta \ln P\rangle_P + c\,\langle\nabla_\theta \ln P\rangle_P\,, \tag{4.5}$$

and $\langle\nabla_\theta \ln P\rangle_P = \int dx\, P\nabla_\theta \ln P = \int dx\, \nabla_\theta P = \nabla_\theta \int dx\, P = \nabla_\theta(1) = 0$.

Corollary 2: If $\mathscr{F}[P]$ is linear with respect to $\ln P$, then the multiplier factor $\tilde{\mathscr{F}}$ can be replaced by $\mathscr{F}$. The proof is simply done by using Corollary 1 in (4.3).

Corollary 3: The proposed online stochastic gradient rule (Eq. (4.2)) can be transformed to a covariance learning rule $\Delta\theta \propto Cov(\tilde{\mathscr{F}}, \nabla_\theta \ln P)$. This is done by subtracting a constant $c = \langle\tilde{\mathscr{F}}\rangle_P$ from the third factor (according to Corollary 1), i.e.,

$$\langle\Delta\theta\rangle_P = \langle\big(\tilde{\mathscr{F}} - \langle\tilde{\mathscr{F}}\rangle_P\big)\,\nabla_\theta \ln P\rangle_P = \langle\tilde{\mathscr{F}}\nabla_\theta \ln P\rangle_P - \langle\tilde{\mathscr{F}}\rangle_P\,\langle\nabla_\theta \ln P\rangle_P\,. \tag{4.6}$$

Cor. 1 - Cor. 3 are generalizations of the policy-gradient method [Baxter et al., 2001]. We want to stress that our proposed learning rule (4.2) can indeed be embedded in the class of biologically plausible 3-factor learning rules, if a neural network is used for parametrization. Detailed examples will be shown below, but the two generic aspects are: (i) the term $\tilde{\mathscr{F}}$ represents a globally modulating third factor. We note that to evaluate $\tilde{\mathscr{F}}$ we need information from all neurons in the ensemble. Importantly, we need to evaluate $\tilde{\mathscr{F}}$ only once and this information is then used in the update step for all neurons; (ii) the term $\nabla_\theta \ln P$ represents a local Hebbian term, and depends on both pre-synaptic and post-synpatic neural activity as shown now.

The two local factors

As an example, we use a population of spiking neurons for learning the density function P such that the spontaneous activity of the neural population at each time step represents a sample drawn from that distribution. Importantly, we assume that there are no “hidden” neurons so that the spike trains of all neurons are observable and part of the density function $P(x)$. The neuron model that we use here is a generalized linear model (GLM). This model has the form of a Spike Response Model (SRM) with escape noise [Pillow et al., 2008, Jolivet et al., 2006]. The membrane potential $u_i(t)$ of neuron i at time t is given as

$$u_i(t) = \sum_j w_{ij}(X_j * \phi)(t) + \eta_i(t), \tag{4.7}$$

where w_{ij} is the synaptic efficacy between pre-synaptic neuron j and post-synaptic neuron i, $X_j(t) = \sum_f \delta(t - t_j^f)$ denotes the presynaptic spike train, $\phi(t)$ is the somatic

Excitatory Post Synaptic Potential (EPSP), and $\eta_i(t) = -\eta_0 \int_0^t ds\ e^{-\frac{t-s}{\tau_a}}\ X_i(s)$ is the adaptation potential (η_0 and τ_a are constants). The spikes are then generated by a stochastic Point process using an exponential intensity $\rho_i(t)$ [Jolivet et al., 2006] conditioned on the membrane potentials,

$$\rho_i(t) = \rho_0 \exp(\frac{u_i(t) - \theta}{\Delta U}), \tag{4.8}$$

where θ and ΔU are physical constants of the neuron. The synaptic efficacies w_{ij} between neurons are free parameters $\theta \in \mathbb{R}^n$ and parametrize P. A set of spike trains $\{X_i\}$ generated by all the neurons in a time interval T represents a data sample x. Therefore, the relative frequency of the occurrence of a given set of spike trains $x = \{X_i\}$ compared to all other possible sets of spike trains represents the corresponding probability density $P(x)$.

The likelihood of a particular spike train $x = \{X_i\}$ which is observed in the interval $[0, T]$ can be written as [Pfister et al., 2006, Rezende et al., 2011]

$$\ln P(x) = \sum_k \int_0^T dt\ [\ln \rho_k(t) X_k(t) - \rho_k(t)], \tag{4.9}$$

and its gradient with respect to the particular synaptic weight w_{ij} is calculated as (see [Pfister et al., 2006, Rezende et al., 2011, Rezende and Gerstner, 2014] for details)

$$\nabla_{w_{ij}} \ln P(x) = \frac{1}{\Delta U}\ (X_j * \phi)(t)\ [X_i(t) - \rho_i(t)]. \tag{4.10}$$

Therefore, we conclude that an update of synaptic weights w_{ij} according to gradient ascent $\Delta w_{ij} \propto \nabla_{w_{ij}} \ln P(x)$ can be calculated locally and is written as a product of two local (Hebbian) factors: $(X_j * \phi)(t)$ which depends on the firing times of the pre-synaptic neuron j and $[X_i(t) - \rho_i(t)]$ that depends on the state of the post-synaptic neuron i. Note that similar derivations can be performed for simpler neuronal models such as binary neurons without refractoriness [Xie and Seung, 2004].

4.2.2 Functional gradient rule: Applications

The functional online gradient rule of Eq. (4.2) can be applied to a wide range of learning problems in different contexts. In this section, we first review two of the existing learning algorithms (policy gradient methods and variational learning methods) and predict how their corresponding online learning rule in a neural network should look like. We show that our prediction (using the functional gradient rule of

Eq. (4.2)) is consistent with the existing models that have been used for the neural implementation of such learning methods.

We then apply our technique to the constraint surprise minimization problem, from which the SMiLe rule was derived, to provide an appropriate online rule for the neural implementation of the surprise belief update algorithm. In Section 4.3 we will use that online rule in a spiking neural network to simulate the aforementioned maze-exploration task in a neural system.

Policy gradient (review)

Reward maximization in the context of reinforcement learning is formulated as finding an action policy $\pi(a|s)$ that maximizes the expected reward $\langle R(s,a)\rangle_{\pi(a|s)f(s)}$, where $R(s,a)$ denotes the reward for taking action a in state s and $f(s)$ is the density function of state space. Policy gradient methods [Baxter et al., 2001, Florian, 2007, Peters and Schaal, 2006, Pfister et al., 2006, Williams, 1992, Xie and Seung, 2004, Sutton et al., 1999] are well-established iterative algorithms to address the reward maximization problems. They iteratively learn the optimal action policy $\pi(a|s)$ by modifying the model parameters in the direction of the gradient of the expected reward function $\langle R(s,a)\rangle_{\pi(a|s)f(s)}$ with respect to the corresponding parameters.

Policy gradient methods have been applied to spiking neural networks [Pfister et al., 2006, Florian, 2007, Xie and Seung, 2004, Vasilaki et al., 2009]. To keep things simple, we assume that the state s of the environment and the action a that the agent chooses are respectively determined by the spike trains of *place cells* and *action cells* as two separate neuronal populations. The synaptic weights w_{ij} between the place cells j and the action cells i are then used as free parameters that encode the action policy $\pi(a|s)$. Policy gradient method for the network of spiking neurons suggests reward maximization (also known as R-max [Frémaux et al., 2010]) learning rule that is generally expressed as (see [Pfister et al., 2006, Vasilaki et al., 2009] for the details of derivation)

$$\begin{aligned} \frac{de_{ij}}{dt} &= -\frac{e_{ij}}{\tau} + (X_j * \phi)(t)\ [X_i(t) - \rho_i(t)], & (4.11) \\ \frac{dw_{ij}}{dt} &= \eta(R(t) - b)e_{ij}(t). & (4.12) \end{aligned}$$

R-max learning rule in Eqs. (4.11),(4.12) is an example of three-factor learning rule. The *eligibility trace* e_{ij} [Sutton and Barto, 1998b] defined as a low-pass filter of co-activity between neurons j and i combines the two local Hebbian factors. The *success signal*

$R(t)-b$ modulates the direction and speed of weight update. The constant baseline b in the success signal $R(t)-b$ is often replaced by the average reward $\bar{R}$ [Frémaux et al., 2010].

We emphasize again that the amount of change in w_{ij} (Eq. (4.12)) does not only depend on the local Hebbian factors (i.e., the eligibility trace e_{ij}), but it also depends on the reward $R(t)$ delivered at time t. This is consistent with the functional gradient rule (**Theorem 1**) in combination with Eq. (4.10), which predict that the change in the model parameters should have the following form:

$$\Delta w_{ij} \propto (R \pm c)\, \nabla_{w_{ij}} \ln \pi \overset{(4.10)}{=} (R \pm c)\, (X_j * \phi)(t)\, [X_i(t) - \rho_i(t)]. \tag{4.13}$$

Note that for maximizing the expected reward $\langle R(s,a)\rangle_{\pi(a|s)f(s)}$, the third factor $\tilde{\mathscr{F}}$ [Eq. (4.3)] may be equal to $R := R(s,a)$ or $R \pm c$ because the reward does not explicitly depend on the policy π (and the model parameters w_{ij}) and so $\tilde{\mathscr{F}} = \mathscr{F}$ (according to **Corollary 2**) or $\tilde{\mathscr{F}} = \mathscr{F} \pm c$ (according to **Corollary 1**). The shift by an arbitrary amount c is a well-known result for policy gradient rules [Baxter et al., 2001, Sutton et al., 1999]

Variational learning (review)

Variational methods are typically used in complex statistical models which are defined by a joint distribution $p(v,h)$ over a set of observed (visible) variables v and latent (hidden) variables h. The joint distribution p can be interpreted as a generative model for the input statistics $p(v)$ governed by some adaptive parameters $\theta \in \mathbb{R}^n$. Variational methods are used in machine learning to approach two important aims: first, to analytically approximate the posterior distribution $p(h|v)$ of hidden variables (for statistical inference over them); second to derive a lower bound for a marginal likelihood $p(v) = \sum_h p(v,h)$ of the visible variables (usually for model selection). We focus on this second aim. A computationally tractable lower bound $\mathscr{L}(q; w, \theta)$ for the marginal likelihood $p(v)$ of the visible variables is calculated by using an auxiliary distribution $q(h|v)$:

$$\begin{aligned} \ln p(v) &= \ln \sum_h p(v,h) = \ln \sum_h q(h|v)\, \frac{p(v,h)}{q(h|v)} \\ &\geq \sum_h q(h|v)\, \ln \frac{p(v,h)}{q(h|v)} := \mathscr{L}(q; w, \theta), \end{aligned} \tag{4.14}$$

where we have applied Jensen's inequality. Here $w \in \mathbb{R}^m$ denotes adaptive parameters used for expressing $q(h|v)$. The difference between the true log likelihood $\ln p(v)$ and its approximated lower bound $\mathscr{L}(q; w,\theta)$ is

$$\ln p(v) - \mathscr{L}(q; w,\theta) = \sum_h q(h|v) \ln \frac{q(h|v)}{p(h|v)} := D_{KL}(q||p). \tag{4.15}$$

Therefore, maximizing the lower bound $\mathscr{L}(q; w,\theta)$ with respect to w is equivalent to minimizing the Kullback-Leibler divergence $D_{KL}(q||p)$ between the true posterior distribution $p(h|v)$ and the approximated one $q(h|v)$. The lower bound $\mathscr{L}(q; w,\theta)$ is known as (negative) variational free energy $\mathbb{F}[q; v]$ in statistical learning [MacKay, 2003] and can be expressed as

$$\mathbb{F}[q; v] = -\mathscr{L}(q; w,\theta) = \langle -\ln p(v,h)\rangle_q - H(q) = \langle -\ln p(v,h) + \ln q(h|v)\rangle_q. \tag{4.16}$$

The variational free energy $\mathbb{F}[q; v]$, for a given observed sample v, is an estimate of the Shannon surprise $S_{Sh}(v) = -\ln \int_h p(v,h)dh = -\ln p(v)$ [Eq. (2.1)] [Friston, 2010]. Therefore, it indicates how surprising a new observed sample v is perceived for a subject whose internal model of the external world is modeled by a generative model $p(v,h)$ and an auxiliary distribution $q(h|v)$.

Let us now relate the above results to our **Theorem 1** [Eqs. (4.2), (4.3)]. We can express the variational free energy $\mathbb{F}[q; v]$ as an expected quantity $\langle \mathscr{F}[q; v]\rangle_q$, where $\mathscr{F}[q; v] = -\ln p(v,h) + \ln q(h|v)$. We introduce weights w which parametrize the distribution $q(h|v)$. Therefore we can apply our technique for approximating the optimal solution. The online learning rule, suggested by **Theorem 1**, for variational free energy minimization is then given by

$$\Delta w \propto -(\mathscr{F} \pm c)\nabla_w \ln q, \tag{4.17}$$

where the minus sign arises because of the minimization. Here $\mathscr{F} = \mathscr{F}[q; v]$ is the point-estimate of free energy for the observed sample v: note that the modulation factor $\mathscr{F}$ in Eq. (4.17) is $-\ln p(v,h) + \ln q(h|v)$ evaluated at a randomly sampled h from $q(h|v)$. According to **Corollary 2**, the multiplicative factor $\tilde{\mathscr{F}}$ in Eq. (4.2) is equal to $\mathscr{F}[q; v]$ since $\mathscr{F}[q; v]$ is linear in $\ln q$.

The learning rule Eq. (4.17) suggests that the amount of change in model parameters w is proportional to an estimate of the Shannon surprise $S_{Sh}(v)$. In other words, the surprise signal measured as the instantaneous free energy (an estimate of the Shannon surprise) modulates the learning rate such that more surprising samples v result in a larger change in model parameters. A practical example of this technique has been reported in [Rezende and Gerstner, 2014], where the same quantity as in Eq. (4.10) is used for the Hebbian term and the third factor is considered to be a novelty-related signal $e(t) = \mathscr{F} - \overline{\mathscr{F}}$, consistent with what we suggested in Eq. (4.17).

Surprise minimization

In Chapter 3, we derived the posterior belief $q_\gamma(\theta)$ from the SMiLe rule [Eq. (3.4)] as a solution to the constraint surprise minimization problem in Eq. (3.3). As before, given a data sample X, we define the objective function to be the confidence-corrected surprise $S_{corr}(X;q) = D_{KL}[q(\theta)||\hat{p}_X(\theta)]$ of the data sample X, under a parametrized belief $q(\theta)$ about the latent variables θ. Before observing X, we have $q(\theta) = \pi_0(\theta)$, where $\pi_0(\theta)$ is the current belief about the model parameters, built from previous samples. If we minimize the above objective function under the constraint $D_{KL}[q(\theta)||\pi_0(\theta)] \leq B$, we will have $q(\theta) = q_\gamma(\theta)$ [Eq. (3.4)].

Instead of solving for the exact solution, we can approximate it iteratively using stochastic gradient descent on the following Lagrangian:

$$\begin{aligned}\mathbb{L}[q] &= D_{KL}[q(\theta)||\hat{p}_X(\theta)] + \frac{1}{\lambda}\big(D_{KL}[q(\theta)||\pi_0(\theta)] - B\big) - \lambda'\left(\int_\theta q(\theta)d\theta - 1\right) \\ &= \Big\langle \ln q(\theta) - \ln \hat{p}_X(\theta) + \frac{1}{\lambda}\ln\frac{q(\theta)}{\pi_0(\theta)} - \lambda' \Big\rangle_q - \frac{B}{\lambda} + \lambda', \qquad (4.18)\end{aligned}$$

where $\frac{1}{\lambda}$ is the Lagrange multiplier for the constraint on bound B, and λ' is the Lagrange multiplier for constraint on $q(\theta)$ to be a probability density function that integrates to 1. We expressed the first Lagrange multiplier as $\frac{1}{\lambda}$ (and not λ) just to be consistent with our notation in Chapter 3 (see Eq. (3.9)). We emphasize that the Lagrangian Eq. (4.18) is expressed for a single data sample X, and not an average (or summation) over a set of data samples.

We now apply **Theorem 1** and find an update Δw for the parameters w that control the current belief:

$$\Delta w = -\eta(\mathscr{F}[q] \pm c)\nabla_w \ln q. \qquad (4.19)$$

The third factor $\mathscr{F}[q]$ in Eq. (4.19) is equal to

$$\mathscr{F}[q] = \ln q(\theta) - \ln \hat{p}_X(\theta) + \frac{1}{\lambda}\ln\frac{q(\theta)}{\pi_0(\theta)}. \qquad (4.20)$$

The third factor $\mathscr{F}[q]$ in Eq. (4.20) consists of three terms. The first two terms (i.e., $\ln q(\theta) - \ln \hat{p}_X(\theta)$) quantify a point-estimate of the confidence-corrected surprise (using a sample θ from the latent variable space), and the third term $\frac{1}{\lambda} \ln \frac{q(\theta)}{\pi_0(\theta)}$ controls the stickiness of the model parameters to its previous values. The degree of stickiness is determined by the Lagrange multiplier $\frac{1}{\lambda}$.

Note that in Eqs. (4.19) and (4.20), $q(\theta)$ is *not* the updated belief, which is obtained only *after* updating w using Eq. (4.19). We can think of $q(\theta)$ as the *current approximate* of the posterior belief after the data sample X is observed. Therefore, one could also replace $q(\theta)$ with $\pi_0(\theta)$ (as the initial guess for the posterior belief) and update w using the online rule Eq. (4.19). Such as assumption simplifies all the above derivations.

Note that in all expressions above, we can replace the scaled likelihood $\hat{p}_X(\theta)$ with the likelihood $p(X|\theta) = \hat{p}_X(\theta) \left(\int_\theta' p(X|\theta') d\theta'\right)$, because they differ only in a multiplicative factor $\int_\theta' p(X|\theta') d\theta'$ which does not depend on θ and is absorbed during normalization. We emphasize again that since the Lagrangian Eq. (4.18) is expressed for a single data sample X, the scale factor $\int_\theta' p(X|\theta') d\theta'$ is a constant. We can further add other constant terms to the online rule to ensure that it works in practice.

4.3 Neural implementation of the SMiLe rule

4.3.1 Neural network model

We propose a neural network model that can be used for the neural implementation of the maze-exploration task, introduced in Chapter 3. Our model consists of a *recognition* network and a *prediction* network (see Fig. 4.1).

Recognition network

The recognition network is a two-layer feed-forward spiking network whose aim is to correctly recognize the room from which environmental inputs are received. The input layer consists of 784 neurons whose activities $y_j,\ j \in \{1, \ldots, 784\}$ stand for the neural representation of the environmental inputs (sensory cues). The output layer consists of 16 excitatory neurons corresponding to 16 available rooms. Given an input vector $\vec{y}$ (a sensory stimulus from the current room), only one of the output neurons $k \in \{1, \ldots, 16\}$ is allowed to be active in a time step. We imagine this to be neurally implemented by a winner-take-all (WTA) framework. This neuron determines the most likely room that the agent visits at that time. Such WTA assumption provides computational benefits for the neural implementation of Bayesian modeling and simplifies analysis of such neural networks [Nessler et al., 2013, Kappel et al., 2014].

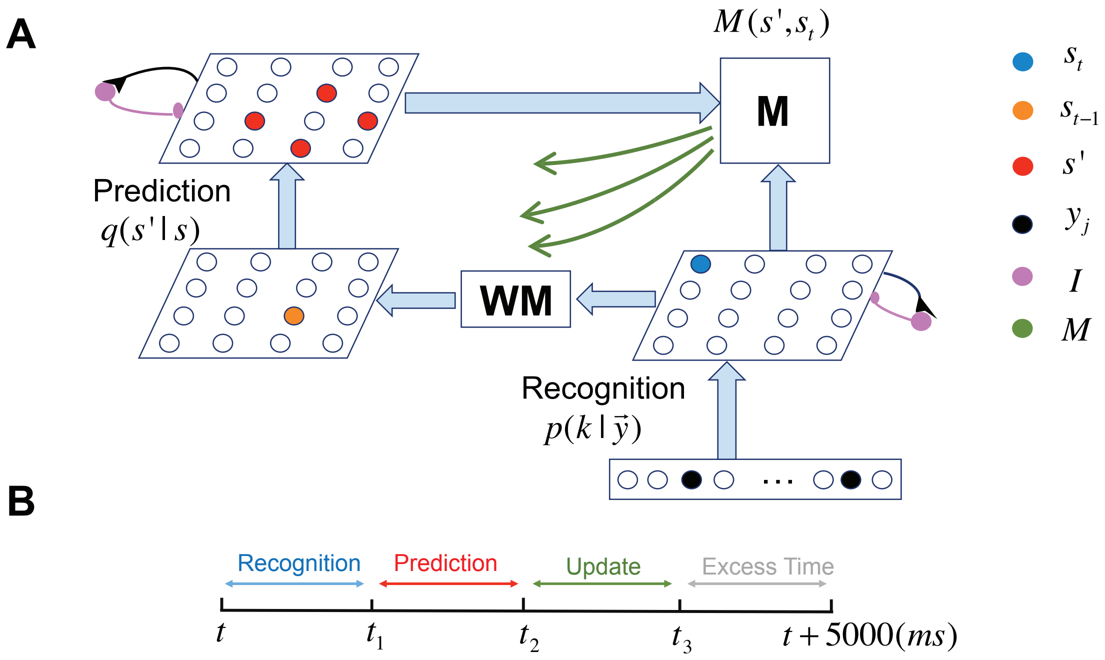

Figure 4.1 – **Neural network model for the maze exploration task. A.** The model consists of a recognition and a prediction network as well as a sub-network M evaluating surprise. The neural representation of the current state s_t is given as binary input vectors $\vec{y}$. One of the neurons in the output layer of the recognition network responds and indicates to which state that input belongs. The previously visited state is recalled by working memory (WM) which clamps one neuron (orange) in the first layer of the prediction network to the active state. The agent's model about state transition probabilities $q(s'|s)$ is encoded in synaptic strengths of the prediction network. Multiple samples (red) are selected via a soft-max spiking probability rule and communicate information to a modulatory sub-network M (green). Once surprise is calculated by M, the global signal is propagated through the network and affects plasticity in the prediction model. In both recognition and prediction networks a WTA framework is employed using an abstract inhibitory neuron (magenta). **B.** Time schedule of network operation during the time spent in a state (lasting for 5 seconds). Once the agent enters a state, the recognition phase starts. Then it starts predicting the next states until the surprise is calculated (at the end of prediction phase). The recognition and the prediction phases could also be in parallel. By release of the global modulatory surprise signal, synaptic strengths change during an update phase. Excess time remaining before entering the next state may be used by the agent for consolidation or advance prediction of the next visited state.

We consider a simplified Spike Response Model for the neurons in the output layer of the recognition model, where the instantaneous firing rate $v_k = exp(u_k)$ of each neuron k is exponentially linked to its membrane potential $u_k = E_k - I$. Here, $E_k = \sum_j w_{kj} y_j$ denotes the total excitatory input received by neuron k and I is the common inhibitory input to all neurons $k \in \{1, ..., 16\}$ in the output layer. Following [Nessler et al., 2013], inhibitory neurons are not modeled explicitly but calculated algorithmically as

$$I = \ln \sum_k e^{E_k}. \tag{4.21}$$

Such an assumption results in soft-max spiking probability rule $v_k = \frac{e^{E_k}}{\sum_k e^{E_k}}$ and normalization of the firing rates, i.e., $\sum_k v_k = 1$ (see [Nessler et al., 2013, Kappel et al., 2014]). The aim of learning in the recognition model is to modify the synaptic strengths between the neurons such that the agent correctly recognizes the room from which environmental inputs are received. We use the following online Hebbian rule (derived from Likelihood maximization) to learn the recognition model (see **Materials and Methods** for derivation):

$$\Delta w_{kj} = \eta \delta_{kk^*} (y_j - e^{w_{kj}}). \tag{4.22}$$

Here y_j is the state of the presynaptic neuron, δ_{kk*} denotes Kronecker delta function and k^* indicates the index of winner neuron in the WTA network of the output layer of the recognition network. The online rule Eq. (4.22) has a simple interpretation. If the post-synaptic neuron k is inactive, the strength of none of its afferent synapses change. If the post-synaptic and pre-synaptic neurons are both active, then LTP occurs. Otherwise (i.e., if only post-synaptic neuron is active) LTD occurs. The term $e^{w_{ij}}$ in Eq. (4.22) stands for heterosynaptic plasticity that is naturally derived as a result of normalization. The online rule Eq. (4.22) enables our recognition network to correctly learn the association between the environmental inputs and their corresponding sources that generated them (see Fig. 4.2). The rule Eq. (4.22) is identical to that of Nessler et al. [Nessler et al., 2013], but our derivation is more direct (see **Materials and Methods**). After learning, a neuron with index k codes for a state (or room) s and another neuron $\tilde{k}$ for a different state $\tilde{s}$. The neuron k in the recognition network is linked (via a working memory) to a neuron with index s in the prediction network.

Prediction network

The prediction network is also a two-layer feed-forward network but with 16 neurons in each layer. The activity of the input layer of this network is driven by working memory (see Fig. 4.1). Working memory recalls the last visited room and enforces neuron s to be the only active neuron at time t in the first layer of the prediction network, if room k was visited at time $t-1$. The activity of the output layer is then driven by neuron s. A neuron with index s' in the output layer of the prediction network indicates a state that is predicted to be visited in the next time step. The agent's belief

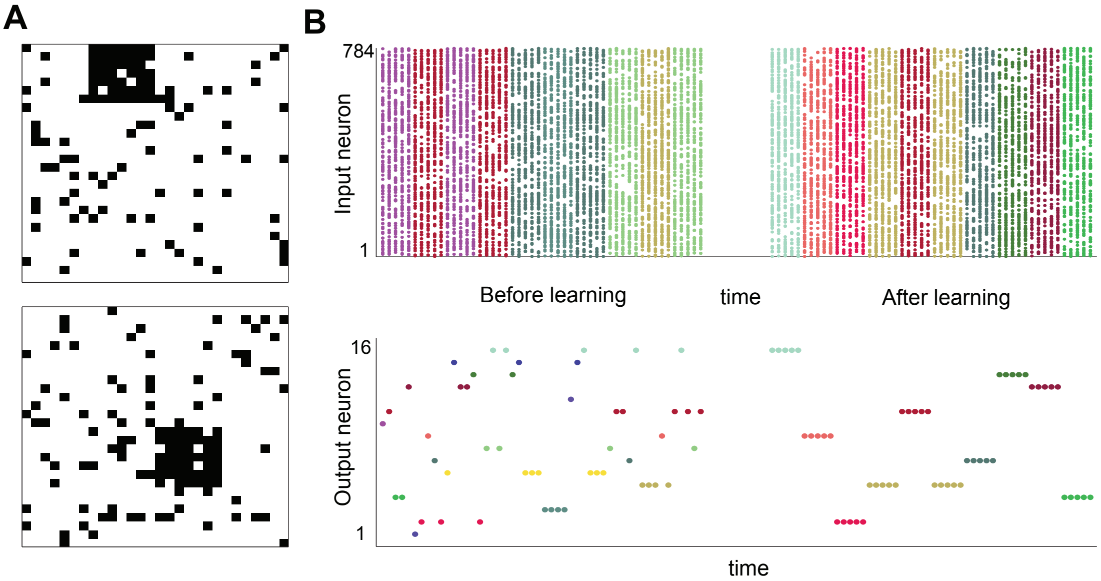

Figure 4.2 – **Inputs and outputs of the recognition network before and after learning. A.** An example of noisy environmental inputs the agent receives from state 2 (top) and state 11 (bottom). The majority ($= \frac{15}{16} \times 784$) of entries of an input vector $\vec{y}$ (visualized as 28×28 square) are independently and identically drawn from a binomial distribution with probability of generating a black pixel equal to 0.1. The remaining entries correspond to the spatial position of the current state and have a black pixel with probability 0.9. The indices of the pixels are randomly shuffled in input vector $\vec{y}$. **B.** The raster plot (spiking activity) of the input and output neurons in the recognition model before and after learning. After the network learns the hidden statistical structure of the input data set, a unique neuron is assigned to the class of input vectors that are generated by a same source corresponding to a particular state ("room"). After learning, the network correctly recognizes the state from which an input vector is presented to the network.

about the state transition probabilities is modeled by synaptic weights $w_{s's}$ between the two layers. Note that our prediction network also uses a WTA framework to ensure that at each time step, only one neuron in each layer remains active. The WTA of the prediction network is implemented analogously to that of recognition network.

In the prediction network, we use the same neuronal model as in the recognition network, i.e., a simplified Spike Response Model, where the instantaneous firing rate $v_{s'} = exp(u_{s'})$ of each neuron s' is exponentially linked to its membrane potential $u_{s'} = E_{s'} - I$, with $E_{s'} = \sum_s w_{s's}\delta_{ss_{t-1}} = w_{s's_{t-1}}$ and $I = \ln \sum_{s'} e^{E_{s'}} = \ln \sum_{s'} e^{w_{s's_{t-1}}}$. Note that in above equations, we assumed that the synaptic weights in the recognition network have converged to their stationary values, and the recognition network is now capable of correctly recognizing the current state, given environmental inputs. Therefore, the state that is recalled from working memory at time t indeed corresponds to previously visited state s_{t-1}. The same assumption will be used in following derivations.

We introduce a surprise-related modulatory signal $M(s', s_t)$ expressed as

$$M(s', s_t) = w_{s' s_{t-1}} - \ln\,(1 + \delta_{s' s_t}) + \frac{1}{mS(X)}(w_{s' s_{t-1}} - w^{old}_{s' s_{t-1}}), \tag{4.23}$$

where s' denotes a *predicted* next state (using the prediction network with parameters $w_{s's}$), given the last visited state was s_{t-1}. The probability of predicting s' as the next state is determined by the current model $q_{t-1}(s')$, and is equal to $e^{w_{s' s_{t-1}}}$. The modulatory signal $M(s', s_t)$ in Eq. (4.23) depends on whether the predicted state s' is the same as the *real* visited state s_t via $\delta_{s' s_t}$. Therefore, it is linked to a notion of *prediction error signal.*

The first two terms (i.e., $w_{s' s_{t-1}} - \ln\,(1 + \delta_{s' s_t})$) in the modulatory expression $M(s', s_t)$ [Eq. (4.23)], in fact, quantifies a point-estimate of the confidence-corrected surprise (i.e., $\langle \ln q(\theta) - \ln \hat{p}_X(\theta) \rangle_{q(\theta)}$ in Eq. (4.20), where q is expressed in terms of synaptic weights $w_{ss'}$). The third term $\frac{1}{mS(X)}(w_{s' s_{t-1}} - w^{old}_{s' s_{t-1}})$ controls the stickiness of the model parameters at their previous values. The degree of stickiness is determined by the surprise $S(X)$ of the most recent state transition $X : s_{t-1} \to s_t$, such that if X is more surprising than X', then the updated belief moves more towards $\hat{p}_X(\theta)$ after observing X than after observing X'.

We emphasize that $S(X)$ stands for "prior" surprise (i.e., the surprise of data sample X before the model is updated). In derivation of the SMiLe rule in Chapter 3, however, we minimized "posterior" surprise (i.e., the surprise of data sample X after belief update) with a constraint that was linked to the "prior" surprise $S(X)$. Therefore, we can ignore any dependency between $S(X)$ and current model parameters $w_{s's}$ (i.e., $\frac{\partial S(X)}{\partial w_{s's}} = 0$).

The plasticity rule that we use in the prediction model is derived by applying stochastic gradient descent to the Lagrangian Eq. (4.18). We emphasize that the functional Eq. (4.18) is expressed for a "single" data sample X, and not an average over a set of data samples. The online rule (see **Materials and Methods** for derivation) is analogous to Eq. (4.19):

$$\Delta w_{s''s} = \eta \delta_{s s_{t-1}} \Big(\langle M(s', s_t) \delta_{s' s''} \rangle_{q(s')} - \langle M(s', s_t) \rangle_{q(s')} \langle \delta_{s' s''} \rangle_{q(s')} \Big). \tag{4.24}$$

The online rule Eq. (4.24) is a *covariance* learning rule (see **Corollary 3**) in which the strength $w_{s''s}$ of connection between the pre-synaptic neuron $s \in \{1, \ldots, 16\}$ and the post-synaptic neuron $s'' \in \{1, \ldots, 16\}$ changes as a function of covariance between the activity of the post-synaptic neuron (i.e., $\delta_{s's''}$) and the surprise-related modulatory signal $M(s', s_t)$, where covariance is approximated by averaging over multiple predicted states s'. Note that the online rule Eq. (4.24) also depends on the activity of the presynaptic neuron s via $\delta_{ss_{t-1}}$, indicating that only the efferent synapses of neuron $s = s_{t-1}$ that is activated by working memory are under the influence of the online rule Eq. (4.24) at time t.

Surprise-related sub-network

The online rule Eq. (4.24) requires a covariance term to be approximated. Therefore modules in our neural network need to calculate three separate quantities: (1) a correlation term $\langle M(s', s_t)\delta_{s's''}\rangle_q$, (2) an estimate of the firing rate $\langle \delta_{s's''}\rangle_q$ of each neuron s'', and (3) the average modulatory signal $\langle M(s', s_t)\rangle_q$. While the first two quantities are neuron-specific parameters, the third one does not depend on the post-synaptic neuron s''. In what follows we explain how these three quantities can be implemented by artificial neurons.

When neuron $s = s_{t-1}$ in the first layer of the prediction network is clamped to be active, multiple samples s' are drawn from the second layer using the current model $q(s')$. Whenever a neuron is active, it sends a signal to the modulatory sub-network M (see Fig.4.1). A signal from M is fed back to the prediction network, but it only affects the presently active neuron s'. Note that at each time step, only one of the neurons in the second layer becomes active (because of WTA). Once a neuron s' fires, three neuron-specific traces $e_1(s'), e_2(s')$, and $e_3(s')$ as well as a modulatory trace $e(M)$ will be updated according to the following rules (see also Fig. 4.3)

$$\dot{e}_1(s') = -\frac{e_1(s')}{\tau_b} + \sum_{t_f(s')\in T_p} \delta(t - t_f(s')) \tag{4.25}$$

$$\dot{e}_2(s') = -\frac{e_2(s')}{\tau_s} + \sum_{t_f(s')\in T_p} \delta(t - t_f(s')) \tag{4.26}$$

$$\dot{e}_3(s') = -\frac{e_3(s')}{\tau_b} + e_2(s')M(s', s_t) \sum_{t_f(s')\in T_p} \delta(t - t_f(s') - dt) \tag{4.27}$$

$$\dot{e}(M) = -\frac{e(M)}{\tau_b} + \sum_{s'} e_2(s')M(s', s_t) \sum_{t_f(s')\in T_p} \delta(t - t_f(s')). \tag{4.28}$$

The trace $e_1(s')$ in Eq. (4.25) approximates the firing rate of neuron s' during the prediction phase (in a time interval T_p), with a long time constant τ_b. On a short time scale τ_s, the trace $e_2(s')$ in Eq. (4.26) tags the currently active neuron to ensure that the "global" feedback $M(s', s_t)$ only affects the tagged neuron s'. We assume that $e_2(s')$ has a very small time constant τ_s and decays back to zero before the next sample appears. The third trace $e_3(s')$ in Eq. (4.27) estimates the correlation term $\langle M(\tilde{s}, s_t)\delta_{\tilde{s}s'}\rangle_{q(\tilde{s})}$ for neuron s' using multiple samples $\tilde{s}$. The time delay dt in Eq. (4.27) is because of the delay in receiving the feedback from the modulatory sub-network M. Note that the first and the the third traces $e_1(s')$ and $e_3(s')$ have big time constant $\tau_b >> \tau_s$ and act as integrators. Finally, the trace $e(M)$ in Eq. (4.28) integrates $M(s', s_t)$ over all samples s'. We emphasize that during prediction phase, the subnetwork M fires whenever *any* neuron s' in the second layer fires. At the end of the prediction phase, all the local traces $e_1(s'), e_2(s'), e_3(s')$ [Eqs. (4.25)-(4.27)], as well as the global trace $e(M)$ [Eq. (4.28)] will be used for updating the model parameters, i.e., the synaptic strengths. Upon presentation of the next state, all traces will reset to zero.

The online rule Eq. (4.24), therefore, can be transformed to

$$\Delta w_{s''s} = \eta\delta_{ss_{t-1}}\left(e_3(s'') - e(M)e_1(s'')\right). \tag{4.29}$$

Three-cycle regime

We assume that the time spent in each room (i.e., the time between visiting s_{t-1} and s_t) is 5 seconds. This time will be further divided into 4 smaller time duration as follows (see Fig. 4.1). Once the agent enters a room, it takes some time to *recognize* what the state s_t is (the recognition phase). Then the last visited state s_{t-1} is recalled from working memory, and the agent starts thinking about those states that were most likely to be visited at time t. This phase is called prediction phase during which neuron s_{t-1} is clamped to be active in the first layer of the prediction network and multiple sample states s' are drawn from the second layer according to the soft-max spiking probability rule. Neurons s' leave traces when they fire. Surprise of the observed state transition $s_{t-1} \rightarrow s_t$ is then calculated by a modulatory trace $e(M)$ in the sub-network M. Note that there is a non-plastic 1-to-1 connection from any neuron s' to modulatory sub-neuron M, such that M fires whenever a neuron in the second layer of the prediction neuron fires.

Once surprise is calculated (at the end of the prediction phase), the synaptic strengths will be ready to change. The strength $w_{s''s}$ of each synaptic connection in the prediction network is then modified using both local factors (presynaptic activity $\delta_{ss_{t-1}}$, and post-synaptic traces $e_1(s''), e_3(s'')$) and global factors (modulatory trace $e(M)$ and learning rate η) according to Eq. (4.29). All synaptic changes occur during the update

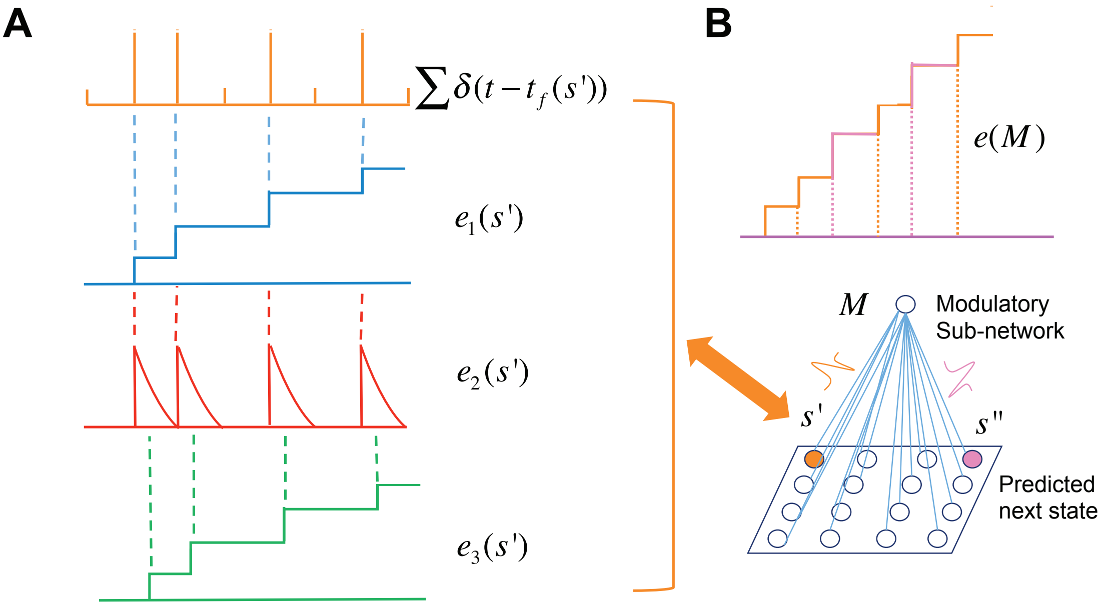

Figure 4.3 – **Local and global traces required for plasticity. A.** Each one of the states s' is equipped with three local traces. The first trace (blue) estimates the firing rate of the neuron by integrating over the spike train (orange) of the neuron. The second trace has a very fast decay and is used to ensure that at the time of feedback from the modulatory sub-network M, the only neuron that is affected by that feedback is the one that is recently fired. The third trace approximates the correlation term between the activity of the post-synaptic neuron and the modulatory signal $M(s', s_t)$. **B.** The modulatory sub-network M fires whenever one of the neurons in the second layer of the prediction network (e.g., s' or s'') fires. The trace $e(M)$ increases by $M(s', s_t)$ whenever neuron s' fires and it increases by $M(s'', s_t)$ whenever neuron s'' fires. Therefore, the change size at each time step is different and depends on the neuron that has just fired in the prediction network.

phase. After these three phases (i.e., recognition, prediction, and update phases) the agent may still have some excess time that can be used for the consolidation of the recent updates or for developing a prior expectation of what it may visit as next time. But these ideas have not been implemented. Note that whether the network is in the phase of recognition, prediction, or update can be determined by a background signal that controls the timing of regimes in which the network operates (like a three-cycle clock).

4.3.2 Simulation results

The setting of the maze exploration we neurally simulate is exactly the same as in Chapter 3. We use the exact parameters as before to generate the same sequence of states that had been visited by the agent in our previous simulation. However, in the neural network implementation of this task, the agent is not aware of the exact state, and it has to recognize the true state using the recognition network.

As described earlier, each time step between states s_{t-1} and s_t is divided into recognition, prediction, and update phases. Once the agent enters state s_t and during recognition phase, 50 input vectors $\vec{y} \in \mathbb{R}^{784}$ are randomly generated and given to the recognition network (see Fig. 4.2). These vectors correspond to the neural representation of the sensory cues within state s_t. In the early phase of learning, the recognition network does not perform well. However, as time goes by, the network becomes capable of recognizing the true state (the current room) by assigning a unique neuron k to the set of all input vectors that are generated by the same source (see Fig. 4.2).

During the prediction phase, the previous state is first recalled from working memory and its corresponding neuron is clamped to be the only active neuron in the first layer of the prediction model. Then we sample 10 times from the prediction network, where at each time one of the neurons $s' \in \{1,...,16\}$ can be selected (according to the softmax spiking probability, discussed earlier). Fig. 4.3 depicts a schematic representation of traces obtained at the end of prediction phase.

The estimation error in the model parameters is depicted in Fig. 4.4 indicating that the neural network quickly adapts to the parameters of the new environment once a switch occurs. Our estimation error graphs are similar to Fig. 3.4 in Chapter 3. The surprise-related modulatory signal $M(s', s_t)$ averaged over multiple samples using current model $q(s')$ responds to unexpected switch points (see Fig. 4.4).

To analyze the sensitivity of network performance to the number of samples at each state, we simulated the network with two different settings. In the first setting, we reduced the number of samples in the recognition network to 1. This increases the time required for the recognition network to learn the state (room) given the one sensory input vector $\vec{y}$. As such the network does not perform well (see Fig. 4.5A). In the second setting, we reduced the number of samples in the prediction network to 1. Here the recognition network quickly learns the statistical structure of the feature space, but using online rule Eq. (4.24) the prediction network cannot learn at all, because when only one s^* is sampled as the predicted next state (in the second layer of the prediction network), then $\Delta w_{s's} = 0$ for all s' in the second layer (Fig. 4.5B). To resolve this issue (when there is only one sample s^* in the prediction network), we suggest to replace the online rule Eq. (4.24) with a learning rule of the form $pre.\overline{post}.(M - \overline{M})$, where $\overline{post} = \langle \delta_{s's''} \rangle_{q(s')} = e^{w_{s''s}}$ is explicitly linked to the synaptic strengths $w_{s''s}$. Fig. 4.5C shows that using this online rule, the network now starts learning but still it does not perform very well. Therefore, a sufficient number of samples in both recognition

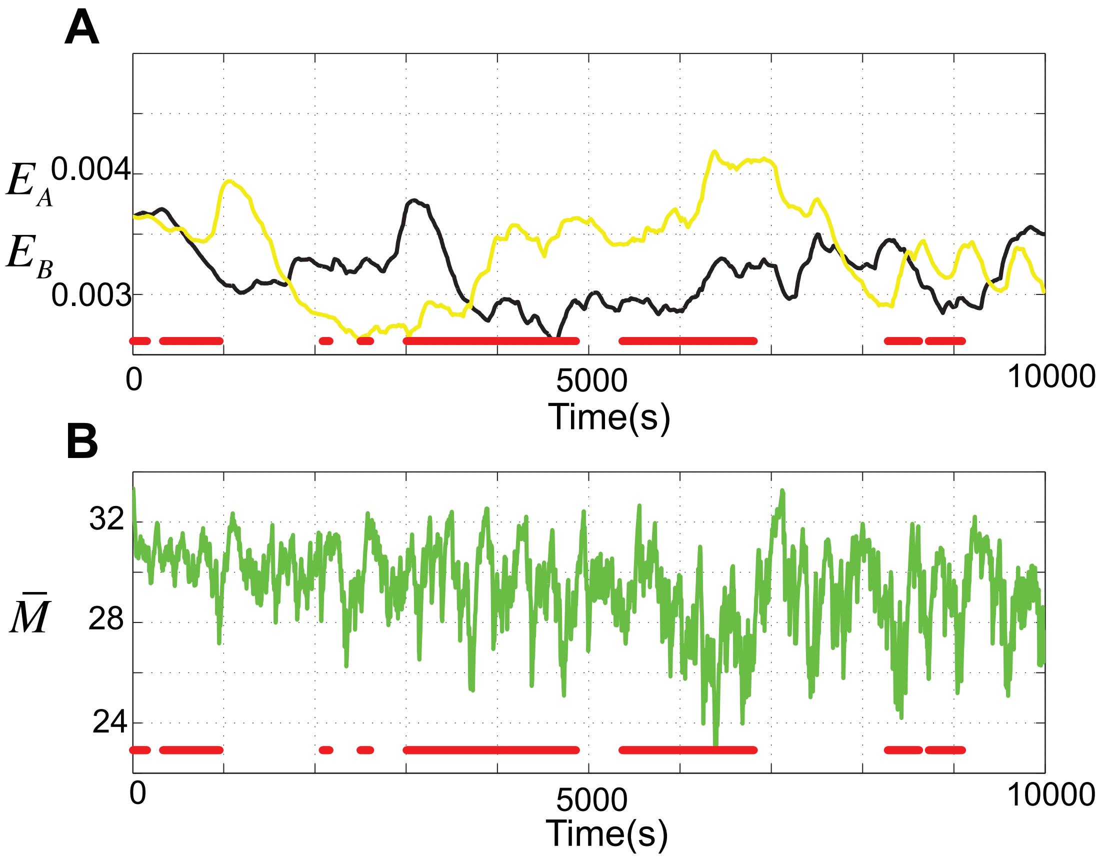

Figure 4.4 – **The estimation error and the modulatory signal in spiking network during maze-exploration task A.** The normalized estimation error $E_{\mathscr{A}}$ and $E_{\mathscr{B}}$ when the estimated probabilities are compared with the true parameters in environment $\mathscr{A}$ (black) and in environment $\mathscr{B}$ (yellow). The red bars indicate the time steps the agent stays in environment $\mathscr{A}$. All the setting parameters are chosen as before (see details of the maze-exploration task in Chapter 3). During 5 seconds of staying in a given state (room), the number of samples for the recognition and the prediction networks are selected as 50, and 10, respectively. The estimation error $E_{\mathscr{A}}$ during exploration of environment $\mathscr{A}$ decreases and then it starts increasing when the agent enters the environment $\mathscr{B}$ **B.** The average modulatory signal $\overline{M} = \langle M(s', s_t) \rangle_{q(s')}$ [Eq. (4.23)] (green) which depends on the point-estimate of surprise (but, is not exactly the same as surprise) increases at switch points.

network and prediction network is necessary for the neural network to work properly.

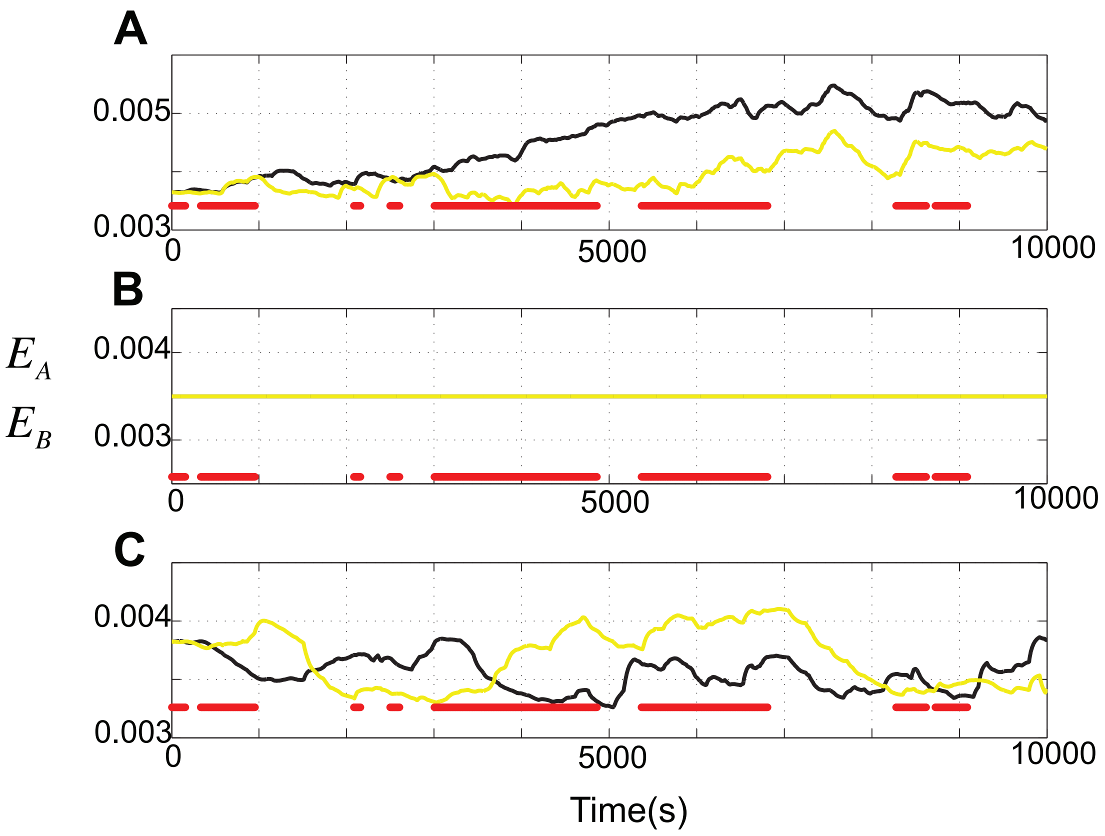

Figure 4.5 – **The effect of number of samples in the recognition network and the prediction network on the learning performance** The estimation errors $E_{\mathscr{A}}$ (black) and $E_{\mathscr{B}}$ (yellow) during maze exploration are depicted. **A.** When the number of samples in the recognition network is set to 1, the network requires much more time than before to correctly recognize the visited state given one environmental input. Therefore, no learning happens in the early phase of simulation. **B.** If the number of samples for the prediction network is reduced to 1, then Eq. (4.24) implies no change in all synaptic strengths. This is because for (the "only") sampled state s^* in the prediction network, $\Delta w_{s''s} \propto M(s^*, s_t) - < M(s', s_t) >= 0$ and for all other states $s'' \neq s^*$ we have $\delta_{s's''} = 0$ (see Eq. (4.24)). The estimation errors $E_{\mathscr{A}}$ (black) and $E_{\mathscr{B}}$ (yellow) coincide. **C.** For the case of "one sample" in the prediction network, learning can be (partially) recovered if the online rule is changed to the form of $pre.\overline{post}.(M - \overline{M})$, where we directly express the firing rate $\overline{post} = < \delta_{s''s'} >_{q(s')}$ of neuron s'' as the probability of that neuron to fire, i.e., $\overline{post} = e^{w_{s''s}}$.

4.4 Discussion

We proposed a general framework for approximating the exact solution of a functional that is expressed as an average quantity. The proposed learning rule is derived from stochastic gradient ascent and has the form of a 3-factor learning rule. The proposed online rule depends on both local Hebbian factors as well as global modulatory factors, if a neural network is used for learning. As such it benefits from a certain degree of biological plausibility and can be applied to various neural network models.

We showed that well-known existing learning methods such as policy gradient and variational learning fall into this general class of 3-factor learning rules. We further applied our model to the same functional from which the SMiLe rule was derived, and showed that it is transformed to a covariance learning rule. We used that learning rule in a spiking neural network to demonstrate that the SMiLe rule can be neurally implemented.

Maximum likelihood estimate as minimization of the Shannon surprise

Maximum likelihood is often used to find the most likely model under which a set of observed input samples/vectors could have been generated by that model. In this method we usually maximize a sum over the log-likelihood of data samples. This is in fact a variant of surprise minimization technique in which the aim of learning is to find a set of parameters under which the observed input samples on average become less surprising than before. The surprise function that is used in maximum likelihood method is the subjective Shannon surprise Eq. (2.1).

Applying stochastic gradient descent for likelihood maximization results in deriving a pure Hebbian online rule. An example was Eq. (4.22) used for training the recognition network. However, if we replace the subjective Shannon surprise by our confidence-corrected surprise, then the online rule becomes a covariance learning rule. In other words, if we apply stochastic gradient on the latter we derive an online rule similar to Eq. (4.24), except that the modulatory term $M(s', s_t)$ [Eq. (4.23)] becomes only the point-estimate of the confidence-corrected surprise, i.e., $\ln q(\theta) - \ln \hat{p}_X(\theta)$ without additional term that controls the stickiness of the model parameters i.e., $\frac{1}{\lambda} \ln \frac{q(\theta)}{\pi_0(\theta)}$ (see Eq. (4.20)). Using confidence-corrected surprise opens a gate to a series of covariance online rules that are driven by a surprise-related signal.

Estimation of the covariance rule

The online rule Eq. (4.24) is a covariance learning rule. For the neural implementation of the maze-exploration task we approximated the covariance using multiple samples randomly drawn from the network. Alternatively, one could approximate the covariance using only one sample. The resulting learning rule would be a modulated

Hebbian rule of the following form

$$\Delta w_{s''s} = \eta \delta_{s_{t-1}s} M(s^*, s_t)(\delta_{s''s^*} - e^{w_{s''s}}), \tag{4.30}$$

where s^* denotes the only sampled state. Although the learning rule above is not capable of learning the maze-exploration task (see Fig. 4.5), it represents a general form of modulated Hebbian rules in which the modulatory factor is the point-estimate of the confidence-corrected surprise where it is approximated only by one sample s^*.

Another extreme case for the implementation of the covariance rule Eq. (4.24) is to analytically calculate the covariance term. This is equivalent to the case in which all states s' had the chance of being sampled with the correct probability. This method results in an online rule of the following form

$$\Delta w_{s''s} = \eta \delta_{s_{t-1}s} \langle \delta_{s''s'} \rangle_{q(s')} \Big(M(s'', s_t) - \langle M(s', s_t) \rangle_{q(s')} \Big), \tag{4.31}$$

where $\langle \delta_{s''s'} \rangle_{q(s')} = e^{w_{s''s}}$ represents the firing rate of post-synaptic neuron.

In other words the covariance rule in Eq. (4.24) has a form of $pre.(post - \overline{post}).(M - \overline{M})$, but the two most extreme cases for approximation of the covariance (i.e., one-sample-based approximation Eq. (4.30) and many-sample-based approximation Eq. (4.31)) have the online forms $M.pre.(post - \overline{post})$ and $pre.\overline{post}.(M - \overline{M})$, respectively. All proposed learning rules above can be considered as variants of 3-factor learning rules.

Relation to a generalized theory of expected utility maximization

Inspired by our method for maximizing a density-dependent functional, we also propose (in **Appendix B**) an information theoretical equivalent of existing models in expected utility maximization, as a standard model of decision making, to incorporate both individual preferences and choice variability. We then show that its exact solution (as the optimal decision making policy) can also be approximated by our functional gradient rule through an online three-factor learning rule.

Learning associations with a novelty signal

In the neural network that we proposed for the recognition network, we assumed that the number of rooms is known for the agent. Therefore, the only thing that the recognition network has to learn is the association between the environmental inputs and the current room (state). In reality, however, the agent does not know the exact number of rooms and it requires to form new memories once a new room is visited for the first time. We can imagine this problem as a clustering task, where the number of clusters is unknown for the agent. Inspired by the Adaptive Resonance Theory (ART) [Carpenter and Grossberg, 1988], we propose a model in **Appendix C** that uses a neurally computed novelty signal to enable the agent to add more clusters whenever it judges an input sample to be novel.

The proposed model requires making a decision about whether an input sample is novel or not. If the new sample is “sufficiently” different from the previous samples, then the new sample is considered to be novel. The problem is then to determine a threshold that indicates the border between the class of familiar samples and the novel samples. In **Appendix D**, we discuss this problem, but in a different context. We propose a theoretical argument that explains why it is difficult for a Bayesian critic (in the context of reinforcement learning) to distinguish similar tasks. Our analysis provides a preliminary hypothesis about the existence of an optimal Bayesian threshold that may determine the border of novelty and familiarity.

4.5 Materials and Methods

Derivation of the online rule for the recognition model

Environmental inputs that the agent receives from rooms are neurally represented as an ensemble of binary vectors $\{\vec{y}(a)\}$. We assume that these vectors encode several environmental features, where each feature uses a 1-out-of-n code. For example if feature number 3 takes the second out of 5 possibilities we would write $(0,1,0,0,0)$ to encode that particular feature. Feature number 3 may correspond to the color of a cue that may take one of 5 possibilities of being blue, red, green, yellow, or black. In the above example, the cue is red. We assume that each input vector $\vec{y}$ encodes m features (e.g., color, shape, size, etc.). Therefore, each $\vec{y}$ is a binary vector with mn entries, where m entries are equal to 1 and the rest of entries are equal to 0. This assumption helps us to easily derive the online learning rule by which we can solve unsupervised clustering tasks [Nessler et al., 2013]. The following paragraph follows unpublished notes of W. Gerstner “Local learning rules for a generative model: Remarks on a paper of Nessler, Pfeiffer and Maass”.

We assume that the inputs are generated by a mixture of multinomial distributions with $k = 16$ components. Each source k generates an input vector $\vec{y}$ with probability $p(\vec{y}|k)$ using the following rule. The probability that the q-th feature of vector $\vec{y}$ (generated by source k) takes the i-th possibility is equal to $\mu_{(i,q)}(k)$, where $\mu_{(i,q)}(k)$ is unknown to the agent. Therefore, the total probability of generating a vector $\vec{y}$ is

$$p(\vec{y}) = \sum_k p(k)p(\vec{y}|k) = \sum_{k=1}^{16} p(k)\Pi_{q=1}^{m}\Pi_{i=1}^{n}\big(\mu_{(i,q)}(k)\big)^{y_{(i,q)}}. \tag{4.32}$$

Here, $y_{(i,q)} := y_j$ denotes the j-th entry of the input vector $\vec{y}$, where j is uniquely determined by pair (i,q) via a 1-to-1 mapping and a shuffling operator.

The aim is to estimate the unknown probabilities $\mu_{(i,q)}(k)$ from the set of data vectors $\{\vec{y}(1), \vec{y}(2), ...\}$ using a neural network. To do so, we parametrize the unknown probabilities $\mu_{(i,q)}(k)$ by synaptic weights $w_{k(i,q)}$ using the transform $\mu_{(i,q)}(k) = exp(w_{k(i,q)})$. Note that $w_{k(i,q)}$ denotes the synaptic weight between neuron $1 < j < mn$ in the first layer, whose index j is uniquely determined by its corresponding pair (i,q), and neuron k in the second layer. Mixture proportions $p(k)$ are also parametrized via $p(k) = exp(w_{k0})$. This gives

$$p(\vec{y}) = \sum_k exp\left(w_{k0} + \sum_{i,q} w_{k(i,q)}y_{(i,q)}\right) = \sum_k exp(u_k). \tag{4.33}$$

The parameter u_k is interpreted as the membrane potential of neuron k in the second layer that sums over all the inputs $y_{(i,q)}$ from the first layer using the weights $w_{k(i,q)}$. We need to choose synaptic weights $w_{k(i,q)}$ so as to maximize the likelihood that the set of observed data vectors $\{\vec{y}(1), \vec{y}(2), ...\}$ could have been generated by our mixture model. We maximize the log-likelihood $L = \sum_a \ln p(\vec{y}(a))$ under the constraint $\sum_i \mu_{(i,q)}(k) = \sum_i exp(w_{k(i,q)}) = 1$ for any k, q.

The maximum of the constraint log-likelihood

$$\tilde{L} = \sum_a \ln p(\vec{y}(a)) - \sum_{k,q} \lambda_k^q \left(\sum_i exp(w_{k(i,q)}) - 1\right), \tag{4.34}$$

where λ_k^q denote Lagrange multipliers, is found by setting the derivatives to zero

$$\begin{aligned} 0 = \frac{\partial \tilde{L}}{\partial w_{k'(i,q)}} &= \sum_a y_{(i,q)}(a)\frac{exp(u_k)}{p(\vec{y}(a))} - \lambda_{k'}^q exp(w_{k'(i,q)}) \\ &= \sum_a y_{(i,q)}(a)p(k'|\vec{y}(a)) - \lambda_{k'}^q exp(w_{k'(i,q)}), \end{aligned} \tag{4.35}$$

where we have introduced

$$p(k'|\vec{y}(a)) = \frac{p(k')p(\vec{y}(a)|k')}{p(\vec{y})(a)} = \frac{e^{u_{k'}(a)}}{\sum_k e^{u_k(a)}}. \tag{4.36}$$

The expression in Eq. (4.36) is interpreted as the *soft-max spiking probability rule* in a stochastic network with lateral interactions that make neuron k' to compete with its neighbors $k \neq k'$.

We sum Eq. (4.35) over i which gives us an explicit expression for the Lagrange multipliers

$$\lambda_{k'}^{q} = \sum_a \sum_i y_{(i,q)}(a) p(k'|\vec{y}(a)) = \sum_a p(k'|\vec{y}(a)). \tag{4.37}$$

Because of the 1-out-of-n coding in each feature of $\vec{y}$, we have $\sum_i y_{(i,q)}(a) = 1$ so that the expression for the Lagrange multipliers and so the online rule (as shown later) become very simple. Note that all the Lagrange multipliers $\lambda_{k'}^{q}$ that depend on source k' are the same, i.e., they are independent of feature $q \in \{1, ..., m\}$ (see Eq. (4.37)). That means it is not necessary to distinguish which input neurons are linked to a particular feature q.

Instead of going directly to the minimum of the constraint log-likelihood $\tilde{L}$ in Eq. (4.34), one can also do a 1-step gradient descent. From Eqs. (4.35),(4.37) we get

$$\begin{aligned} \Delta w_{k(i,q)} = \eta \frac{\partial \tilde{L}}{\partial w_{k(i,q)}} &= \eta \sum_a y_{(i,q)}(a) p(k|\vec{y}(a)) - \sum_a p(k|\vec{y}(a)) exp(w_{k(i,q)}) \\ &= \eta \sum_a p(k|\vec{y}(a)) \big(y_{(i,q)}(a) - exp(w_{k(i,q)}) \big) \end{aligned} \tag{4.38}$$

The weight update in Eq. (4.38) can be easily transformed to our desired learning rule in Eq. (4.22). To do so, we first go from batch to online and neglect the sum over the patterns a. This is a standard step done for all online learning rules in neural network theory. Second, upon repeated representations of pattern a, the factor $p(k|\vec{y}(a))$ gives exactly the probability that the post-synaptic neuron fires. Therefore, thanks to the WTA framework in the second layer, we can replace $p(k|\vec{y}(a))$ with a Kronecker delta function δ_{kk^*}, where k^* denotes the index of winner neuron upon presentation of input $\vec{y}(a)$. The Kronecker delta δ_{kk^*} is in fact a point-estimate of the probability $p(k|\vec{y}(a))$.

Finally if we replace each pair (i, q) with its corresponding index j, we will arrive in the online learning rule $\Delta w_{kj} = \eta \delta_{kk^*}(y_j(a) - exp(w_{kj}))$ that is presented in Eq. (4.22).

Derivation of the online rule for the prediction model

In Chapter 3, our belief about the state transition probabilities from state s to $s' \in \{1, ..., 16\} \setminus s$ was described by a set of parameters $\alpha(s, s')$. These parameters are used for describing 16 Dirichlet distributions, where each distribution is used to model state transition probabilities from a particular state s. We parametrize the model parameters $\alpha(s, s')$ by synaptic weights $w_{s's}$ using the transform

$$w_{s's} = \ln \frac{\alpha(s, s')}{\sum_{s''} \alpha(s, s'')}, \tag{4.39}$$

where $w_{s's}$ denotes the strength of connection from pre-synaptic neuron s in the input layer of the prediction network to the post-synaptic neuron s' in the output layer of the prediction network.

To derive the online rule Eq. (4.24) we need to apply the stochastic gradient descent on the constraint Lagrangian Eq. (4.18), which we re-express it as

$$\tilde{L} = \sum_{\theta} q_{t-1}(\theta) \left(\ln q_{t-1}(\theta) - \ln \hat{p}_X(\theta) + \frac{1}{\lambda} \ln \frac{q_{t-1}(\theta)}{q_{t-2}(\theta)} \right) - \lambda' \left(\sum_{\theta} q_{t-1}(\theta) - 1 \right), \tag{4.40}$$

where we denoted $q(\theta)$ and $\pi_0(\theta)$ in Eq. (4.18) by q_{t-1} and q_{t-2} in Eq. (4.40), respectively, to emphasize that these two distributions correspond to our belief in subsequent time steps. We reserve q_t to denote the updated belief after using the online rule Eq. (4.24).

In Eq. (4.40), $q_{t-1}(\theta)$ denotes a Dirichlet distribution and describes our most recent belief about the state transitions from the last observed state s_{t-1}. The support of $q_{t-1}(\theta)$ is the set of 15-dimensional vectors θ whose entries are real numbers in the interval $(0, 1)$ with the sum of the coordinates is 1. Since our neural network cannot handle the distributions with a “continuous” support, we replace $q_{t-1}(\theta)$ with a suitable categorical distribution (with a “discrete” support):

$$q_{t-1}(\theta) = \Pi_{s'} \left(\frac{\alpha(s_{t-1}, s')}{\sum_{s''} \alpha(s_{t-1}, s'')} \right)^{[\theta = \mathbb{1}_{s'}]} = \Pi_{s'} exp(w_{s's_{t-1}}[\theta = \mathbb{1}_{s'}]), \tag{4.41}$$

where $\mathbb{1}_{s'} \in \mathbb{R}^{15}$ denotes a unity vector with entry s' equal to 1, and $[\theta = \mathbb{1}_{s'}]$ denotes Iverson bracket (a variable that is 1 if the condition $\theta = \mathbb{1}_{s'}$ is fulfilled, and is 0 otherwise). The support of vectors θ now becomes a set of 16 unity vectors $\{\mathbb{1}_{s'}\}_{s'=1}^{16}$, where each unity vector is chosen with the probability

$$q_{t-1}(\theta = \mathbb{1}_{s'}) = \frac{\alpha(s_{t-1}, s')}{\sum_{s''} \alpha(s_{t-1}, s'')} = exp(w_{s' s_{t-1}}). \tag{4.42}$$

The scaled likelihood (which is originally a Dirichlet distribution) can also be replaced by its corresponding categorical distribution, indicating that the likelihood of visiting state s_t if the latent parameter is $\theta = \mathbb{1}_{s'}$ is equal to

$$\hat{p}_X(\theta = \mathbb{1}_{s'}) = \frac{1 + \delta_{s' s_t}}{16}. \tag{4.43}$$

In the surprise modulated belief update (Algorithm 1 in Chapter 3), the parameter γ increases with the surprise $S(X)$ of the newly acquired sample X. Without loss of generality, we assume that $\gamma = \frac{mS(X)}{1+mS(X)}$, where m determines the propensity of a subject to change his belief (see Eq. (3.5)). The Lagrange multiplier parameter $\frac{1}{\lambda}$ in Eq. (4.18) is explicitly linked to the parameter γ that is used in the SMiLe rule via $\gamma = \frac{\lambda}{1+\lambda}$ (see Eq. (3.12)). As such, the Lagrange multiplier λ is linearly linked to the surprise of the new data sample via $\lambda = mS(X)$. Therefore, in combination with Eqs. (4.42), (4.43), the Lagrangian Eq. (4.40) can be expressed as

$$\tilde{L} = \sum_{s'} e^{w_{s' s_{t-1}}} \left(w_{s' s_{t-1}} - \ln \frac{1 + \delta_{s' s_t}}{16} + \frac{w_{s' s_{t-1}} - w^{old}_{s' s_{t-1}}}{mS(X)} \right) - \lambda' \left(\sum_{s'} e^{w_{s' s_{t-1}}} - 1 \right). \tag{4.44}$$

Setting the derivative to zero

$$\begin{aligned} \frac{\partial \tilde{L}}{\partial w_{s'' s_{t-1}}} &= \sum_{s'} \delta_{s' s''} e^{w_{s' s_{t-1}}} \left(w_{s' s_{t-1}} - \ln \frac{1 + \delta_{s' s_t}}{16} + \frac{w_{s' s_{t-1}} - w^{old}_{s' s_{t-1}}}{mS(X)} + (1 + \frac{1}{mS(X)}) \right) \\ &- \lambda' \sum_{s'} \delta_{s' s''} e^{w_{s' s_{t-1}}} = 0, \end{aligned} \tag{4.45}$$

and summing Eq. (4.45) over s'' gives an explicit expression for the Lagrange multiplier

$$\lambda' = \sum_{s'} e^{w_{s' s_{t-1}}} \left(w_{s' s_{t-1}} - \ln \frac{1 + \delta_{s' s_t}}{16} + \frac{w_{s' s_{t-1}} - w^{old}_{s' s_{t-1}}}{mS(X)} + (1 + \frac{1}{mS(X)}) \right) \tag{4.46}$$

Therefore, we can derive the online rule by inserting Eq. (4.46) in Eq. (4.45):

$$\frac{\partial \tilde{L}}{\partial w_{s''s_{t-1}}} = \sum_{s'} e^{w_{s's_{t-1}}} \left(\delta_{s's''} \tilde{M}(s', s_t)\right) - \left(\sum_{s'} e^{w_{s's_{t-1}}} \tilde{M}(s', s_t)\right) \left(\sum_{s'} e^{w_{s's_{t-1}}} \delta_{s's''}\right), \tag{4.47}$$

where

$$\tilde{M}(s', s_t) = w_{s's_{t-1}} - \ln \frac{1 + \delta_{s's_t}}{16} + \frac{w_{s's_{t-1}} - w^{old}_{s's_{t-1}}}{mS(X)} + (1 + \frac{1}{mS(X)}). \tag{4.48}$$

From there there is only a few minor steps to derive the online rule in Eq. (4.24). First, $\sum_{s'} e^{w_{s's_{t-1}}} f(s')$ is replaced by an average $\langle f(s') \rangle_{q_{t-1}}$ because $q_{t-1}(\theta = \mathbb{1}_{s'}) = e^{w_{s's_{t-1}}}$. Second, the modulatory term $M(s', s_t)$ in Eq. (4.23) differs from $\tilde{M}(s', s_t)$ in Eq. (4.48) only in constant terms which creates no problem (according to **Corollary 2**). In all derivations above we assumed that $s = s_{t-1}$. Therefore, the dependency to the activity of the pre-synaptic neuron is manually inserted by $\delta_{ss_{t-1}}$ in Eq. (4.24).

5 Future Works

5.1 Dissociation between negative and positive surprise

Surprise may result either from the occurrence of an unexpected event or the non-occurrence of an expected one. These two cases are known in the literature as positive and negative surprise, respectively, and different neural substrates have been suggested to be responsible for encoding each of them [Alexander and Brown, 2011]. Positive surprise is often more “surprising” than negative surprise consistent with our theory of surprise, where we assign bigger surprise to zero-probability events than low-probability events.

Note that whether surprise is positive or negative does not determine the actual valance of the surprise consequent. In other words, the consequence to surprise can be neutral, pleasant, or unpleasant regardless of surprise being positive or negative. Negative surprise corresponds to the non-occurrence of an expected event. If our expectation is rewarding and the reward is not delivered (i.e., omission of an expected reward), we experience a negative surprise with a negative (unpleasant) valence. If we expect a punishment and it does not occur, we experience a negative surprise but with a positive (pleasant) valence. Positive surprise corresponds to the occurrence of an unexpected outcome. In this case, an event occurs with no prior expectation. If an unexpected reward is delivered we experience a positive surprise with a positive (pleasant) valence, but if we encounter an unexpected punishment then we experience a positive surprise with a negative (unpleasant) valence.

Which computational models can clearly dissociate these two types of surprises is an interesting question that may worth to be studied in future. Moreover, surprise is usually followed by emotional states such as joy or confusion. How we can integrate such emotional responses into the surprise theory is also interesting to be investigated.

5.2 Surprise in language of neurons

Although surprise is often studied in the behavioral level, we argue that surprise is a neural phenomenon which is reflected globally, in emotions among other things. What psychologists call surprise is really something that arises when surprise does not just happen at the sensory level but propagates to higher cognitive and emotion areas and as a result is reflected even in physiological responses and affects behavior.

In the microscopic level of neural activity, surprise might also be viewed as a drive factor that makes a neuron fire (the idea is originally proposed in [Palm, 2012]). Most of the neurons in the brain fire, if they are sufficiently excited or dis-inhibited. One could interpret these excitation and dis-inhibition, as two variants of unexpected surprising events.

In one hand, if a neuron is used to not receive so much excitation from its afferent neurons, then sudden bursts of spikes in its excitatory afferents are unexpected. Occurrence of such unexpected event may surprise the neuron leading to its firing (corresponding to the positive surprise). Many neurons in striatum (a part of basal ganglia) behave like what was just explained. They fire because of sudden excitation they receive from cortical area.

On the other hand, if a neuron is used to receive lots of inhibition from its afferent neurons, then the omission of inhibitory inputs, or a sudden relaxation of such inhibition may lead to its firing (corresponding to the negative surprise). Unexpected omission of inhibitory inputs from the complex of substantia nigra pars reticulata (SNr) and globus pallidus internus (GPi) to the thalamus causes the thalamic neurons to fire, consistent with what mentioned above.

In fact, a neuron fires whenever it is surprised either by excitation (positive surprise) or dis-inhibition (negative surprise); but how surprise can efficiently be calculated in the subthreshold regime of neural activity, and how it is linked to other neural parameters (such as membrane potential) are not yet resolved.

It has been hypothesized that the actual information that is propagated in the brain is linked to surprise [Palm, 2012]. The argument behind this hypothesis is the “sparseness” of the neuronal activity, where it is compared to the volume of the information that is required to internally describe the state of the external world, suggesting that our external world is described only by unexpected and surprising information. This hypothesis would need to be more scientifically investigated.

5.3 Relation between neuromodulators and the global factors

There are different neuromodulators in the brain such as Dopamine (DA), noradrenaline (NE), acetylcholine (ACh), serotonin (5-HT), etc. They play essential roles for the modulation of neural activity all over the brain regions. On the other hand, multi-factor learning rules may depend on several global factors corresponding to the activity of these neuromodulatory systems. An interesting question then would be how different neuromodulators are mapped / associated to the global factors that are theoretically derived.

For instance, we do not know whether assigning reward to DA and surprise to NE is a correct assumption. DA and NE might both encode reward as well as surprise using non-linear mappings. Furthermore, neuromodulators significantly interact with each other, a fact that should be incorporated in theories about global factors.

5.4 Interplay between surprise and different forms of uncertainty

Surprise interplay with different forms of uncertainty. We already saw in our proposed theory that surprise decreases with model uncertainty, and the model uncertainty increases after surprising events. However, a more-in-depth study may be required to understand the exact way that different forms of uncertainty affect each other and surprise. In the following we briefly describe some of the forms of uncertainty:

(1) *World uncertainty* quantifies inherent stochasticity of the environment. This uncertainty is irreducible, even if a perfect model of the environment is available. In Bayesian framework, the *conditional entropy* $\mathscr{H}(x|\theta)$ is a measure of world uncertainty. This quantity calculates how much uncertainty remains in a random variable x, without considering any uncertainty for the model of the world, i.e., the model parameters θ.

(2) *Model uncertainty* determines how uncertain we are about the current model of the environment. This incorporates uncertainty about model parameters or their estimation. Despite of the world uncertainty, the model uncertainty is reducible if additional information is provided. In Bayesian setting, the model uncertainty is quantified by the entropy $\mathscr{H}(\theta)$ of the model parameters. Note that $\mathscr{H}(\theta|x)$ or $\mathscr{H}(\theta|X)$ also quantifies model uncertainty; but after a random variable x or a data sample X has been used to modify our knowledge about the model parameters.

(3) *Expected uncertainty* is the amount of uncertainty that we expect about the outcome of a stochastic world. The expected uncertainty can be quantified by $\mathscr{H}(x)$, where it implicitly incorporates both world uncertainty as well as model uncertainty. An alternative to the expected uncertainty is the *total uncertainty* which is defined as the sum of the world uncertainty and model uncertainty. The total uncertainty is quantified by the joint entropy $\mathscr{H}(x,\theta) = \mathscr{H}(x|\theta) + \mathscr{H}(\theta)$. Although the expected uncertainty is different from the total uncertainty, they both increases with model uncertainty and the world uncertainty but in different ways. Total uncertainty is linearly expressed in terms of world uncertainty and model uncertainty, but the expected uncertainty is non-linearly related to them.

(4) *Unexpected uncertainty* is a different form of uncertainty that is linked to the occurrence of an unexpected event beyond the known stochasticity of the environment. Unexpected uncertainty in our terminology corresponds to surprise, as a result of a violation in a subjective expectation about the environmental outcome. Unexpected uncertainty quantifies the uncertainty that a subject has about whether a fundamental change in the environment is occurred or the unexpected observation is just because of an outlier without any change in the context.

A Palm Theory of Surprise

Uncertainty and subjectivity are amongst essential requirements for evaluating surprise. They may be originated from different sources including uncertainty about *message perception* and the subjective element of *interestingness* in describing the outcomes of a probabilistic world.

Gunther Palm [Palm, 2012] studied novelty and surprise using a new approach to information theory. He has modified the classic information theory to consider the process of perception or formation of the message as well as to incorporate the subjective elements of interestingness by proposing two interesting concepts: *description* and *repertoire*. Here we briefly “review” his proposed method for quantification of novelty and surprise.

In classic information theory, any uncertainty in the perception of a message stems from a noise that is added to the original message through communication channel, i.e., the channel noise. There is no uncertainty about how the message was encoded and thus no uncertainty about how it should be decoded. In other words, both sender and receiver have agreed on a common language (a set of codes) that is used for message transmission. The receiver just need to find out that the received noisy message is indicative of which code. However in reality, different individuals may use different expressions for describing a same message. As a result, we usually receive information through channels in which there is not only channel noise for corruption of the message, but also we suffer from lacking a common agreed language for perception of the message. Different channels provide different amount of information when we have no direct access to the outcome of the stochastic world

A.1 Description and repertoire

Throughout this section, we use terminologies from classic probability theory. The *probability space* (Ω, Σ, p) is a mathematical construct which is used to model real-world stochastic processes. It consists of three segments: (1) a sample space Ω which is defined as the set of all possible *outcomes* $\omega \in \Omega$, (2) a set Σ of *events* defined as the set of all subsets of the sample space Ω. where each event is a subset of Ω containing zero or more outcomes ω, and (3) a *probability* function $p : \Sigma \rightarrow [0, 1]$ that assigns a positive number from 0 to 1 to each event $A \in \Sigma$. For example, in a probabilistic task such as dice throwing, the sample space is all possible numbers which can be appeared, i.e, $\Omega = \{1, 2, 3, 4, 5, 6\}$. The appearance of an even number is an event $A = \{2, 4, 6\}$, and the probability of this event is $p(A) = 0.5$.

Description

Different individuals may use different expressions to describe a same outcome $\omega \in \Omega$. An expression that is used for describing ω may also be correct for a different outcome $\omega' \neq \omega$. Therefore, given an expression, we may not be able to distinguish whether the true outcome is ω or ω'. For instance, if the outcome of a thrown dice is expressed by "an even number", then we cannot distinguish whether the actual outcome is 2, 4, or 6.

In Palm theory of novelty and surprise, a concept called *description* is defined in order to distinguish the way different people describe outcomes of a random process. A description $d : \Omega \rightarrow \Sigma$ is a mapping from the sample space Ω to the event space Σ. Each outcome ω is *described* by an expression $d(\omega) \in \Sigma$ (as a subset of sample space Ω) which is fulfilled by all the elements which comprise it including the outcome $\omega \in d(\omega)$.

Here we provide three different ways of describing the actual outcome of a thrown dice. The person number 1 describes the actual outcome by telling you exactly the number that is appeared when the dice is thrown. The description function that the person number 1 uses for describing the outcomes is $d_1(\omega) = \{\omega\}\ \forall \omega \in \Omega$, indicating that each proposition $d_1(\omega)$ is fulfilled only by the outcome ω. The person number 2 describes each outcome just by letting you know whether the number is even or odd (i.e., $d_2(\omega) = \{1, 3, 5\}\ \forall \omega \in \{1, 3, 5\}$ and $d_2(\omega) = \{2, 4, 6\}\ \forall \omega \in \{2, 4, 6\}$). So, the proposition $d_2(\omega = 4)$ which is used for describing the outcome $\omega = 4$ is fulfilled by all outcomes $\omega \in \{2, 4, 6\}$, because they are all even numbers. The third person describes each outcome by telling that the appeared number is bigger than zero and less than seven (a description that provides no useful information because all the possible outcomes fulfill this condition). This way of description corresponds to $d_3(\omega) = \Omega\ \forall \omega \in \Omega$ (see

Fig. A.1).

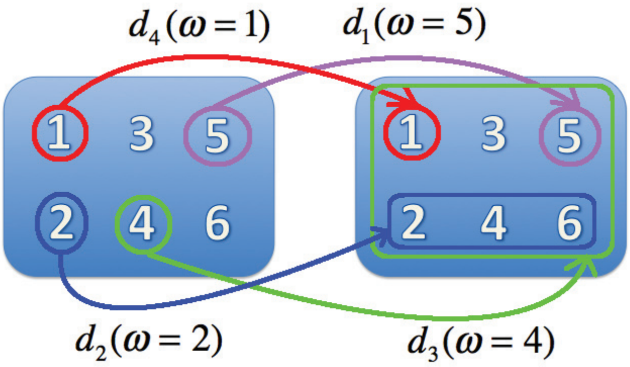

Figure A.1 – **Description functions.** Each outcome ω is described differently using different description functions $d_i(\omega)$. Each description is identified by a color. Blue boxes indicate the set of outcomes Ω, where each description maps an outcome from left box to a subset of outcomes in the right box. For instance, description number 3 (green) describes outcome $\omega = 4$ by a statement that is fulfilled by all outcomes $\{1,...,6\}$.

A description d is *finer* than a description d' (written as $d \subseteq d'$), if $d(\omega) \subseteq d'(\omega)$, $\forall \omega \in \Omega$. In fact, the finer description d more precisely describes the outcomes than d'. In the example above, $d_1 \subseteq d_2 \subseteq d_3$, meaning that d_1 is the most informative way of describing the outcomes and d_3 is the least informative one. Here, informativeness corresponds to the level of uncertainty (about the actual outcome) that is reduced by a description d.

Novelty and Surprise

The *novelty provided by ω for the description d* is defined as $N_d(\omega) = -\log p(d(\omega))$. Note that $N_d : \Omega \to \mathbb{R}^+$ is a random variable. The expected value of this random variable determines the *average novelty* $\mathcal{N}(d) = \mathbb{E}(N_d)$ of the description d, that quantifies, on average, the usefulness of the information provided by a description d in reducing the uncertainty about recognizing the true outcome. In the example above, the average novelty of each description is equal to $\mathcal{N}(d_1) = \log 6$, $\mathcal{N}(d_2) = \log 2$, and $\mathcal{N}(d_3) = 0$.

To quantify how *surprising* the information provided by a description d is, we first construct a new description $\vec{d}$, defined as $\vec{d}(\omega) = \{\omega' : p(d(\omega')) \leq p(d(\omega))\}\ \forall \omega \in \Omega$. The description $\vec{d}$ describes each outcome ω by an expression $\vec{d}(\omega)$ that is fulfilled by all the outcomes ω' for which $p(d(\omega')) \leq p(d(\omega))$. The *surprise provided by ω for the description d* is defined as the novelty provided by ω for description $\vec{d}$, i.e., $S_d(\omega) = N_{\vec{d}}(\omega)$. Just like novelty, surprise $S_d(\omega)$ is also a random variable by which we can define the *average surprise* $\mathscr{S}(d)$ of a description d as the average of this random variable which is equal to the average novelty of directed description $\vec{d}$, i.e., $\mathscr{S}(d) = \mathbb{E}(N_{\vec{d}})$. The average surprise $\mathscr{S}(d)$ of a description measures how different size of all propositions $\{d(\omega_i)\}_{i=1}^{n}$ (used for describing different outcomes $\{\omega_i\}_{i=1}^{n}$) is in comparison to each other.

We already saw that d_1, d_2, and d_3 have different average novelties corresponding to different amounts of usefulness in reduction of uncertainty about the outcomes in the example of dice throwing ($\mathscr{N}(d_1) > \mathscr{N}(d_2) > \mathscr{N}(d_3)$). However, all these descriptions are equally surprising with the average surprise equal to zero, i.e., $\mathscr{S}(d_i) = 0$ for $i = 1,2,3$. The reason is that in all these descriptions $d_i \in \{d_1, d_2, d_3\}$, the novelty $N_{d_i}(\omega)$ is the same for all the outcomes $\omega \in \Omega$. In other words, all the outcomes are *similarly* described for a given description. We as subjects who receive information about actual outcomes through these channels will never be surprised as there is no unpredictability in the way that they describe different outcomes. Mathematically, this is because $\vec{d_i}(\omega) = \Omega \ \forall \omega \in \Omega$ and $i = 1,2,3$.

Now consider a fourth description d_4 which describes all the outcomes except for $\omega = 1$, for which he describes it perfectly. Such describer is modeled as $d_4(\omega) = \Omega \ \forall \omega \in \Omega \setminus \{1\}$ and $d_4(\omega = 1) = \{1\}$. Here the channel is surprising because some one who receives information from this channel faces an unexpected proposition (when outcome $\omega = 1$ happens). For a person who is often used to receive no useful information through this channel, the average surprise of the fourth describer is equal to $\mathscr{S}(d_4) = \frac{1}{6}\sum_{i=1}^{6} S_{d_4}(\omega = i) = \frac{1}{6}\log 6 > 0$ because

$$\begin{aligned} S_{d_4}(\omega = 1) &= N_{\vec{d}_4}(\omega = 1) \\ &= -\log p(\vec{d}_4(\omega = 1)) \\ &= -\log p(\{\omega' : p(d_4(\omega')) \le p(d_4(\omega = 1))\}) \\ &= -\log p(\{\omega' = 1\}) \\ &= \log 6, \end{aligned} \tag{A.1}$$

and $S_{d_4}(\omega) = -\log p(\Omega) = 0 \ \forall \omega \in \Omega \setminus \{1\}$.

Repertoire

Palm has proposed the concept *repertoire*. For a probability space (Ω, Σ, p), a repertoire $\alpha = \{A^i\}_{i=1}^{m}$ (as a subset of $\Sigma \setminus \{\emptyset\}$) is a collection of non-empty expressions A^i that their union covers the whole sample space Ω (i.e., $\cup_{i=1}^{m} A^i = \Omega$, where $\cup$ denotes the union operator). A repertoire α is basically a collection of expressions A^i by which a subject describes different outcomes $\omega \in \Omega$. Each subject may have his own repertoire α as it is determined by his own personal desires and interests.

In the example of game of lotto, imagine some one is interested in those configurations in which 3 or more numbers come in a row. This interest in configurations can be expressed by repertoire $\alpha = \{A_3, A_4, A_5, A_6, \Omega\}$ where $A_n = [n$ numbers in a row$]$ for $n = 3,4,5,6$ is a proposition which is fulfilled by all configurations in which n numbers are in a row. Note that $\cup_\alpha = \Omega$. It means that if she wants to describe configuration $X = \{11,22,33,47,48,49\}$ in example above, she can use $A_3 \in \alpha$ and $\Omega \in \alpha$ as statements she is interested in for describing the outcomes. The way that she perceives both configurations $Y = \{7,16,23,35,40,48\}$ and $W = \{2,7,19,23,37,43\}$ is that they are ordinary configurations. In fact they will be described by her using proposition $\Omega \in \alpha$ which means they do not have any specific property (as she perceived) except that they are ordinary members of sample space Ω with no specific relation between their elements. However, configuration $Z = \{1,2,3,4,5,6\}$ can be described by all propositions in repertoire $\alpha = \{A_3, A_4, A_5, A_6, \Omega\}$ as all the propositions in α is fulfilled by configuration Z.

If we use the propositions within α to describe lottery drawings, we first observe that α is no partition of a sample space Ω as there is overlap between the propositions; on the contrary, $A_6 \subseteq A_5 \subseteq A_4 \subseteq A_3 \subseteq \Omega$. So, if a particular drawing ω fulfills A_6 (like configuration Z in example above), it also fulfills A_5, A_4, A_3, and Ω. Therefore it may be correctly described by all of existing propositions in α. However, a proper choice from α always describes ω by A_6 if there are 6 numbers in a row because it is the most accurate and informative description for such outcome among all the propositions in α. One may also calculate the average novelty and surprise for repertoire α as $\mathscr{N}(\alpha) = \mathscr{N}(d_\alpha)$ and $\mathscr{S}(\alpha) = \mathscr{S}(d_\alpha)$, respectively.

If there is no description, we can also think of the novelty and surprise of a repertoire per se. The idea is that if we were able to be informed of the actual outcomes with no restriction of receiving information from a describer, how should we describe the events in a most accurate way with respect to a given repertoire? For an event $\omega \in \Omega$, the set of all possible descriptions of ω in α is denoted by $\alpha_\omega = \{A^i \in \alpha : \omega \in A^i\}$. What we assume is that a *minimal* proposition in α_ω should be chosen to describe ω. The statement A is minimal if there is no other proposition in α_ω that is contained in A. So a minimal proposition describes the event ω in a most accurate way compared to any other proposition in α_ω. Indeed, for a repertoire α, it would be reasonable to define the novelty of repertoire α provided by ω as the maximal novelty of all propositions in

α_ω, i.e.,

$$N_\alpha(\omega) = \max\{\mathscr{N}(A) : A \in \alpha_\omega\}. \tag{A.2}$$

A description d is called a *proper choice* from α if for every $\omega \in \Omega$, there is no proposition $A \in \alpha$ such that $\omega \in A \subset d(x)$. In fact, a proper choice d has this property that it describes each event ω by a minimal proposition $A \in \alpha_\omega$. In general, there may be several proper choices in α. We denote by $D(\alpha)$ the set of all proper choices from α. The average novelty $\mathscr{N}(\alpha)$ and surprise $\mathscr{S}(\alpha)$ of a repertoire α is defined as:

$$\begin{aligned} \mathscr{N}(\alpha) &= \max\{\mathscr{N}(d) : d \in D(\alpha)\} && \text{(A.3)} \\ \mathscr{S}(\alpha) &= \max\{\mathscr{S}(d) : d \in D(\alpha)\}. && \text{(A.4)} \end{aligned}$$

A repertoire α is called *tight* if there is exactly one proper choice from it. In this case, this choice is denoted by d_α and so $D(\alpha) = \{d_\alpha\}$.

Given a predefined repertoire α for a person, we may also be interested in measuring the amount of novelty or surprise that is provided by an event ω through a description d for that specific person characterized by α. The average novelty or usefulness of a description d with respect to a repertoire α is defined as $\mathscr{N}_\alpha(d) = \mathbb{E}(N_d^\alpha)$ where

$$N_d^\alpha = \max\{\mathscr{N}(A) : d(\omega) \subseteq A \in \alpha\}. \tag{A.5}$$

The novelty of description d with respect to repertoire α, provided by event ω, is considered as a random variable $N_d^\alpha : \Omega \to \mathscr{R}$ and is computed as the maximum possible novelty that you may receive by a proposition $A \in \alpha$ which contains $d(\omega)$. The average novelty, then would be the expectation of such random variable. For any finite repertoire α and any description d, we have $\mathscr{N}_\alpha(d) \leq \mathscr{N}(d)$ and $\mathscr{N}_\alpha(d) = \mathscr{N}(d)$ if and only if $R(d) \subseteq \alpha$ where $R(d)$ denotes the range of description d. In other words, the average novelty of a description with respect to a repertoire is always less than or at most equal to the average pure novelty of that description. The equality holds whenever all the propositions used in description for describing the events is a member of the repertoire, meaning that we are interested in all of the propositions $d(\omega) \in R(d)$. One could also similarly define the average surprise of description d with respect to repertoire α as $\mathscr{S}_\alpha(d) = \mathbb{E}(S_d^\alpha)$ where

$$S_d^\alpha = \max\{\mathscr{S}(A) : d(\omega) \subseteq A \in \alpha\}. \tag{A.6}$$

B Generalized Expected Utility Maximization for Decision Making

A standard model of decision making is *expected utility maximization* [Meyer, 1987] in which a decision maker selects a choice $x^* \in \mathcal{X}$ with the highest *subjective expected utility* $U(x^*)$ among all other alternatives $x \in \mathcal{X}$. In a probabilistic framework, it can be interpreted as selecting choice x^* with probability 1 and choosing the rest with probability 0 (i.e. $P(x) = \delta(x - x^*)$ is the corresponding choice selection density function which determines the likelihood of selecting different choices, where $\delta(.)$ denotes the Kronecker delta function). The density function $P(x) = \delta(x - x^*)$ maximizes the expected value of the utility function $\langle U(x)\rangle_P$ among all possible density functions $P(x)$ because

$$\langle U(x)\rangle_P = \sum_x U(x)P(x) \le \sum_x U(x^*)P(x) = U(x^*) = \langle U(x)\rangle_{\delta(x-x^*)} \,. \tag{B.1}$$

In reinforcement learning, however, choosing the action with the highest value function does not allow for sufficient exploration; this requires choice variability, e.g., by adding noise. Furthermore, individual preferences should be incorporated into the decision making processes, such as action selection in RL.

Expected utility theory accounts for individual differences by explicitly modeling different beliefs about the probabilities of different outcomes. Instead of using a stochastic action selection function (such as a sigmoid) we propose an information theoretical equivalent of existing models to incorporate both individual preferences and choice variability. As such we do not need to impose any specific form of constraints existing in different models. In contrast to maximizing just the average utility

$\langle U(x)\rangle_P$, maximizing the functional

$$\begin{aligned} \mathbb{F}[P] &= \langle U(x)\rangle_P + \frac{1}{\lambda_1}H(P) - \frac{1}{\lambda_2}H(P,P_0) \\ &= \left\langle U(x) - \frac{1}{\lambda_1}\ln P(x) + \frac{1}{\lambda_2}\ln P_0(x)\right\rangle_P, \end{aligned} \tag{B.2}$$

yields a choice selection density function $P(x)$ which not only leads to a relatively high average utility, but also allows exploration. Further, it does not allow the solution to be highly different from a reference P_0. While the entropy $H(P) = \langle -\ln P(x)\rangle_P$ of a density function P in Eq. (B.2) models choice variability, the relative entropy $H(P,P_0) = \langle -\ln P_0(x)\rangle_P$ models subjectivity by involving a subjective reference density function P_0. The minus sign in the last term of the first line in Eq. (B.2) penalizes those density functions that are highly different from a subjective reference P_0. The parameters λ_1 and λ_2 control the fuzziness of the solution by changing the weights of the second and third terms, respectively. By taking the derivative of Eq. (B.2) with respect to P and setting it equal to zero, one could find that the functional in Eq. (B.2) is maximized by

$$P^*(x) = \arg\max_P \mathbb{F}[P] = \frac{P_0(x)^{\frac{\lambda_1}{\lambda_2}}\, e^{\lambda_1 U(x)}}{Z(\lambda_1,\lambda_2)}, \tag{B.3}$$

where $Z(\lambda_1,\lambda_2) = \sum_x P_0(x)^{\frac{\lambda_1}{\lambda_2}}\, e^{\lambda_1 U(x)}$ is the normalizing factor. Equation (B.3) resembles a (modified) Bayes' rule in the sense that the effect of utility U in making the posterior density function P^* is controlled by a free parameter λ_1 and a prior belief P_0 that is affected by the ratio $\frac{\lambda_1}{\lambda_2}$. Although the optimal density P^* that yields the maximal functional value $\mathbb{F}[P^*] = \frac{1}{\lambda_1}\ln Z(\lambda_1,\lambda_2)$ is explicitly derived in Eq. (B.3), it can also be learned using the functional gradient rule Eq. (4.2). This is because the functional Eq. (B.2) can be expressed as $\langle \mathscr{F}[P]\rangle_P$ where $\mathscr{F}[P] = U(x) - \frac{1}{\lambda_1}\ln P(x) + \frac{1}{\lambda_2}\ln P_0(x)$ is a density-dependent functional.

If the maximizer P^* is approximated by any other density $\tilde{P}$, then its corresponding functional value $\mathbb{F}[\tilde{P}]$ differs from its maximal value $\mathbb{F}[P^*]$ in proportion to the KL divergence $D_{KL}(\tilde{P}||P^*) \geq 0$. This is because,

$$\begin{aligned}
\mathbb{F}[P^*] - \mathbb{F}[\tilde{P}] &= \frac{1}{\lambda_1}\ln Z(\lambda_1,\lambda_2) - \langle U(x) - \frac{1}{\lambda_1}\ln\tilde{P}(x) + \frac{1}{\lambda_2}\ln P_0(x)\rangle_{\tilde{P}} \\
&= \frac{1}{\lambda_1}\langle \ln Z(\lambda_1,\lambda_2) - \ln e^{\lambda_1 U(x)} + \ln\tilde{P}(x) - \frac{\lambda_1}{\lambda_2}\ln P_0(x)\rangle_{\tilde{P}} \\
&= \frac{1}{\lambda_1}\langle \ln\frac{Z(\lambda_1,\lambda_2)\tilde{P}(x)}{e^{\lambda_1 U(x)} P_0(x)^{\frac{\lambda_1}{\lambda_2}}}\rangle_{\tilde{P}} = \frac{1}{\lambda_1}\langle \ln\frac{\tilde{P}(x)}{P^*(x)}\rangle_{\tilde{P}} = \frac{1}{\lambda_1} D_{KL}\left(\tilde{P}||P^*\right) \qquad \text{(B.4)}
\end{aligned}$$

B.1 Derivation of the soft-max rule from expected utility maximization

As an example, we investigate a binary decision making task (such as the two-armed bandit problem) in which a subject has to make a decision between two alternatives $x = 1$ and $x = 0$. The probability $P(x)$ of making decision x is modeled by a Bernoulli distribution parametrized by θ such that $P(x) = \theta^x(1-\theta)^{(1-x)}$. We use $P_0(x) = 0.5$ to incorporate no prior preference in making different decisions. We further assume that $\lambda_1 = \lambda_2 = \lambda$ to make the formula simpler. As such, the optimal probability of making the decision $x = 1$ in our binary example is equal to

$$P^*(x=1) = \frac{e^{\lambda U(x=1)}}{e^{\lambda U(x=1)} + e^{\lambda U(x=0)}} = \frac{1}{1+e^{-\lambda\Delta U}}, \qquad \text{(B.5)}$$

where $\Delta U = U(x=1) - U(x=0)$ is the difference between the decisions' utilities. If $U(x=1) > U(x=0)$, the probability $P^*(x=1)$ of making decision $x = 1$ in Eq. (B.5) is greater than 0.5. The parameter λ then determines how big that probability should be for different values of ΔU. Note that the stochastic (sigmoid) action selection function, which is used in the expected utility theorem for modeling choice variability, is explicitly derived in Eq. (B.5) as the optimal solution in the sense that it maximizes the functional Eq. (B.2).

C Learning Associations with a Neurally Computed Global Novelty Signal

Here we propose a model to measure novelty based on the activity of decision units in a decision making process in neural networks. We also propose a way by which novelty can implicitly modulate (gate) Hebbian plasticity to control learning at underlying synapses. We apply our model to a clustering task, in which the total number of clusters is initially unknown to the network. The proposed model is able to add more clusters whenever it judges an input sample to be novel, i.e., a sample which may belong to none of the existing clusters. This model represents an agent that is able to generate (trigger) new states, an essential feature for learning new environments. The proposed model can be used to explain how humans and animals efficiently encode and discover the inherent characteristics of the environment with which they interact.

We argue that the novelty signal in this framework can be interpreted as a (global) modulatory signal, corresponding to the diffusion of a non-specific neuromodulator (e.g., norepinephrine (NE) released from locus coeruleus (LC) neurons) and can modulate the local Hebbian factors (i.e., the coactivity of pre- and post-synaptic neurons) in synaptic plasticity rules. As such, it can be considered as a biologically plausible third factor in multi-factor learning rules.

C.1 A neural network model

We use a two-layer feed-forward neural network with Oja's learning rule in the framework of competitive learning (winner-takes-all (WTA)) to classify input patterns (see Fig. C.1A). The network consists of $m_{max} = m_{used} + m_{free}$ output units, corresponding to the maximum number of clusters the system can learn. Each of the m_{used} units represents a previously learned cluster (e.g., known fruits such as apple and banana). m_{free} is the number of *loser* units which have never been used for describing environmental features. These are called loser units as they always lose the competition in

WTA network [Hertz et al., 1991] unless a novel pattern (e.g., an unknown fruit such as rambutan or jabuticaba) is presented.

Each output unit $x_i = g(h_i)$ is fully connected to a set of inputs x_j via excitatory connections $w_{ij} \geq 0$. Here $g : \mathbb{R}^+ \to [0,1]$ is the activation function and

$$h_i = b + E_i - I_c - I_n \, \Theta(\bar{x}_i), \tag{C.1}$$

is the net input to neuron i, where $b \geq 0$ is a common excitatory input (constant baseline activity), $0 \leq E_i = \sum_j w_{ij} x_j \leq 1$ is the normalized forward excitatory input, I_c stands for the competitive inhibition (required for implementing WTA) and

$$I_n = c \, \mathcal{N}(X) \geq 0, \tag{C.2}$$

with a constant c is the absolute value of the inhibitory signal proportional to the novelty $0 \leq \mathcal{N}(X) \leq 1$ of each presented input pattern X. The Heaviside function $\Theta(.)$ ensures that the novelty-related inhibitory signal is applied only to *active* output units (i.e., neurons with $\bar{x}_i > 0$, where $\bar{x}_i$ is the mean activity, averaged over a large number of past examples). In what follows, we assume that

If the presented input pattern X is *not* novel, then we can neglect the influence of inhibition I_n [Eq. (C.2)] on the net input h_i [Eq. (C.1)]. Therefore, one of m_{used} output units (denoted by i^*) that receives *maximal* excitation E_i remains activated. All other output units becomes silent as a result of WTA architecture. The set of afferent weights to unit i^*, denoted by $W_{i^*} = (w_{i^*1}, \ldots, w_{i^*n})$, represents the prototype of a cluster to which input X belongs.

Dynamics of WTA network causes only one output unit (the winner i^*) to remain active. This makes sure that only the set of afferent weights to the winner unit, i.e., W_{i^*}, is modified. This is a consequence of the Oja's learning rule:

$$\Delta w_{ij} = \eta x_i (x_j - x_i w_{ij}), \tag{C.3}$$

in which learning is gated by the activity x_i of the post-synaptic neuron. That is if the post-synaptic neuron is not activated (i.e., $x_i = 0$), there is no change in its afferent synaptic weights ($\Delta w_{ij} = 0, \ \forall j$).

According to Oja's learning rule [Eq. C.3], the set of afferent weights to the winner unit i^* (with $x_{i^*} = 1$) is changed by $\Delta W_{i^*} = \eta(X - W_{i^*})$ where η is a small learning rate. Such an update shifts the prototype W_{i^*} a bit towards X. After many trials, the prototype approaches to the center of all data samples X that are classified in the corresponding cluster.

If data sample X is *sufficiently* novel, then the inhibitory term I_n will exceed the excitatory term E_i. Therefore, the net input h_i in Eq. (C.1), for all m_{used} output units (with $\bar{x}_i > 0$), falls below the baseline $b - I_c$. The net input to all m_{free} loser units (with $\bar{x}_i = 0$), however, is bigger than $b - I_c$ because they never receive such an inhibition unless they are used for learning environmental features. Therefore, in the presence of a novel sample, one of the loser units finds the chance of winning the competition in the WTA network. As such only one of m_{free} units finally remains active in the output layer. A learning gate is then open for the afferent weights of that unit, corresponding to creation of a new prototype. Once a loser unit is used for representing a new cluster for describing environmental features, it no longer belongs to the set of loser units.

C.2 Our measure of novelty

The novelty signal $\mathcal{N}(X)$ in our model is negatively proportional to the maximum excitatory input to the neurons in the output layer i.e.,

$$\mathcal{N}(X) = 1 - max_i[E_i]. \tag{C.4}$$

The idea behind this measure is that if *none* of the decision units in the output layer is sufficiently activated, there is not enough evidence that the input pattern belongs to one of existing clusters that they represent. Therefore, the input sample is novel with respect to what the system has learned (encoded as synaptic efficacies between input and output layers).

The inhibitory signal I_n in Eq. (C.2) not only depends on the novelty signal [Eq. C.4], but also it depends on a constant c. We propose to choose

$$c = \frac{T}{1-T}, \; 0 \le T < 1. \tag{C.5}$$

Here, T is indicative of a threshold. If the activity of *none* of m_{used} output units is above T (i.e., if $max_i \; E_i < T$), then the input data sample X is recognized as sufficiently novel (with $\mathcal{N}(X) > 1 - T$). The net input h_i [Eq. (C.1)] to all m_{used} output units falls below $b - I$ because:

$$h_i = b + E_i - I_c - I_n < b + max_i \; E_i - I_c - c\mathcal{N}(X) < b + T - I - \frac{T}{1-T}(1-T) = b - I. \tag{C.6}$$

As a consequence, a former loser unit now becomes the winner of the WTA competition and a new prototype will be built. The threshold T in Eq. (C.5) is a *subjective* parameter. A larger T results in a *finer* description of the environment. Data samples are more frequently recognized as novel (and so more prototypes or clusters are created) for a large T than a low T. On the other hand the coarsest description of the environmental features is achieved by the smallest threshold, i.e., $T = 0$. In that case, no data sample is recognized as novel and all the inputs are classified into one of the existing clusters.

C.3 Simulation

We apply our proposed model to an online clustering task in which the number of prototypes is unknown to the subject. The aim is to learn the inherent structure of the input data space in an unsupervised fashion.

Two-dimensional input patterns randomly drawn from four multivariate normal distributions with underlying means $(\pm 3, \pm 3)$ and covariance matrix $2.56\, \mathbb{I}_2$ (where $\mathbb{I}_n$ is an $n \times n$ identity matrix) are given to the network in four blocks. Each block contains 100 samples and has a different mean. The model is capable of recognizing all 4 clusters, although from the beginning it was not aware of how many clusters exists in the dataset (see Fig. C.1B). The novelty signal calculated by Eq. (C.4) is depicted in Fig. C.1C. As expected, the novelty signal rises at those time steps where that samples from a different distribution is presented (i.e., $t = 101, 201, 301$).

C.4 Discussion

We proposed a simple model in which a novelty signal is efficiently used for generating new memories without disruption of past learned memories. Novelty signal triggers an inhibitory signal (whenever the input pattern is not familiar for the prototypes that have been learned) and drives learning at synapses that do not yet represent environmental features. Using this method we adapts number of output units we need for describing environmental features.

Once an input pattern is presented to the network, the activity of each output unit determines to what extent it is likely that the input belongs to its corresponding cluster. A higher activity then corresponds to a bigger extent. In classic WTA network, a unit with maximum activity always wins the competition no matter how much it is activated. However, in our framework low-activation is not acceptable for a unit to be considered as a winner in the WTA network. The winner unit does not only have a higher activity compared to the outher output units, but also it has to have an activity that is larger than a given threshold T.

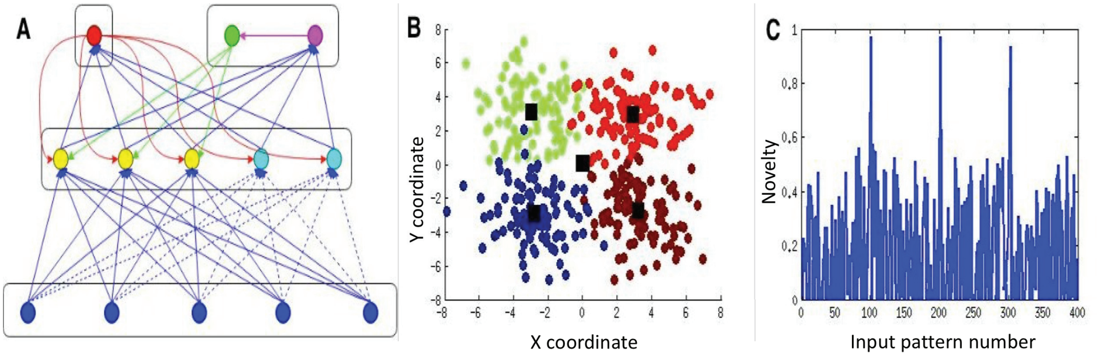

Figure C.1 – **Neural network model for novelty-based clustering. A.** Input units (blue circles in bottom layer) provide excitatory input for excitatory decision units in middle layer consisting of both loser units (cyan circles) and active units (yellow circles) through weak (dashed arrows) and strong (solid arrows) connections, respectively. WTA between decision units is implemented via mutual inhibition through a common inhibitory pool (red circle). Additionally, the decision units project to an inhibitory unit (magenta circle) via pre-defined non-plastic connections to determine their maximum activity. A novelty detector (green circle) strongly inhibits the active units if the maximum activity among decision units is not high enough. It will increase the chance that one of always-loser units wins the competition. Hence, a new cluster is created. **B.** Two-dimensional input patterns (colored dots). The network is able to add new clusters (black squares) whenever a novel pattern is presented. Each color indicates the cluster into which samples are classified. The black squares among colored dots represent the center of each learned cluster. Black squares at $(0,0)$ correspond to weak synapses afferent to loser units which have not yet been used for learning a new cluster. **C.** The measured novelty for each input pattern presented to the network. Whenever the distribution from which samples are drawn is changed (i.e., trials $101, 201$, and 301), the novelty signal increases resulting in creation of a new cluster.

A functional requirement for our model to work properly is that the novelty signal has to be generated fast enough before WTA network (using competitive inhibition I_c) determines the winner neuron. Therefore the dynamics of the novelty-triggered inhibitory signal must be faster than the competitive inhibition. This is a plausible assumption, as there is neurophysiological evidence that the LC-NE system (as neural correlate of novelty) is fast enough to affect the outcome of a decision process (WTA network), in the cortical areas [Usher et al., 1999].

The inhibitory signal I_n in Eq. (C.2) is a *global* signal. The significance of this model is that the novelty signal is globally available for all the output units, and can be considered as a neuromodulatory signal which implicitly affects Hebbian plasticity. Moreover, novelty only depends on the *maximum* activity in the output layer. In other words, we do not need to track the activity of whole network in order to calculate novelty. This is not, however, the case for other novelty measures whose calculation depends on many parameters that have to be collected from the whole network.

The stability of categories formed by competitive learning is not guaranteed as there is always a weight change if a new pattern is presented to the network. This problem can be solved by freezing the learned categories by gradually stopping the learning. However, stopping the ability to learn causes the network to lose its plasticity and not to react to any new data. This is why we need a triggering signal like novelty that opens a gate for a set of synapses that should be used for learning new memories. However, too much plasticity in the neural system results in forgetting past memories. Adaptive resonance theory (ART) [Carpenter and Grossberg, 1988] has been proposed to address the stability-plasticity dilemma by proposing a set of neural network models to address the problem of pattern recognition. The idea behind the ART is very similar to what we proposed in our model. However, ART models mainly use artificial neural networks with algorithmic approach to address the problem, while our proposed model tries to explain how novelty can be neurally calculated and how it can be incorporated into plasticity rules. Although our current model is not yet fully implemented in a biological plausible setting, we believe that our proposed model can provides some insights about how memories can be formed in our brain.

An application of the proposed model is to address one of the challenges in clustering problem: how many prototypes is needed to efficiently encode the input data space? By efficiency, we mean trying not to use too many prototypes for describing an environment, if a few is enough for doing so. In fact, the number of prototypes one need for encoding the environment should be adaptively adjusted according to the level of environment's complexity. Despite of classic clustering algorithms like K-means in which the total number of clusters is initially known to the learning agent (a machine or a neural system), our brain does not consider a pre-defined number of prototypes for describing the world. Instead, it gradually increases the number of prototypes as the complexity of the environment increases.

D What Hinders Bayesian Optimal Critic to Distinguish Similar Tasks?

In the framework of reinforcement leaning guided by a critic [Sutton and Barto, 1998a, Doya, 2000, Frémaux et al., 2013], simple running average can be used to estimate the expected reward corresponding to a single task which is repeated several times during training. However, if multiple tasks have to be learned in parallel, then the critic has to distinguish different stimuli corresponding to different tasks in order to estimate the expected reward separately for each underlying task [Frémaux et al., 2010].

Psychophysical experiments on perceptual learning have shown that learning two similar tasks in parallel is impossible [Herzog et al., 2012, Tartaglia et al., 2009]. It suggests that the brain has an *imperfect critic* to calculate the expected reward in a sense that if task A and task B are highly similar, the critic can no longer distinguish them. As such, it is not able to correctly estimate the expected reward for each of the two tasks, separately.

Here we would like to address this problem from a theoretical point of view. We formulate this problem as a simple mathematical problem in the framework of statistical decision theory, to give an abstract theoretical answer to experimental finding above. We argue that if the two tasks are sufficiently similar, reporting them as a single task would be less riskier than reporting them as two separate tasks. Therefore, it makes sense (from the view of decision making theory) for the subject not to make a riskier decision (i.e., reporting them as separate tasks).

D.1 Argument 1: using statistical decision theory

Imagine that there are two different tasks A and B that must be learned in parallel. We model different stimuli corresponding to tasks A and B as realizations of two Gaussian distributions $\mathcal{N}(\mu, 1)$ and $\mathcal{N}(-\mu, 1)$, respectively. Here, $\mu \geq 0$ is the absolute mean value of the underlying distributions. For the simplicity we fixed the standard deviation σ of Gaussian distributions to 1 in order that similarity between task A and task B is determined just by μ; otherwise the ratio $\frac{\mu}{\sigma}$ determines how much two tasks are similar. Smaller μ indicates more similarity because it causes the Gaussian distributions to have overlap (see Fig. D.1). In other words, when μ is small, the distinction between the two tasks becomes difficult.

We assume that tasks A and B are interleaved during training; called *roving* condition [Herzog et al., 2012, Tartaglia et al., 2009]. That is at each time step, either task A or B is chosen with 50% chance. Then a random number corresponding to a noisy stimulus associated to the chosen task is presented accordingly. As such, stimulus x which represents either task A or task B can be modeled as a random variable with probability density function

$$P(x) = 0.5\,\mathcal{N}(\mu, 1) + 0.5\,\mathcal{N}(-\mu, 1). \tag{D.1}$$

Using observed samples $\mathbf{X} = \{x_i\}_{i=1}^{n}$, a decision rule $\delta(\mathbf{X})$ is defined as a *statistics* which can be used for estimating the true unknown parameter μ. For a given observation set $\mathbf{X}$, two hypotheses (represented as two decision rules) should be compared. If the critic believes in existence of a single task (the first hypothesis), then the sample mean

$$\delta_1(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{D.2}$$

is a good candidate for estimation of the underlying parameter μ. If it believes in two different tasks (the second hypothesis), then decision rule

$$\delta_2(\mathbf{X}) = 0.5\,\Big(\frac{1}{|\mathscr{A}^+|}\sum_{i\in\mathscr{A}^+} x_i - \frac{1}{|\mathscr{A}^-|}\sum_{i\in\mathscr{A}^-} x_i\Big), \tag{D.3}$$

could be considered as a simple and rational method for the point estimation of unknown parameter μ. Here $\mathscr{A}^+$ and $\mathscr{A}^-$ denote the set of indices for which samples are non-negative ($x_i \geq 0$) or negative ($x_i < 0$) values, respectively. Equation (D.3) calculates half of distance between the sample means of positive and negative samples; a quantity which can be used for estimation of μ.

A key notion of decision theory is that decision rules (here $\delta_1(\mathbf{X})$ and $\delta_2(\mathbf{X})$) should be compared by their *risk* functions. A less riskier decision is more appreciated [Pratt et al., 1995]. The risk of a decision rule is defined as the *expected loss* incurred when that decision rule is employed. Here we use a quadratic loss function

$$\mathscr{L}(\mu, \delta(\mathbf{X})) = (\mu - \delta(\mathbf{X}))^2, \tag{D.4}$$

which indicates how much critic loses because of its estimation method $\hat{\mu} = \delta(\mathbf{X})$ as a decision rule, if true parameter is μ. Since loss function $\mathscr{L}$ in Eq. (D.4) depends on observation $\mathbf{X}$, its expectation with respect to underlying distribution $P(x)$ in Eq. (D.1) is defined as risk

$$R(\mu, \delta) = \mathbb{E}[\mathscr{L}(\mu, \delta(\mathbf{X}))]. \tag{D.5}$$

Note that the risk function in Eq. (D.5) does not depend on observation $\mathbf{X}$, and it solely depends on true parameter μ and applied decision rule $\delta(.)$ as a point estimation method. Fig. D.2 depicts risks $R(\mu, \delta_1)$ and $R(\mu, \delta_2)$ associated with decision rules in Eqs. (D.2) and (D.3), respectively (details of calculation are given at the end of Appendix).

We argue that this finding is actually an answer to explain why critic is not able to distinguish similar tasks. The reason is that if two tasks are sufficiently similar (which is the case if $\mu < \mu^*$ where μ^* is associated with the intersection point in graph of Fig. D.2)), decision rule Eq. (D.2) (i.e., no distinction between the two tasks) is less riskier than its alternative in Eq. (D.3) (i.e., decision on their distinctions). Hence, if critic decides not to distinguish two similar tasks, it might be because of a rational decision making strategy, that prefers less riskier decisions than more riskier ones [Pratt et al., 1995].

D.2 Argument 2: using Bayesian reasoning

We show that Bayesian theory can also explain such a phenomenon. Let us assume that stimuli are generated by a mixture of two Gaussian distributions

$$P(x|\mu, \sigma) = 0.5\,\mathcal{N}(\mu, \sigma) + 0.5\,\mathcal{N}(-\mu, \sigma), \tag{D.6}$$

where mean μ and standard deviation σ are unknown parameters. A uniform prior $\pi(\mu, \sigma)$ over two-dimensional space (μ, σ) is assumed. Using Bayes' rule we calculated posterior distribution over parameters (μ, σ) after observation of $n = 1000$ iid samples, where the true parameters (μ, σ) were $(0.5, 1)$ (two similar tasks) and $(2, 1)$ (two different tasks).

Joint posterior distribution $\pi(\mu, \sigma|\mathbf{X})$ is depicted in Fig. D.3 and Fig. D.4 when true parameters are $(0.5, 1)$ and $(2, 1)$, respectively. Despite Fig. D.4 (case of different tasks) in which the posterior distribution $\pi(\mu, \sigma|\mathbf{X})$ looks like a bump around true parameters $(0.5, 1)$, in Fig. D.3 (case of similar tasks) the posterior distribution is bended over a wider region. It shows that if two tasks are similar (i.e., when $\mu < \mu^*$), Bayesian framework is not also capable of distinguishing two tasks because region around $(0, 1.2)$ (one task with a bigger standard deviation than the true value) is as likely as region around $(0.5, 1)$ (two different tasks).

We also numerically calculated the amount of uncertainty in final estimation (measured as the entropy of joint posterior distribution $\pi(\mu, \sigma|\mathbf{X})$) for each true mean value μ in the range of $[0, 2]$. Fig. D.5 shows that for smaller μ estimation uncertainty is much higher than that for bigger μ; consistent with the results inferred from Fig. D.3 and Fig. D.4.

We would like to emphasize that the experimental results in [Herzog et al., 2012, Tartaglia et al., 2009] are consistent with the logic of a critic. A critic which distinguishes *all* different stimuli suffers from a specific form of overfitting and would never generalize. At the other extreme case, a critic which maintains just a single running average for all stimuli is not able to learn those tasks in parallel. Therefore, an optimal critic must detect the essence of stimulus and neglect the noise, which requires building clusters on input data set. If the stimuli from two different tasks are very similar, then it makes sense if the critic assign only one cluster to all of them, and not distinguishing as separate tasks.

D.3 Calculation of the risk functions

$$
\begin{aligned}
R_1(\mu) &= R(\mu, \delta_1) = \mathbb{E}[(\mu - \delta_1(\mathbf{X}))^2], \\
&= \mathbb{E}[(\mu - \frac{1}{n}\sum_i x_i)^2] \\
&= \mu^2 + \frac{1}{n^2}\mathbb{E}[\sum_i x_i^2 + \sum_{i \neq j} x_i x_j] - \frac{2\mu}{n}\mathbb{E}[\sum_i x_i] \\
&= \mu^2 + \frac{1}{n^2}\Big(n\mathbb{E}[x^2] + n(n-1)\mathbb{E}^2[x]\Big) - \frac{2\mu}{n} n\mathbb{E}[x] \\
&= (1 + \frac{1}{n})\mu^2 + \frac{1}{n} \\
&\overset{n\to\infty}{\approx} \mu^2.
\end{aligned}
\tag{D.7}
$$

$$
\begin{array}{rcl}
R_2(\mu) & = & R(\mu,\delta_2) = \mathbb{E}[(\mu - \delta_2(\mathbf{X}))^2], \\
 & = & \mathbb{E}[(\mu - \frac{0.5}{|\mathscr{A}^+|}\sum_{i\in\mathscr{A}^+} x_i + \frac{0.5}{|\mathscr{A}^-|}\sum_{i\in\mathscr{A}^-} x_i)^2] \\
 & \overset{(i)}{\approx} & \mathbb{E}[(\mu - \frac{1}{n}\sum_{i\in\mathscr{A}^+} x_i + \frac{1}{n}\sum_{i\in\mathscr{A}^-} x_i)^2] \\
 & = & \mu^2 + \frac{1}{n^2}\mathbb{E}[\sum_{i\in\mathscr{A}^+} x_i^2 + \sum_{i\neq j\in\mathscr{A}^+} x_i x_j] + \frac{1}{n^2}\mathbb{E}[\sum_{i\in\mathscr{A}^-} x_i^2 + \sum_{i\neq j\in\mathscr{A}^-} x_i x_j] \\
 & - & \frac{2\mu}{n}\mathbb{E}[\sum_{i\in\mathscr{A}^+} x_i] + \frac{2\mu}{n}\mathbb{E}[\sum_{i\in\mathscr{A}^-} x_i] - \frac{2}{n^2}\mathbb{E}[\sum_{i\in\mathscr{A}^+} x_i \sum_{i\in\mathscr{A}^-} x_i] \\
 & = & \mu^2 + \frac{1}{n^2}\Big(\frac{n}{2}\mathbb{E}[x_+^2] + \frac{n}{2}(\frac{n}{2}-1)\mathbb{E}^2[x_+]\Big) + \frac{1}{n^2}\Big(\frac{n}{2}\mathbb{E}[x_-^2] + \frac{n}{2}(\frac{n}{2}-1)\mathbb{E}^2[x_-]\Big) \\
 & - & \frac{2\mu}{n}\frac{n}{2}\mathbb{E}[x_+] + \frac{2\mu}{n}\frac{n}{2}\mathbb{E}[x_-] - \frac{2}{n^2}\frac{n^2}{4}\mathbb{E}[x_+]\mathbb{E}[x_-] \\
 & \overset{(ii)}{=} & \mu^2 + \frac{1}{n}\mathbb{E}[x_+^2] + (1-\frac{1}{n})\mathbb{E}^2[x_+] - 2\mu\mathbb{E}[x_+] \\
 & \overset{(iii)}{=} & \mu^2 + \frac{1}{2n}\Big((1+\mu^2)[1-\Phi(-\mu)] + \mu\phi(-\mu) + (1+\mu^2)[1-\Phi(\mu)] - \mu\phi(\mu)\Big) \\
 & + & \frac{1}{4}(1-\frac{1}{n})\Big(\mu[1-\Phi(-\mu)] - \phi(-\mu) - \mu[1-\Phi(\mu)] - \phi(\mu)\Big)^2 \\
 & - & \mu\Big(\mu[1-\Phi(-\mu)] - \phi(-\mu) - \mu[1-\Phi(\mu)] - \phi(\mu)\Big) \\
 & \overset{(iv)}{=} & \mu^2 + \frac{1}{2n}(1+\mu^2)[2-\Phi(\mu)-\Phi(-\mu)] \\
 & + & \frac{1}{4}(1-\frac{1}{n})\Big(\mu[\Phi(\mu)-\Phi(-\mu)] - 2\phi(\mu)\Big)^2 \\
 & - & \mu\Big(\mu[\Phi(\mu)-\Phi(-\mu)] - 2\phi(\mu)\Big) \\
 & \overset{(v)}{=} & (1+\frac{1}{2n})\mu^2 + \frac{1}{2n} + \frac{1}{4}(1-\frac{1}{n})[\mu\, erf(\frac{\mu}{\sqrt{2}}) - 2\phi(\mu)]^2 \\
 & - & \mu[\mu\, erf(\frac{\mu}{\sqrt{2}}) - 2\phi(\mu)] \\
 & = & (1+\frac{1}{2n})\mu^2 + \frac{1}{2n} + \frac{1}{4}(1-\frac{1}{n})\Psi^2(\mu) - \mu\Psi(\mu) \\
 & \overset{n\to\infty}{\approx} & \frac{1}{4}\Psi^2(\mu) - \mu\Psi(\mu) + \mu^2 \qquad \text{(D.8)}
\end{array}
$$

In calculation above, we used approximation $|\mathscr{A}^+| \approx |\mathscr{A}^-| \approx \frac{n}{2}$ in (i) and equalities $\mathbb{E}[x_-^2] = \mathbb{E}[x_+^2]$ and $\mathbb{E}[x_-] = -\mathbb{E}[x_+]$ in (ii). In (iii) we used the fact that $\mathbb{E}[x_+] = 0.5[f(\mu)+$

$f(-\mu)]$ and $\mathbb{E}[x_+^2] = 0.5[g(\mu) + g(-\mu)]$ where

$$f(\mu) = \int_0^\infty x\mathcal{N}(\mu,1)\,dx = \mu[1-\Phi(-\mu)] - \phi(-\mu), \tag{D.9}$$

$$g(\mu) = \int_0^\infty x^2\mathcal{N}(\mu,1)\,dx = (1+\mu^2)[1-\Phi(-\mu)] + \mu\phi(-\mu). \tag{D.10}$$

Here $\phi(x) = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$ is the standard normal probability density function and $\Phi(x) = \int_{-\infty}^{x}\phi(t)dt = \frac{1}{2}\left(1 + erf(\frac{x}{\sqrt{2}})\right)$ is its corresponding cumulative distribution function where $erf(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$ is the error function. We further used equalities $\phi(\mu) = \phi(-\mu)$ in (iv) and $\Phi(\mu) + \Phi(-\mu) = 1$ as well as $\Phi(\mu) - \Phi(-\mu) = erf(\frac{\mu}{\sqrt{2}})$ in (v) to simplify the final expression written as a function of μ and $\Psi(\mu) := \mu\, erf(\frac{\mu}{\sqrt{2}}) - 2\phi(\mu)$.

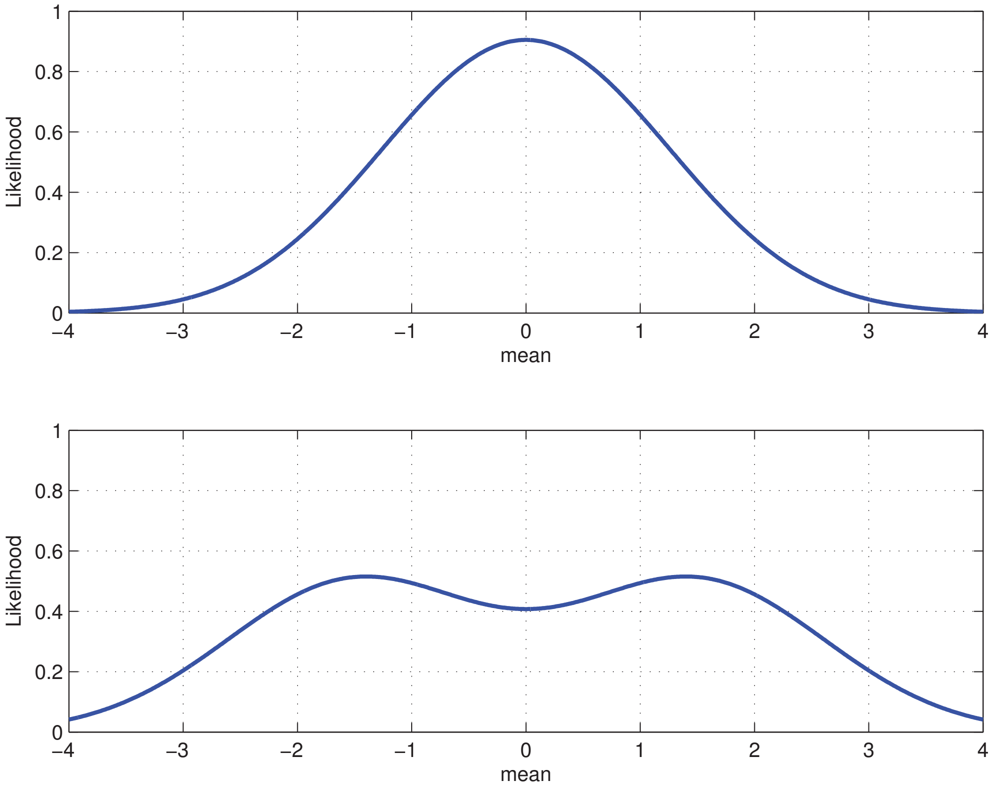

Figure D.1 – **Mixture of two Gaussian distributions.** Probability density function $P(x) = 0.5\,\mathcal{N}(\mu,1) + 0.5\,\mathcal{N}(-\mu,1)$ for $\mu = 0.5$ (top) and $\mu = 2$ (bottom). Two Gaussian distributions have overlap when true mean μ is small. It makes difficult to distinguish two similar tasks corresponding to each of Gaussian distributions $\mathcal{N}(\mu,1)$ and $\mathcal{N}(-\mu,1)$.

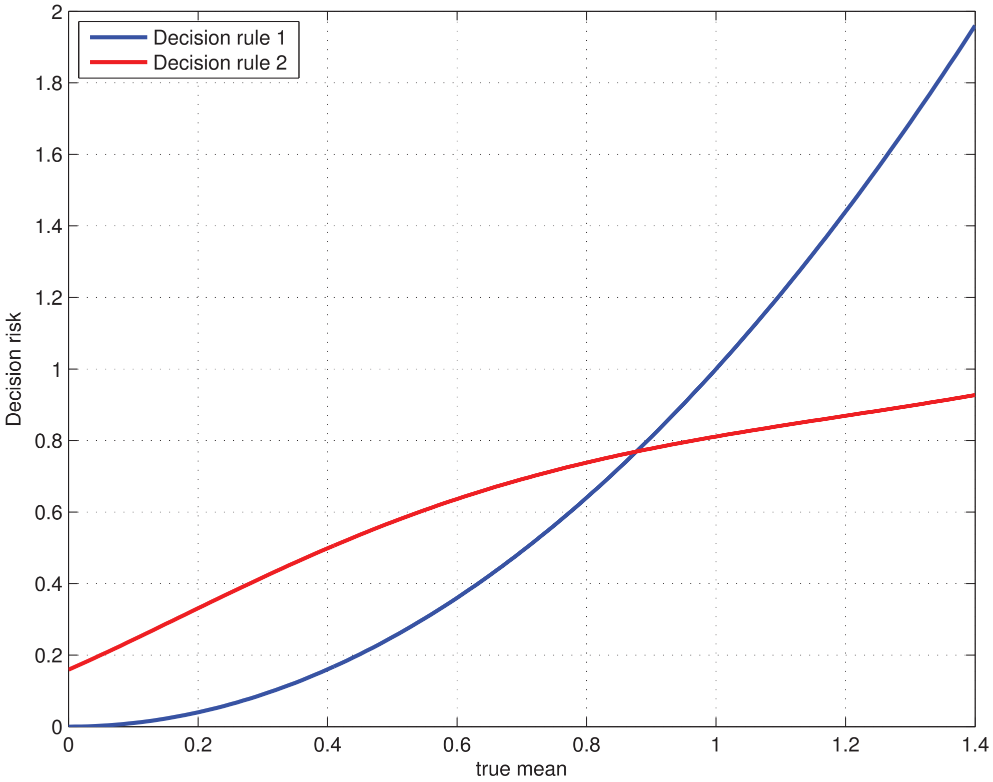

Figure D.2 – **Decision risks for optimal Bayesian critic**. Risk functions corresponding to two decision rules Eq. (D.2) in blue and Eq. (D.3) in red. For similar tasks (i.e., for $\mu < \mu^*$ where μ^* corresponds to the intersection point), deciding not to distinguish two tasks is less riskier than deciding to distinguish them. As such, a rational critic (which minimizes its decision risk) must not distinguish the two tasks if they are sufficiently similar.

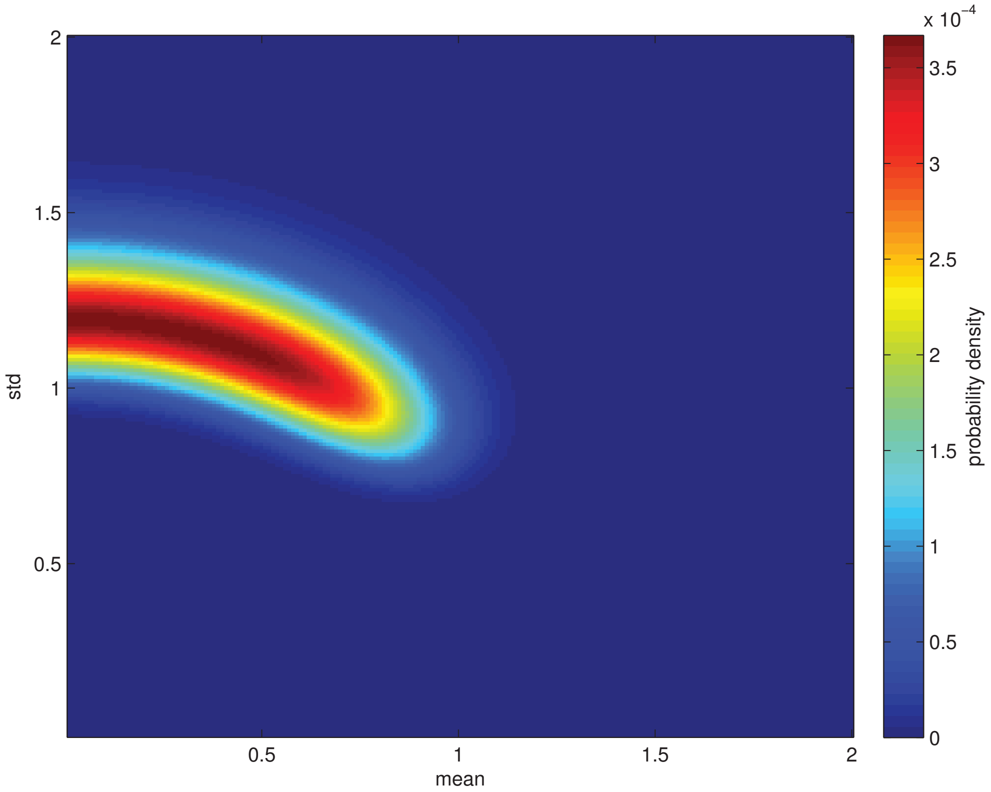

Figure D.3 – **Bayesian decision for identification of two similar tasks.** The joint posterior distribution $\pi(\mu, \sigma|\mathbf{X})$ obtained by the Bayes' rule (after 50 samples) when the true mean (horizontal axis) is $\mu = 0.5$ and the true standard deviation (std, vertical axis) is $\sigma = 1$. Since the true mean is small (two tasks are similar), the Bayesian agent cannot distinguish whether the samples are generated by one Gaussian component (with higher std) or two Gaussian components (with smaller std).

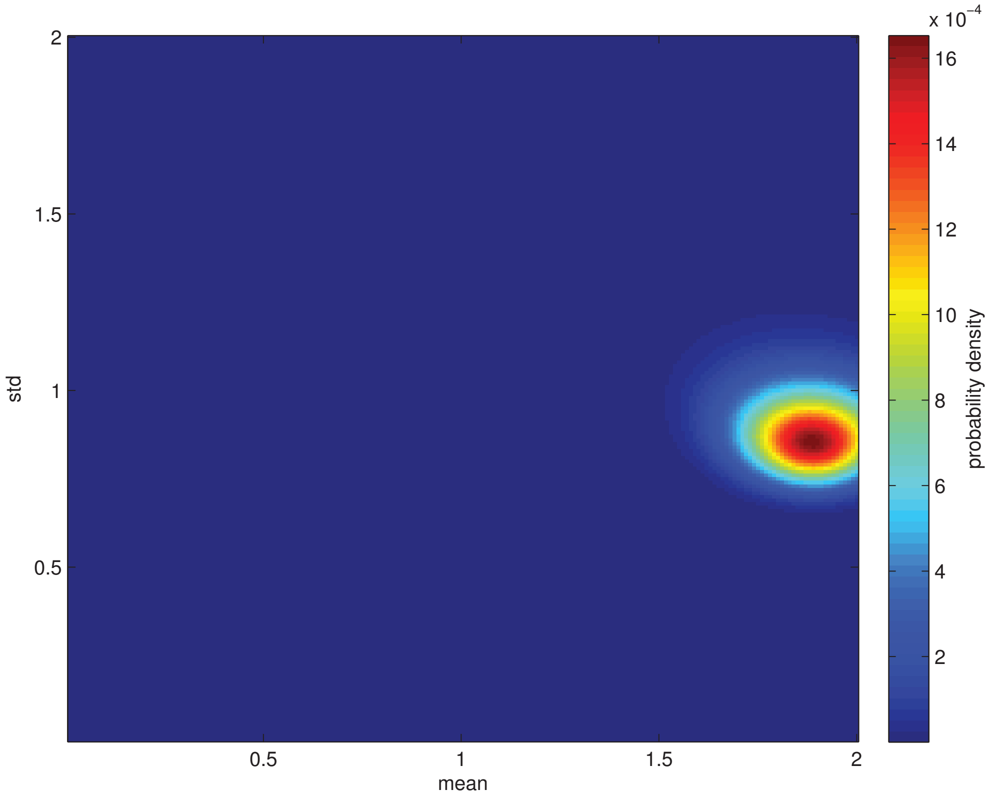

Figure D.4 – **Bayesian decision for identification of two different tasks.** The joint posterior distribution $\pi(\mu,\sigma|\mathbf{X})$ obtained by the Bayes' rule (after 50 samples) when the true mean (horizontal axis) is $\mu = 2$ and the true standard deviation (std, vertical axis) is $\sigma = 1$. Since the true mean $\mu = 2$ is relatively large, the Bayesian agent does not have a problem in estimating the true parameters. In fact it recognizes that there are two sufficiently different tasks.

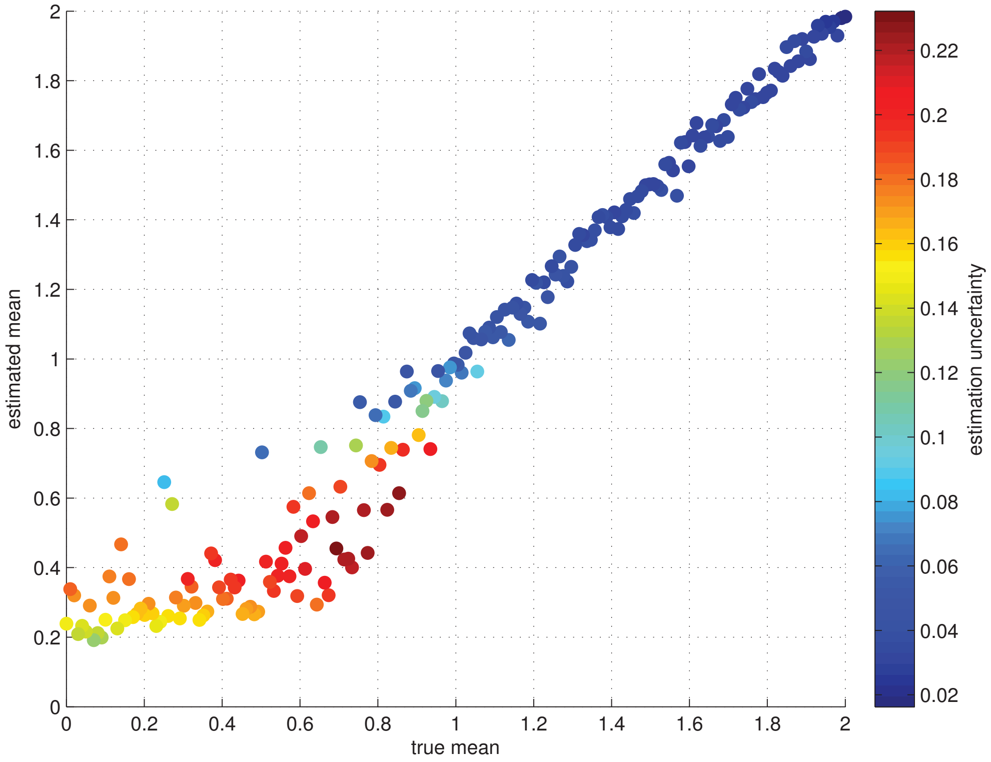

Figure D.5 – **Estimation uncertainty for different mean values.** The uncertainty (indicated by color intensity) about the estimated mean $\hat{\mu}$ (vertical axis) is smaller for the larger true means $\mu > 1$ (horizontal axis) than for smaller true means $\mu < 1$. This indicates that the estimation uncertainty decreases as the similarity between the two tasks decreases. The graph also shows that the accuracy of estimation is higher for larger μ than smaller μ (because the dots are less distributed around the diagonal for larger μ than smaller μ). The estimated mean $\hat{\mu}$ is calculated by averaging over the joint distribution (i.e., $\hat{\mu} = \int d\mu \left(\mu \int d\sigma \; \pi(\mu, \sigma|X)\right)$). The estimation uncertainty is calculated by the entropy of the joint distribution $\pi(\mu, \sigma|X)$.

Contributions

This section summarizes my contribution to each of the preceding chapters.

Chapter 1: I wrote the introduction. It was then edited by Wulfram Gerstner (WG) and Kerstin Preuschoff (KP). The idea of the multi-factor learning rules and how they can be linked to the activity of neuromodulatory system in the brain is proposed by WG.

Chapter 2: I introduced the confidence-corrected surprise and I did the mathematical derivations. The theory was then refined by WG. The consistency of the proposed model with the conceptual and computational characteristics of surprise has been checked and discussed with KP. The text was written by me in collaboration with WG and KP.

Chapter 3: I formulated the problem of learning through surprise minimization. I did the mathematical derivations as well as the simulations. I performed data analysis and produced the figures. I wrote the text in collaboration with WG and KP. The dynamic decision-making task was proposed by KP, and the high-dimensional maze-exploration task was proposed by WG, and was inspired by the previous work of Danilo Rezende. I derived and implemented the hierarchical Bayesian model in collaboration with Johanni Brea.

Our results in Chapters 2 and 3 were presented in multiple conferences on computational neuroscience (either as a poster or a talk), and are ready for submission to a journal under the name:

"Balancing New against Old Information: The Role of Surprise"
M. Faraji, K. Preuschoff, and W. Gerstner (arXiv 1606.05642 [stat.ML]).

Chapter 4: WG and I derived the learning rules. The text was written by me and was edited by WG and KP. I did the simulations and produced the figures. This work was inspired by a previous work of Danilo Rezende. The primary results of this chapter have been documented as an extended abstract for the International Conference on Reinforcement Learning and Decision Making (RLDM 2015). The work was then extended to incorporate the spiking neural network implementation. A manuscript is in preparation and will be submitted soon under the name:

“ A Biologically Plausible 3-Factor Learning Rule from Gradient Descent Optimization” M. Faraji, K. Preuschoff, and W. Gerstner.

Appendix A: This section was a review of the work on surprise by Gunther Palm. I had no personal contribution in the derivation of his theory, and I only summarized it as a ready reference.

Appendix B: I did the derivations and wrote the text, edited by KP. The results of this section were part of the aforementioned extended abstract at RLDM 2015.

Appendix C: The idea behind this model followed from a discussion with WG. Inspired by Grossberg’s theory of ART, WG and I developed the proposed model in this section. I performed the simulations and wrote the text. The results of this section were presented as a poster at Cosyne 2015.

Appendix D: I did the mathematical derivations and performed simulations. I presented my results as a talk in Jan. 2015 for a group of researchers from different Swiss universities working on a collaborative project (Sinergia).

Funding: During my PhD study, I have been financially supported by the European Research Council (ERC, grant agreement number 268689), and the Human Brain Project (HBP, grant agreement number 604102).

Bibliography

[Adams and MacKay, 2007] Adams, R. P. and MacKay, D. J. (2007). Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*.

[Alexander and Brown, 2011] Alexander, W. H. and Brown, J. W. (2011). Medial prefrontal cortex as an action-outcome predictor. *Nature neuroscience*, 14(10):1338–1344.

[Aston-Jones and Cohen, 2005] Aston-Jones, G. and Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annu. Rev. Neurosci.*, 28:403–450.

[Aston-Jones et al., 1997] Aston-Jones, G., Rajkowski, J., and Kubiak, P. (1997). Conditioned responses of monkey locus coeruleus neurons anticipate acquisition of discriminative behavior in a vigilance task. *Neuroscience*, 80(3):697–715.

[Baldi and Itti, 2010] Baldi, P. and Itti, L. (2010). Of bits and wows: a bayesian theory of surprise with applications to attention. *Neural Networks*, 23(5):649–666.

[Balleine and Dickinson, 1998] Balleine, B. W. and Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology*, 37(4):407–419.

[Balleine and O'Doherty, 2010] Balleine, B. W. and O'Doherty, J. P. (2010). Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology*, 35(1):48–69.

[Barber, 2012] Barber, D. (2012). *Bayesian reasoning and machine learning.* Cambridge University Press.

[Baxter et al., 2001] Baxter, J., Bartlett, P. L., and Weaver, L. (2001). Experiments with infinite-horizon, policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:351–381.

[Bear and Singer, 1986] Bear, M. F. and Singer, W. (1986). Modulation of visual cortical plasticity by acetylcholine and noradrenaline. *Nature*, 320:172–176.

[Beck et al., 2008] Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., Shadlen, M. N., Latham, P. E., and Pouget, A. (2008). Probabilistic population codes for bayesian decision making. *Neuron*, 60(6):1142–1152.

[Behrens et al., 2007] Behrens, T. E., Woolrich, M. W., Walton, M. E., and Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nature neuroscience*, 10(9):1214–1221.

[Bishop, 1994] Bishop, C. M. (1994). Novelty detection and neural network validation. In *Vision, Image and Signal Processing, IEE Proceedings-*, volume 141, pages 217–222. IET.

[Bouret and Sara, 2004] Bouret, S. and Sara, S. J. (2004). Reward expectation, orientation of attention and locus coeruleus-medial frontal cortex interplay during learning. *European Journal of Neuroscience*, 20(3):791–802.

[Bouret and Sara, 2005] Bouret, S. and Sara, S. J. (2005). Network reset: a simplified overarching theory of locus coeruleus noradrenaline function. *Trends in neurosciences*, 28(11):574–582.

[Brea et al., 2013] Brea, J., Senn, W., and Pfister, J.-P. (2013). Matching recall and storage in sequence learning with spiking neural networks. *The Journal of Neuroscience*, 33(23):9565–9575.

[Bush et al., 2000] Bush, G., Luu, P., and Posner, M. I. (2000). Cognitive and emotional influences in anterior cingulate cortex. *Trends in cognitive sciences*, 4(6):215–222.

[Carpenter and Grossberg, 1988] Carpenter, G. A. and Grossberg, S. (1988). The art of adaptive pattern recognition by a self-organizing neural network. *Computer*, 21(3):77–88.

[Carter et al., 1998] Carter, C. S., Braver, T. S., Barch, D. M., Botvinick, M. M., Noll, D., and Cohen, J. D. (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science*, 280(5364):747–749.

[Chaloner and Verdinelli, 1995] Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, pages 273–304.

[Clayton et al., 2004] Clayton, E. C., Rajkowski, J., Cohen, J. D., and Aston-Jones, G. (2004). Phasic activation of monkey locus ceruleus neurons by simple decisions in a forced-choice task. *The Journal of neuroscience*, 24(44):9914–9920.

[Cohen et al., 2007] Cohen, J. D., McClure, S. M., and Angela, J. Y. (2007). Should i stay or should i go? how the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 362(1481):933–942.

[Costa and Rudebeck, 2016] Costa, V. D. and Rudebeck, P. H. (2016). More than meets the eye: the relationship between pupil size and locus coeruleus activity. *Neuron*, 89(1):8–10.

[Couey et al., 2007] Couey, J. J., Meredith, R. M., Spijker, S., Poorthuis, R. B., Smit, A. B., Brussaard, A. B., and Mansvelder, H. D. (2007). Distributed network actions by nicotine increase the threshold for spike-timing-dependent plasticity in prefrontal cortex. *Neuron*, 54(1):73–87.

[Dayan et al., 2000] Dayan, P., Kakade, S., and Montague, P. R. (2000). Learning and selective attention. *nature neuroscience*, 3:1218–1223.

[Decker and McGaugh, 1991] Decker, M. W. and McGaugh, J. L. (1991). The role of interactions between the cholinergic system and other neuromodulatory systems in learing and memory. *Synapse*, 7(2):151–168.

[Donchin et al., 1978] Donchin, E., Ritter, W., McCallum, W. C., et al. (1978). Cognitive psychophysiology: The endogenous components of the erp. *Event-related brain potentials in man*, pages 349–411.

[Doya, 2000] Doya, K. (2000). Reinforcement learning in continuous time and space. *Neural computation*, 12(1):219–245.

[Ebstein et al., 1996] Ebstein, R. P., Novick, O., Umansky, R., Priel, B., Osher, Y., Blaine, D., Bennett, E. R., Nemanov, L., Katz, M., and Belmaker, R. H. (1996). Dopamine d4 receptor (d4dr) exon iii polymorphism associated with the human personality trait of novelty seeking. *Nature genetics*, 12(1):78–80.

[Fairhall et al., 2001] Fairhall, A. L., Lewen, G. D., Bialek, W., and van Steveninck, R. R. d. R. (2001). Efficiency and ambiguity in an adaptive neural code. *Nature*, 412(6849):787–792.

[Fletcher et al., 2001] Fletcher, P., Anderson, J., Shanks, D., Honey, R., Carpenter, T., Donovan, T., Papadakis, N., and Bullmore, E. (2001). Responses of human frontal cortex to surprising events are predicted by formal associative learning theory. *Nature neuroscience*, 4(10):1043–1048.

[Florian, 2007] Florian, R. V. (2007). Reinforcement learning through modulation of spike-timing-dependent synaptic plasticity. *Neural Computation*, 19(6):1468–1502.

[Frank et al., 2013] Frank, M., Leitner, J., Stollenga, M., Förster, A., and Schmidhuber, J. (2013). Curiosity driven reinforcement learning for motion planning on humanoids. *Frontiers in neurorobotics*, 7.

[Frémaux et al., 2010] Frémaux, N., Sprekeler, H., and Gerstner, W. (2010). Functional requirements for reward-modulated spike-timing-dependent plasticity. *The Journal of Neuroscience*, 30(40):13326–13337.

[Frémaux et al., 2013] Frémaux, N., Sprekeler, H., and Gerstner, W. (2013). Reinforcement learning using a continuous time actor-critic framework with spiking neurons. *PLoS Comput Biol*, 9(4):e1003024.

[Friston, 2010] Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138.

[Friston and Kiebel, 2009] Friston, K. and Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 364(1521):1211–1221.

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

[Gil et al., 1997] Gil, Z., Connors, B. W., and Amitai, Y. (1997). Differential regulation of neocortical synapses by neuromodulators and activity. *Neuron*, 19(3):679–686.

[Gillner and Mallot, 1998] Gillner, S. and Mallot, H. A. (1998). Navigation and acquisition of spatial knowledge in a virtual maze. *Journal of Cognitive Neuroscience*, 10(4):445–463.

[Gu, 2002] Gu, Q. (2002). Neuromodulatory transmitter systems in the cortex and their role in cortical plasticity. *Neuroscience*, 111(4):815–835.

[Hasselmo, 1999] Hasselmo, M. E. (1999). Neuromodulation: acetylcholine and memory consolidation. *Trends in cognitive sciences*, 3(9):351–359.

[Hasselmo et al., 1996] Hasselmo, M. E., Wyble, B. P., and Wallenstein, G. V. (1996). Encoding and retrieval of episodic memories: role of cholinergic and gabaergic modulation in the hippocampus. *Hippocampus*, 6(6):693–708.

[Hayden et al., 2011] Hayden, B. Y., Heilbronner, S. R., Pearson, J. M., and Platt, M. L. (2011). Surprise signals in anterior cingulate cortex: neuronal encoding of unsigned reward prediction errors driving adjustment in behavior. *The Journal of Neuroscience*, 31(11):4178–4187.

[Hebb, 2002] Hebb, D. O. (2002). *The organization of behavior: A neuropsychological theory*. Psychology Press.

[Hertz et al., 1991] Hertz, J., Krogh, A., and Palmer, R. G. (1991). *Introduction to the theory of neural computation*, volume 1. Basic Books.

[Herzog et al., 2012] Herzog, M. H., Aberg, K. C., Frémaux, N., Gerstner, W., and Sprekeler, H. (2012). Perceptual learning, roving and the unsupervised bias. *Vision research*, 61:95–99.

[Hess and Polt, 1960] Hess, E. H. and Polt, J. M. (1960). Pupil size as related to interest value of visual stimuli. *Science*, 132(3423):349–350.

[Holland, 1997] Holland, P. C. (1997). Brain mechanisms for changes in processing of conditioned stimuli in pavlovian conditioning: Implications for behavior theory. *Animal Learning & Behavior*, 25(4):373–399.

[Hsieh et al., 2000] Hsieh, C. Y., Cruikshank, S. J., and Metherate, R. (2000). Differential modulation of auditory thalamocortical and intracortical synaptic transmission by cholinergic agonist. *Brain research*, 880(1):51–64.

[Itti and Baldi, 2009] Itti, L. and Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision research*, 49(10):1295–1306.

[Itti and Baldi, 2005] Itti, L. and Baldi, P. F. (2005). Bayesian surprise attracts human attention. In *Advances in neural information processing systems*, pages 547–554.

[Janis and Mann, 1977] Janis, I. L. and Mann, L. (1977). *Decision making: A psychological analysis of conflict, choice, and commitment.* Free Press.

[Jaskowski et al., 1994] Jaskowski, P., Wauschkuhn, B., et al. (1994). Suspense and surprise: On the relationship between expectancies and p3. *Psychophysiology*, 31(4):359–369.

[Jepma and Nieuwenhuis, 2011] Jepma, M. and Nieuwenhuis, S. (2011). Pupil diameter predicts changes in the exploration–exploitation trade-off: evidence for the adaptive gain theory. *Journal of cognitive neuroscience*, 23(7):1587–1596.

[Jolivet et al., 2006] Jolivet, R., Rauch, A., Lüscher, H.-R., and Gerstner, W. (2006). Predicting spike timing of neocortical pyramidal neurons by simple threshold models. *Journal of computational neuroscience*, 21(1):35–49.

[Jones and Higgins, 1995] Jones, D. and Higgins, G. (1995). Effect of scopolamine on visual attention in rats. *Psychopharmacology*, 120(2):142–149.

[Kakade and Dayan, 2002] Kakade, S. and Dayan, P. (2002). Acquisition and extinction in autoshaping. *Psychological review*, 109(3):533.

[Kalat, 2012] Kalat, J. (2012). *Biological psychology.* Cengage Learning.

[Kappel et al., 2014] Kappel, D., Nessler, B., and Maass, W. (2014). Stdp installs in winner-take-all circuits an online approximation to hidden markov model learning. *PLoS Comput Biol*, 10(3):e1003511.

[Kennedy et al., 2003] Kennedy, H. J., Evans, M. G., Crawford, A. C., and Fettiplace, R. (2003). Fast adaptation of mechanoelectrical transducer channels in mammalian cochlear hair cells. *Nature neuroscience*, 6(8):832–836.

[Kimura et al., 1999] Kimura, F., Fukuda, M., and Tsumoto, T. (1999). Acetylcholine suppresses the spread of excitation in the visual cortex revealed by optical recording: possible differential effect depending on the source of input. *European Journal of Neuroscience*, 11(10):3597–3609.

[Knight et al., 1996] Knight, R. T. et al. (1996). Contribution of human hippocampal region to novelty detection. *Nature*, 383(6597):256–259.

[Kobayashi et al., 2000] Kobayashi, M., Imamura, K., Sugai, T., Onoda, N., Yamamoto, M., Komai, S., and Watanabe, Y. (2000). Selective suppression of horizontal propagation in rat visual cortex by norepinephrine. *European Journal of Neuroscience*, 12(1):264–272.

[Kolossa et al., 2015] Kolossa, A., Kopp, B., and Fingscheidt, T. (2015). A computational analysis of the neural bases of bayesian inference. *NeuroImage*, 106:222–237.

[Kolter and Ng, 2009] Kolter, J. Z. and Ng, A. Y. (2009). Near-bayesian exploration in polynomial time. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 513–520. ACM.

[Krugel et al., 2009] Krugel, L. K., Biele, G., Mohr, P. N., Li, S.-C., and Heekeren, H. R. (2009). Genetic variation in dopaminergic neuromodulation influences the ability to rapidly and flexibly adapt decisions. *Proceedings of the National Academy of Sciences*, 106(42):17951–17956.

[Kullback and Leibler, 1951] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

[Li et al., 2003] Li, S., Cullen, W. K., Anwyl, R., and Rowan, M. J. (2003). Dopamine-dependent facilitation of ltp induction in hippocampal ca1 by exposure to spatial novelty. *Nature neuroscience*, 6(5):526–531.

[Lin et al., 2003] Lin, Y.-W., Min, M.-Y., Chiu, T.-H., and Yang, H.-W. (2003). Enhancement of associative long-term potentiation by activation of β-adrenergic receptors at ca1 synapses in rat hippocampal slices. *The Journal of neuroscience*, 23(10):4173–4181.

[Little and Sommer, 2011] Little, D. Y. and Sommer, F. T. (2011). Learning in embodied action-perception loops through exploration. *arXiv preprint arXiv:1112.1125*.

[Lusher et al., 2001] Lusher, J., Chandler, C., and Ball, D. (2001). Dopamine d4 receptor gene (drd4) is associated with novelty seeking (ns) and substance abuse: the saga continues... *Molecular psychiatry*.

[Ma et al., 2006] Ma, W. J., Beck, J. M., Latham, P. E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature neuroscience*, 9(11):1432–1438.

[MacKay, 2003] MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

[Meyer, 1987] Meyer, J. (1987). Two-moment decision models and expected utility maximization. *The American Economic Review*, pages 421–430.

[Missonnier et al., 1999] Missonnier, P., Ragot, R., Derouesné, C., Guez, D., and Renault, B. (1999). Automatic attentional shifts induced by a noradrenergic drug in alzheimer's disease: evidence from evoked potentials. *International journal of psychophysiology*, 33(3):243–251.

[Mongillo and Deneve, 2008] Mongillo, G. and Deneve, S. (2008). Online learning with hidden markov models. *Neural computation*, 20(7):1706–1716.

[Morris, 1984] Morris, R. (1984). Developments of a water-maze procedure for studying spatial learning in the rat. *Journal of neuroscience methods*, 11(1):47–60.

[Müller et al., 1999] Müller, J. R., Metha, A. B., Krauskopf, J., and Lennie, P. (1999). Rapid adaptation in visual cortex to the structure of images. *Science*, 285(5432):1405–1408.

[Nassar et al., 2012] Nassar, M. R., Rumsey, K. M., Wilson, R. C., Parikh, K., Heasly, B., and Gold, J. I. (2012). Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature neuroscience*, 15(7):1040–1046.

[Nassar et al., 2010] Nassar, M. R., Wilson, R. C., Heasly, B., and Gold, J. I. (2010). An approximately bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *The Journal of Neuroscience*, 30(37):12366–12378.

[Nelken et al., 1999] Nelken, I., Rotman, Y., and Yosef, O. B. (1999). Responses of auditory-cortex neurons to structural features of natural sounds. *Nature*, 397(6715):154–157.

[Nelson et al., 2004] Nelson, A. L., Grant, E., Galeotti, J. M., and Rhody, S. (2004). Maze exploration behaviors using an integrated evolutionary robotics environment. *Robotics and Autonomous Systems*, 46(3):159–173.

[Nessler et al., 2013] Nessler, B., Pfeiffer, M., Buesing, L., and Maass, W. (2013). Bayesian computation emerges in generic cortical microcircuits through spike-timing-dependent plasticity. *PLoS Comput Biol*, 9(4):e1003037.

[Palm, 2012] Palm, G. (2012). *Novelty, information and surprise.* Springer.

[Payzan-LeNestour and Bossaerts, 2011] Payzan-LeNestour, E. and Bossaerts, P. (2011). Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS computational biology*, 7(1):e1001048.

[Pearce and Hall, 1980] Pearce, J. M. and Hall, G. (1980). A model for pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological review*, 87(6):532.

[Peters and Schaal, 2006] Peters, J. and Schaal, S. (2006). Policy gradient methods for robotics. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2219–2225. IEEE.

[Pfister et al., 2006] Pfister, J.-P., Toyoizumi, T., Barber, D., and Gerstner, W. (2006). Optimal spike-timing-dependent plasticity for precise action potential firing in supervised learning. *Neural computation*, 18(6):1318–1348.

[Pillow et al., 2008] Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E., and Simoncelli, E. P. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–999.

[Pineda et al., 1997] Pineda, J., Westerfield, M., Kronenberg, B., and Kubrin, J. (1997). Human and monkey p3-like responses in a mixed modality paradigm: effects of context and context-dependent noradrenergic influences. *International Journal of Psychophysiology*, 27(3):223–240.

[Posner and Fan, 2004] Posner, M. I. and Fan, J. (2004). Attention as an organ system. *Topics in integrative neuroscience: From cells to cognition*, pages 31–61.

[Pratt et al., 1995] Pratt, J. W., Raiffa, H., and Schlaifer, R. (1995). *Introduction to statistical decision theory*. MIT press.

[Preuschoff and Bossaerts, 2007] Preuschoff, K. and Bossaerts, P. (2007). Adding prediction risk to the theory of reward learning. *Annals of the New York Academy of Sciences*, 1104(1):135–146.

[Preuschoff et al., 2008] Preuschoff, K., Quartz, S. R., and Bossaerts, P. (2008). Human insula activation reflects risk prediction errors as well as risk. *The Journal of Neuroscience*, 28(11):2745–2752.

[Preuschoff et al., 2011] Preuschoff, K., t Hart, B. M., and Einhäuser, W. (2011). Pupil dilation signals surprise: evidence for noradrenaline's role in decision making. *Front Neurosci*, 5:115.

[Rajkowski et al., 1994] Rajkowski, J., Kubiak, P., and Aston-Jones, G. (1994). Locus coeruleus activity in monkey: phasic and tonic changes are associated with altered vigilance. *Brain research bulletin*, 35(5):607–616.

[Ranganath and Rainer, 2003] Ranganath, C. and Rainer, G. (2003). Neural mechanisms for detecting and remembering novel events. *Nature Reviews Neuroscience*, 4(3):193–202.

[Redgrave et al., 2010] Redgrave, P., Rodriguez, M., Smith, Y., Rodriguez-Oroz, M. C., Lehericy, S., Bergman, H., Agid, Y., DeLong, M. R., and Obeso, J. A. (2010). Goal-directed and habitual control in the basal ganglia: implications for parkinson's disease. *Nature Reviews Neuroscience*, 11(11):760–772.

[Reynolds and Wickens, 2002] Reynolds, J. N. and Wickens, J. R. (2002). Dopamine-dependent plasticity of corticostriatal synapses. *Neural Networks*, 15(4):507–521.

[Rezende and Gerstner, 2014] Rezende, D. J. and Gerstner, W. (2014). Stochastic variational learning in recurrent spiking networks. *Frontiers in computational neuroscience*, 8.

[Rezende et al., 2011] Rezende, D. J., Wierstra, D., and Gerstner, W. (2011). Variational learning for recurrent spiking networks. In *NIPS*, pages 136–144.

[Rüter et al., 2012] Rüter, J., Marcille, N., Sprekeler, H., Gerstner, W., and Herzog, M. H. (2012). Paradoxical evidence integration in rapid decision processes. *PLoS Comput Biol*, 8(2):e1002382.

[Sara and Segal, 1991] Sara, S. and Segal, M. (1991). Plasticity of sensory responses of locus coeruleus neurons in the behaving rat: implications for cognition. *Progress in brain research*, 88:571–585.

[Sara, 1998] Sara, S. J. (1998). Learning by neurones: role of attention, reinforcement and behaviour. *Comptes Rendus de l'Académie des Sciences-Series III-Sciences de la Vie*, 321(2):193–198.

[Sara, 2009] Sara, S. J. (2009). The locus coeruleus and noradrenergic modulation of cognition. *Nature reviews neuroscience*, 10(3):211–223.

[Sara et al., 1994] Sara, S. J., Vankov, A., and Hervé, A. (1994). Locus coeruleus-evoked responses in behaving rats: a clue to the role of noradrenaline in memory. *Brain research bulletin*, 35(5):457–465.

[Schölkopf and Smola, 2002] Schölkopf, B. and Smola, A. J. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond.* MIT press.

[Schultz, 2015] Schultz, W. (2015). Neuronal reward and decision signals: from theories to data. *Physiological reviews*, 95(3):853–951.

[Schultz, 2016] Schultz, W. (2016). Dopamine reward prediction-error signalling: a two-component response. *Nature Reviews Neuroscience*.

[Schultz et al., 1997] Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599.

[Schultz and Dickinson, 2000] Schultz, W. and Dickinson, A. (2000). Neuronal coding of prediction errors. *Annual review of neuroscience*, 23(1):473–500.

[Seol et al., 2007] Seol, G. H., Ziburkus, J., Huang, S., Song, L., Kim, I. T., Takamiya, K., Huganir, R. L., Lee, H.-K., and Kirkwood, A. (2007). Neuromodulators control the polarity of spike-timing-dependent synaptic plasticity. *Neuron*, 55(6):919–929.

[Settles, 2010] Settles, B. (2010). Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11.

[Shannon, 1948] Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

[Solomon et al., 2004] Solomon, S. G., Peirce, J. W., Dhruv, N. T., and Lennie, P. (2004). Profound contrast adaptation early in the visual pathway. *Neuron*, 42(1):155–162.

[Steinberg et al., 2013] Steinberg, E. E., Keiflin, R., Boivin, J. R., Witten, I. B., Deisseroth, K., and Janak, P. H. (2013). A causal link between prediction errors, dopamine neurons and learning. *Nature neuroscience*, 16(7):966–973.

[Stern et al., 1996] Stern, C. E., Corkin, S., González, R. G., Guimaraes, A. R., Baker, J. R., Jennings, P. J., Carr, C. A., Sugiura, R. M., Vedantham, V., and Rosen, B. R. (1996). The hippocampal formation participates in novel picture encoding: evidence from functional magnetic resonance imaging. *Proceedings of the National Academy of Sciences*, 93(16):8660–8665.

[Sun et al., 2011] Sun, Y., Gomez, F., and Schmidhuber, J. (2011). Planning to be surprised: Optimal bayesian exploration in dynamic environments. In *Artificial General Intelligence*, pages 41–51. Springer.

[Sutton and Barto, 1998a] Sutton, R. S. and Barto, A. G. (1998a). *Introduction to reinforcement learning.* MIT Press.

[Sutton and Barto, 1998b] Sutton, R. S. and Barto, A. G. (1998b). *Reinforcement learning: An introduction.* Cambridge Univ Press.

[Sutton et al., 1999] Sutton, R. S., McAllester, D. A., Singh, S. P., and Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. In *NIPS*, volume 99, pages 1057–1063.

[Takeuchi et al., 2016] Takeuchi, T., Duszkiewicz, A. J., Sonneborn, A., Spooner, P. A., Yamasaki, M., Watanabe, M., Smith, C. C., Fernández, G., Deisseroth, K., Greene, R. W., et al. (2016). Locus coeruleus and dopaminergic consolidation of everyday memory. *Nature*.

[Tartaglia et al., 2009] Tartaglia, E. M., Aberg, K. C., and Herzog, M. H. (2009). Perceptual learning and roving: Stimulus types and overlapping neural populations. *Vision research*, 49(11):1420–1427.

[Tenenbaum and Griffiths, 2001] Tenenbaum, J. B. and Griffiths, T. L. (2001). Structure learning in human causal induction. *Advances in neural information processing systems*, pages 59–65.

[Tribus, 1961] Tribus, M. (1961). Information theory as the basis for thermostatics and thermodynamics. *Journal of Applied Mechanics*, 28(1):1–8.

[Ulanovsky et al., 2003] Ulanovsky, N., Las, L., and Nelken, I. (2003). Processing of low-probability sounds by cortical neurons. *Nature neuroscience*, 6(4):391–398.

[Usher et al., 1999] Usher, M., Cohen, J. D., Servan-Schreiber, D., Rajkowski, J., and Aston-Jones, G. (1999). The role of locus coeruleus in the regulation of cognitive performance. *Science*, 283(5401):549–554.

[Vankov et al., 1995] Vankov, A., Hervé-Minvielle, A., and Sara, S. J. (1995). Response to novelty and its rapid habituation in locus coeruleus neurons of the freely exploring rat. *European Journal of Neuroscience*, 7(6):1180–1187.

[Vasilaki et al., 2009] Vasilaki, E., Frémaux, N., Urbanczik, R., Senn, W., and Gerstner, W. (2009). Spike-based reinforcement learning in continuous state and action space: when policy gradient methods fail. *PLoS Comput Biol*, 5(12):e1000586.

[Wallenstein et al., 1998] Wallenstein, G. V., Hasselmo, M. E., and Eichenbaum, H. (1998). The hippocampus as an associator of discontiguous events. *Trends in neurosciences*, 21(8):317–323.

[Williams, 1992] Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

[Wilson et al., 1992] Wilson, P. N., Boumphrey, P., and Pearce, J. M. (1992). Restoration of the orienting response to a light by a change in its predictive accuracy. *Quarterly Journal of Experimental Psychology: Section B*, 44(1):17–36.

[Wilson et al., 2013] Wilson, R. C., Nassar, M. R., and Gold, J. I. (2013). A mixture of delta-rules approximation to bayesian inference in change-point problems. *PLoS computational biology*, 9(7):e1003150.

[Xie and Seung, 2004] Xie, X. and Seung, H. S. (2004). Learning in neural networks by reinforcement of irregular spiking. *Physical Review E*, 69(4):041909.

[Yu and Dayan, 2005] Yu, A. J. and Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, 46(4):681–692.

Mohammadjavad Faraji

Personal Data

- **Contact address**: Av. de la Dôle 3, 1005 Lausanne, Switzerland
- **Tel**: (+41) 78 661 35 22, **E-mail**: mjf.faraji@gmail.com
- **DoB**: 29.05.1987, **Sex**: Male, **Marital status**: Married
- **Nationality:** Iran, with **Swiss resident permit B** (since 04.09.2009)

Education

- **École Polytechnique Fédérale de Lausanne**, Lausanne, Switzerland. Sept. 2011 - Dec. 2016
  Ph.D. in Computer and Communication Sciences, Computational Neuroscience.
- **École Polytechnique Fédérale de Lausanne**, Lausanne, Switzerland. Sept. 2009 - July 2011
  M.Sc. in Communication Systems, Signal and Image Processing, GPA: **5.48**/6.
- **Sharif University of Technology**, Tehran, Iran. Sept. 2005 - July 2009
  B.Sc. in Electrical Engineering, Communications, GPA: **17.72**/20.

Honors and Awards

- **Ranked 15th**/400,000 in the nationwide university entrance exam, Iran, 2005.
- Awarded **Excellence Bachelor Scholarship** from the National Elites Foundation of Iran, 2007.
- **Exceptional Talent** at Sharif University of Technology, Iran, 2009.
  (This title is awarded by being exempted from nationwide university entrance exam for master program.)
- Awarded **Master Research Scholarship** from the School of Computer and Communication Sciences, EPFL, 2010.

Research Experience

- **Laboratory of Computational Neuroscience (LCN)** Sept. 2011 - Present
  A research assistant and a PhD student under supervision of Prof. **Wulfram Gerstner** working on theoretical cornerstone of synaptic multi-factor learning rules, computational modelling of novelty and surprise signals and how they affect learning in machines (particularly, the brain).
- **Signal Processing Laboratory (LTS2)** July 2010 - July 2011
  A summer intern and a research assistant under supervision of Prof. **Pierre Vandergheynst** working on multiresolution analysis of graph-based data using wavelets on graphs via spectral graph theory and its applications in semi-supervised learning and transduction.
- **Audiovisual Communications Laboratory (LCAV)** Feb. 2010 - June 2010
  A research assistant under supervision of Prof. **Martin Vetterli** working on the impact of redundancy in pattern representation on classification accuracy.
- **Advanced Communications Research Institute (ACRI)** Feb. 2008 - July 2009
  A research assistant under supervision of Prof. **Farokh Marvasti** working on GWBE codes in overloaded CDMA systems and digital image watermarking emphasizing on reversibility and blindness using time-frequency transformations.

Publications

- **M. Faraji**, K. Preuschoff, W. Gerstner, “Balancing New Against Old Information: The Role of Surprise”, in preparation, arXive 1606.05642 [stat.ML].
- **M. Faraji**, K. Preuschoff, W. Gerstner, “A Biologically Plausible 3-Factor Learning Rule from Gradient Descent Optimization”, in preparation.
- D. I Shuman, **M. Faraji**, P. Vandergheynst, “A Multiscale Pyramid Transform for Graph Signals”, IEEE Transactions on Signal Processing, vol. 64, num. 8, p. 2119 - 2134, 2016.
- D. I Shuman, **M. Faraji**, P. Vandergheynst, “Semi-Supervised Learning with Spectral Graph Wavelets”, in Proceedings of the International Conference on Sampling Theory and Applications (SampTA), Singapore, May 2-6, 2011.
- P. Pad, **M. Faraji**, F. Marvasti, “Constructing and Decoding GWBE Codes Using Kronecker Products”, IEEE Communications Letters, vol. 14, num. 1, p. 1-3, 2010.

Abstracts

- **M. Faraji**, K. Preuschoff, W. Gerstner, “Surprise-Modulated Belief Update: How to Learn within Changing Environments? ”, Computational Neuroscience Meeting (CNS), Jeju Island, South Korea, July 2016.
- **M. Faraji**, K. Preuschoff, W. Gerstner, “A Novel Information Theoretic Measure of Surprise ”, International Conference on Mathematical Neuroscience (ICMNS), Antibes - Juan Les Pins, France, May 2016.
- **M. Faraji**, K. Preuschoff, W. Gerstner, “A Novel Measure of Surprise with Applications for Learning within Changing Environments ”, Computational and Systems Neuroscience (Cosyne), Salt Lake City, Utah, USA, March 2016.
- **M. Faraji**, K. Preuschoff, W. Gerstner, “Surprise Minimization as a Learning Strategy in Neural Networks ”, Computational Neuroscience (CNS), Prague, Czech Republic, July 2015.
- **M. Faraji**, K. Preuschoff, W. Gerstner, “A Biologically Plausible 3-factor Learning Rule for Expectation Maximization in Reinforcement Learning and Decision Making ”, in Proceedings of Reinforcement Learning and Decision Making (RLDM), Edmonton, Canada, June 2015.
- **M. Faraji**, K. Preuschoff, W. Gerstner, “Learning Associations With A Neurally-Computed Global Novelty Signal ”, Computational and Systems Neuroscience (Cosyne), Salt Lake City, Utah, USA, March 2015.
- M. Lehmann, A. Aivazidis, **M. Faraji**, K. Preuschoff, “Bayesian Filtering, Parallel Hypotheses and Uncertainty: a New, Combined Model for Human Learning ”, Computational and Systems Neuroscience (Cosyne), Salt Lake City, Utah, USA, March 2015.
- **M. Faraji**, K. Preuschoff, W. Gerstner, “Neuromodulation by Surprise: A Biologically Plausible Model of the Learning Rate Dynamics.”Computational Neuroscience (CNS), Quebec, Canada, July 2014.
- **M. Faraji**, K. Preuschoff, W. Gerstner, “A Biologically Plausible Model of the Learning Rate Dynamics.”, Gordon Research Conference on Neurobiology of Cognition (GRC), Sunday River Resort - Newry, Maine, USA, July 2014.
- **M. Faraji**, K. Preuschoff, W. Gerstner, “Surprise-based Learning: Neuromodulation by Surprise in Multi-Factor Learning Rules.”, Computational and Systems Neuroscience (Cosyne), Salt Lake City, Utah, USA, Feb. 2014.

Technical Reports

- “Learning with Surprise: Theory and Applications ”, **PhD Thesis**. Dec. 2016
- “Surprise in Decision Making and Interactive Learning”, Semester Project in PhD. July 2012
- “Novelty and Surprise in Reinforcement Learning”, Semester Project in PhD. Jan. 2012
- “A Laplacian Pyramid Scheme in Graph Signal Processing”, **Master Thesis**. June 2011
- “Frame-Based Classification”, Semester Project in Master. June 2010
- “Digital Image Watermarking Based on 2D Transforms ”, **Bachelor Thesis**. June 2009

Teaching Experience

- **Teaching Assistantship, EPFL and Sharif University, (14 courses).** Fall 2007 - Fall 2015

  General Physics II, Biological Modeling of Neural Networks, Embedded Systems, Statistical Signal and Data Processing Through Applications, Neural Dynamics of Single Neurons-EdX, Linear Algebra, Biological Modeling of Neural Networks, Unsupervised and Reinforcement Learning in Neural Networks, Probability and Statistics, Signal Processing for Communications, Basic Circuit Theory II, Multi-Variable Calculus, Course on MATLAB Programming, Basic Circuit Theory I.

Language Skills

- **Persian** : Mother tongue
- **English** : Fluent
- **French** : Intermediate (Level A2)
- **Arabic** : Beginner (Level A1)

Computer Skills

- **Programming Languages**: C, C++, Java, Python, HTML.
- **Operating Systems**: Linux, Mac OSx, Windows.
- **Neural Simulators**: Brian, NEST, Auryn.
- **Softwares**: MATLAB, Simulink, R, Julia, Microsoft Office, LaTeX.