A Paradigm for Autonomous Targeted Interaction with Biological Neuronal Networks

Dissertation
zur Erlangung des Doktorgrades der Technischen Fakultät der
Albert-Ludwigs-Universität Freiburg im Breisgau

Vorgelegt von
Sreedhar Saseendran Kumar

November 2017

Dekan

Prof. Dr. Oliver Paul

Referenten

Prof. Dr. Ulrich Egert

Tag der Abgabe: 23.11.2017

Tag der Prüfung: 12.04.2018

Sreedhar Saseendran Kumar
Laboratory for Biomicrotechnology
Dept. of Microsystems Engineering
Faculty of Engineering
Albert-Ludwigs-Universität Freiburg

Declaration of collaboration

The author (SSK[1,2]) hereby declares that the autonomous stimulation paradigm presented in this thesis was the result of collaborative work with researchers from the Machine Learning Lab, Department of Computer Science, University of Freiburg.

Some of the results presented here (particularly in Chapters 3 and 5) are expected to be included in the PhD thesis of Jan Wülfing (JW[3]: supervised by Martin Riedmiller[3,4] and Joschka Boedecker[3]), to be submitted in 2018.

Contributions to the project are summarized below:

- SSK and JW developed the idea, designed experiments and implemented the framework.
- Neuroscientific experiments were performed by SSK.
- Model development, data and quality analyses were performed by SSK.
- SSK and JW interpreted the results.
- Reinforcement learning (RL) experiments were performed by JW and SSK.
- RL components were conceived and implemented by JW.

[1] Laboratory of Biomicrotechnology, Department of Microsystems Engineering, University of Freiburg.
[2] Bernstein Center Freiburg, University of Freiburg
[3] Machine Learning Lab, Department of Computer Science, University of Freiburg.
[4] Current address: Google DeepMind, London, United Kingdom.

To the memory of my mother

Acknowledgements

My journey as a PhD student is about to end. In hindsight, the experience has been complex: humbling, but also rich and transformative. I have yet to reflect deeply on it. Thankfulness is often a clumsy but helpful first step in such retrospective expeditions. This note of gratitude is exactly that.

Thanks to Ulrich Egert – my teacher, mentor and supervisor – for his guidance, inspiration and encouragement. The enthusiasm he brought to the pursuit of scientific questions was infectious. Uli was always willing to give his time generously. He was candid in both critique and praise and settled for nothing short of my best effort. He was quick to raze a poorly formulated claim or summarily boot a shabby draft. But he always took the effort to elaborately unpack his reasoning. So though it hurt sometimes, I will cherish those lessons in clear scientific thinking. It was his singular confidence in me, even in difficult phases, that gave me the fortitude to persist.

In Jan Wülfing I found a friend, collaborator and 'co-conspirator' from the Machine Learning Lab. Together, we endured the thrills and chills of working with a complex experimental set up. Our frank discussions were instrumental in working our way forward. Thanks also to Martin and Joschka for their help, feedback and support.

I want to thank Samora for his generous support as a friend and mentor. Thanks for those discussions on matters scientific and otherwise. He was always there to lend a helping hand, an encouraging word or a suggestion. Thanks to Greg and Noah for helping me during my first days in the lab. Special thanks to Antje, Ehsan, Maximilian, Diego and the rest of 'Meagroup' for their support, feedback, discussions and laughs.

Thanks to our excellent lab technicians Ute Riede, Patrick Pauli and Alexander Giffey, without whose hard work in the lab our experiments could not have proceeded smoothly. My gratitude extends also to Marion, Janina, Birgit and Anne for their prompt administrative support.

Thanks also to my beloved partner Divya for sharing vicariously in my joy and despair. Your love and companionship continue to inspire me. I am forever grateful to my family. But for your love and sacrifices, I would not have been here today.

Contents

List of Figures

Abbreviations

ANN	artificial neural network
AP	Action potential
BNN	biological neuronal network
DBS	Deep brain stimulation
DIV	day(s) *in vitro*
IBI	inter-burst interval
ISI	inter-spike interval
IstimI	inter-stimulus interval
ITO	indium tin oxide
LSPI	Least-Squares Policy Iteration
MCS	Multi Channel Systems
MDP	Markov decision process
MEA	microelectrode array
MEM	minimum essential medium
MRAS	Model-reference adaptive system
NFQ	Neural Fitted Q-iteration
OCD	Obsessive-compulsive disorder
PD	Parkinson's disease
PEI	polyethyleneimine
PI	Proportional-Integral
PSTH	peri-stimulus time histogram
RE	recording electrode
RL	Reinforcement learning
RS	response strength
SB	spontaneous burst
SE	stimulation electrode
SISO	Single input-single output

Summary

Targeted interaction with networks in the brain holds immense therapeutic potential as a clinical technique for the treatment of neurological disorders like epilepsy and Parkinson's Disease. Aided by technological advances, electrical stimulation of the brain is increasingly explored as a therapeutic strategy. Most current approaches involve continuous stimulation with heuristically chosen parameter settings – the open loop paradigm – where the influence of ongoing neuronal activity on stimulation outcome is discounted. Consequently, they fail to achieve targeted interaction and may even contribute to stimulus-induced side-effects. To fully leverage electrical stimulation as a therapeutic, augmentative or research tool, there is a need to re-imagine it in a closed-loop setting, where the stimulator is able to respond appropriately to the current 'state' of the network.

Closed-loop control of neuronal network activity is, however, a highly non-trivial problem. Control strategies typically rely on a suitable dynamical model of the system. However, a broadly accepted mathematical theory for the collective ongoing activity of large neuronal populations is unavailable. The interaction between external stimuli and ongoing activity is also poorly understood and is characterized by irregular and complex dynamics (stochastic, non-linear and non-stationary). Further, in a clinical context, it is often not feasible to specify explicit responses or output patterns *a priori*, unlike in conventional regulation and tracking problems. Factors like these make it cumbersome to define tractable control problems for biological neuronal networks.

How can the closed-loop methodology be exploited to approach neurobiological control problems? In this thesis we propose a novel autonomous paradigm using methods of Reinforcement Learning (RL) to address this challenge. We implemented the technical framework of the paradigm and experimentally assessed its ability to optimize stimulation strategies and adapt to temporal heterogeneities in neuronal network activity.

RL emphasizes learning from direct interaction with the system without relying on exemplary supervision or complete models. To develop the concept and algorithms, we established the technical framework of the paradigm with generic neuronal networks *in vitro* as a model system. These networks retain the richness of cellular level processes and networked neurophysiological mechanisms characteristic of neuronal ensembles in the brain. They also preserve many of the challenges current RL algorithms would face in a neurotechnological context, such as high-dimensional state and action spaces and non-stationary activity dynamics. Concepts emerging from such co-adaptive interaction schemes are therefore expected to be generalizable. Major challenges for autonomous paradigms in the context of clinical neurostimulation were captured in toy control problems for the model system. With suitably defined target response features, we pursued RL methods to achieve them autonomously, assuming little in *a priori* information. The quality of the autonomously learned solutions were independently assessed.

In neurobiological control problems, desired response patterns are often stated in terms of overall system objectives, since their explicit values are seldom known *a priori*. Autonomous paradigms need to be able to achieve such objectives while optimizing stimulus settings with respect to factors like clinical outcomes or energy consumption. Further, they need to adapt to the non-stationary dynamics of ongoing activity. We translated such challenges into two separate problems formulated for our model system: an extremum seeking problem where the explicit target was unknown *a priori* and an adaptive clamping problem. While the former aimed at maximizing stimulation efficacy by interacting with the neuronal network, the latter involved clamping response strengths to predefined levels for long durations by adapting to the poorly understood temporal non-stationarities and fluctuations in network excitability, typically found in neuronal networks both *in vivo* and *in vitro*.

Using prior studies on the activity dynamics and their interaction with stimuli in generic neuronal networks *in vitro*, we approached the extremum seeking problem numerically and showed that it was well-defined and had a unique solution for each network. Therefore, predictions based on phenomenological models could be used to independently validate autonomously learned policies. They were found to be in very good agreement across networks. Our results offer the first proof-of-principle of an evaluable framework involving an autonomous controller optimizing its interaction with biological neuronal networks.

The adaptive clamping problem was approached using a failure driven algorithm development strategy. An extended history of ongoing and evoked activity was fed online to the RL controller to allow it to track temporal non-stationarities and fluctuations in the interaction model. We showed that long term stable goal-directed interaction was indeed feasible with such a paradigm. Response strengths consistently improved to levels closer to target and response failures were less likely after learning, compared to a random stimulation strategy. However, the temporal extent of the feedback signal and the nature of the interaction model were found to be factors crucial to the stability of the paradigm. Moreover, biological factors like stochastic shifts in the ongoing network mode, also contributed to oscillatory instabilities.

Taken together, in this thesis we present a novel autonomous RL based paradigm to approach closed-loop regulation and tracking problems in biological neuronal networks. Using control problems that share the underlying structure of targeted neurostimulation problems, we demonstrated how some of the challenges involved could be mitigated. Extending the framework may be a promising step forward for clinical applications involving neurostimulation.

Zusammenfassung

Die zielgerichtete Interaktion mit Netzwerken im Gehirn besitzt eine großes therapeutisches Potenzial zur Behandlung von neurologischen Erkrankungen wie Epilepsie und Parkinson. Im Zuge des technologischen Fortschritts wird dabei zunehmend die elektrische Stimulation des Gehirns als therapeutische Strategie erforscht. Die meisten aktuellen Ansätze basieren auf der kontinuierlichen Stimulation mit heuristisch ermittelten Parametern. In solchen sogenannten Open-Loop-Paradigmen (offener Regelkreis) wird der Einfluss fortlaufender neuronaler Aktivität und Plastizität auf das Stimulationsergebnis vernachlässigt. In der Folge, verfehlen solche Ansätze oft den erwünschten Stimulationseffekt und können gar zu unerwünschten Nebenwirkungen führen. Um die Möglichkeiten der elektrischen Nervenstimulation in der Therapie, Prothetik oder Forschung voll zu entfalten wäre es daher notwendig sie in einen geschlossenen Regelkreis (Closed-Loop-Paradigma) einzubetten. In einem solchen reagiert der Stimulator auf den aktuellen SZustandëines Nervennetzwerks und optimiert die Stimulationsparameter kontinuierlich nach, um das Stimulationsziel zu erreichen.

Die Regelung neuronaler Netzwerkaktivität mit Hilfe eines geschlossenen Regelkreises ist jenoch keinsfalls trivial. Regelungsstrategien basieren typischerweise auf einer geeigneten dynamischen Beschreibung des Systems. Eine allgemein akzeptierte mathematische Theorie zur fortlaufenden kollektiven Aktivität in grossen neuronalen Populationen ist bisher jedoch nicht vorhanden. Ebenso ist die Wechselwirkung zwischen äußeren Reizen und fortlaufender Aktivität wenig erforscht und zeichnet sich in der Regel durch unregelmäßige und komplexe (stochastische, nichtlineare und nicht-stationäre) Dynamiken aus. Im klinischen Kontext ist es zudem oft nicht möglich ein erwünschtes Antwort- oder Aktivitätsmuster a priori und explizit zu spezifizieren, anders als dies bei herkömmlichen Regelungs- und Trackingaufgaben der Fall ist. All diese Faktoren spiegeln die Schwierigkeit wider, lösbare Regelungsprobleme für biologische neuronale Netzwerke zu definieren.

In wie weit ist es jedoch überhaupt möglich, mit Hilfe der Closed-Loop-Methodik, neurobiologische Regelungsprobleme zu lösen? In vorliegender Arbeit wird ein neuartiges Paradigma für die autonome dynamische Kontrolle von biologischen neuronalen Netzwerken mit Methoden des Reinforcement-Learning (RL) vorgeschlagen. RL ermöglicht ein zielgerichtetes Lernen aus der direkten Interaktion mit einem System ohne dabei auf vollständige Modelle zurückzugreifen. Ein Regelkreis wurde entsprechend implementiert und die Effizienz der Stimulationsstrategie hinsichtlich ihrer Anpassungsfähigkeit an neuronale Nichtstationaritäten experimentell evaluiert.

Um grundlegende Konzepte und Algorithmen zu entwickeln, wählten wir generische neuronale Netzwerke in vitro als experimentelles Modellsystem. In diesen Netzwerken bleibt eine Fülle zellulärer Prozesse und neurophysiologischer Mechanismen erhalten, die für neuronale Netzwerke im Gehirn charakteristisch sind. Die Netzwerkdynamik *in vitro* bietet daher hinsichtlich ihrer Komplexität eine Problemstellung, die mit derjenigen im neurotechnologischen Kontext vergleichbar ist. Hierzu gehören hochdimensionale Zustands- und Aktionsräume, sowie nicht-stationäre Aktivitätsdynamiken. Wir gehen daher davon aus, dass die aus diesem Ansatz hervorgehenden Ansätze zur koadaptiven Interaktion verallgemeinerbar sind.

Grundlegende Herausforderungen an autonome Paradigmen, die im Kontext klinischer Neurostimulation bestehen, wurden auf vereinfachte Regelungsprobleme abgebildet. Dabei wurden RL-Methoden eingesetzt, um eine geeignet gewählte Regelgröße in der Reizantwort autonom und mit wenig a priori Information zu kontrollieren. Die Qualität der autonom erlernten Lösungen wurde im Anschluss unabhängig überprüft.

In neurobiologischen Regelungsproblemen wird das Stimulationsziel oft bezogen auf die Funktion des Systems in seiner Gesamtheit definiert, da die zugrundeliegenden Dynamiken meist unbekannt sind. Autonome Paradigmen müssen in der Lage sein ein solches Ziel zu erreichen, während Stimulationsparameter bezogen auf Faktoren wie physiologische Auswirkungen oder Energieverbrauch optimiert werden. Desweiteren müssen sie in der Lage sein, sich an die nichtstationäre Dynamik fortlaufender neuronaler Aktivität anzupassen.

Wir übersetzten solche Herausforderungen in zwei gesonderten Problemstellungen an unser Modelsystem: einer Maximierungsaufgabe, in der das erreichbare Stimulationsergebnis a priori unbekannt war, und einer Optimierungsaufgabe, in der die Regelgröße auf einem Sollwert gehalten werden musste. In der ersten Aufgabe war das Ziel, die Stimulationseffizienz durch Interaktion mit dem neuronalen Netzwerk zu maximieren, d.h. möglichst viele Aktionspotenziale auszulösen. In der zweiten Aufgabe sollte die Stärke der Reizanwort (wieder die Anzahl ausgelöster Aktionspotenziale) über längere Zeiträume konstant gehalten werden. Hierbei war eine stetige Anpassung an nicht vorhersehbaren Fluktuationen in der Netzwerkerregbarkeit, die für neuronale Netzwerke in vivo und in vitro typisch sind, notwendig.

Anhand von Vorstudien zur Aktivitätsdynamik in solchen Netzwerken und deren Wechselwirkung mit elektrischen Reizen war es uns möglich das Optimierungsproblem klar zu definieren und durch eine numerische Strategie für jedes Netzwerk eine eindeutige Lösung zu finden. Basierend auf phänomenologischen Modellen konnten zudem Vorhersagen gemacht werden, die eine unabhängige Validierung der erlernten Lösungen erlaubten. Die Ergebnisse zeigten über Netzwerke hinweg eine gute Übereinstimmung. Wir zeigen damit erstmalig, dass die Realisierung eines autonomen Controllers für ein Optimierungsproblem in biologischen neuronalen Netzwerken möglich ist.

Wir näherten uns dem Problem der adaptiven Regelung mit Hilfe einer fehlerbasierten Strategie zur Entwicklung von Algorithmen. Der RL-Controller hatte dabei Zugriff auf den Verlauf der spontanenen und evozierten Aktivität innerhalb eines zurückreichenden Zeitfensters, um zeitliche Nicht-stationaritäten und Fluktationen mit Hilfe des Interaktionsmodells zu verfolgen. Wir zeigten, dass es mit diesem Ansatz möglich ist, über längere Zeitraurme hinweg eine stabile Regulung zu realisieren. Im Vergleich zu einer zufälligen Stimulationsstrrategie verbesserten sich Reizantworten konsistent durch eine stetige Annäherung an den Sollwert. Ebenso wurde das Auftreten von Stimulationsfehlschlägen mit dem Lernen minimiert. Die Wahl des Interaktionsmodells sowie der des Zeitfensters für das Rückkopplungssignal waren für der Stabilität des Paradigmas dabei entscheidend. Intrinsische biologische Faktoren, wie stochastische Zustandswechsel im Netzwerkmodus, konnten oszillatorische Instabilitäten nach sich ziehen.

Zusammengefasst wird mit dieser Arbeit ein neuartiges autonomes RL-Paradigma zur Closed-Loop-Kontrolle von biologischen neuronale Netzen vogestellt. Anhand von Regelungsproblemen die im Wesentlichen ebenso bei der zielgerichteten Neurostimulation im klinischen Kontext auftreten, konnten wir Wege aufzeigen, um einige der Herausforderungen anzugehen. Die Erweiterung des vorgestellten Ansatzes könnte ein vielversprechender Schritt in der klinischen Anwendung der Neurostimulation sein.

Chapter 1

Introduction

Networks in the brain support the dynamic emergence of spatio-temporal electrochemical activity patterns that are thought to form the basis of its information processing capabilities (Bressler, 1995; Mesulam, 1998; McIntosh, 2000; Buzsaki, 2006). The ability to interact with them in a targeted manner – be it to induce or maintain desired patterns, or prevent undesired ones – will be of substantial clinical import. Additionally, controlling neuronal networks opens avenues to characterize them and address fundamental questions on the physiological basis of information processing in the brain (Wallach et al., 2011; Wallach, 2013; Xu and Barak, 2017). Aided by technological advances, neurostimulation has become the basis for a range of effective therapies that alleviate the symptoms of otherwise treatment-resistant neurological disorders. Most approaches rely on the continuous stimulation of a chosen anatomical target. With little dynamic possibilities to adjust stimulation parameters to the ongoing brain activity and the need for stimulation, such techniques fall short of achieving target-oriented interaction and may even contribute to the likelihood of stimulus-induced side-effects (Zhang et al., 2010; Baizabal-Carvallo et al., 2014; Pedrosa et al., 2014).

1.1 Targeted interaction with neuronal networks

Static stimuli to control dynamic brain networks?

Numerous efforts have been made since the late nineteenth century to interact with neuronal networks. Electrical excitability of the brain and the 'surprising' effects of external stimulation were keenly documented by researchers of the time (Fritsch and Hitzig, 1870; Gildenberg, 2005). In modern times, driven by technological advancements, electrical stimulation of the brain has evolved into a viable strategy to manage the symptoms of an increasing range of neurological disorders (Chen et al., 2012; Carron et al., 2013). It has been proven effective in treating movement disorders like essential tremor (Benabid et al., 1991; Benabid et al., 1993; Schuurman et al., 2000; Koller et al., 1999; Rehncrona et al., 2003), generalized dystonia (Coubes et al., 2000; Vidailhet et al., 2005; Isaias et al., 2009) and Parkinson's disease (PD) (Benabid et al., 1988; Benabid et al., 2009; Limousin et al., 1998; Krack et al., 2003; Deuschl et al., 2006; Bittar et al., 2005; Sarem-Aslani and Mullett, 2011; Kringelbach et al., 2007) and has shown considerable promise in treating refractory epilepsy (Fisher et al., 2010; Morrell, 2011; Fridley et al., 2012).

Stimulation of different anatomical targets are being tested for several psychiatric diseases such as Obsessive-compulsive disorder (OCD) (Nuttin et al., 1999; Mallet et al., 2008), Gilles de la Tourette syndrome (Fraint and Pal, 2015) and refractory depression (Mayberg et al., 2005; Holtzheimer and Mayberg, 2011) and has shown promise in treating other pharmaco-resistant brain pathologies like trigemino-dysautonomic headaches (e.g. refractory cluster headaches) (Matharu and Zrinzo, 2010; Franzini et al., 2003; Leone et al., 2006). Neurostimulation has also been pursued as a means to artificially inject information into neural circuits, e.g. towards neuroprosthetic devices capable of sensory feedback (Raspopovic et al., 2014). In most cases, the basis of the strategy is the repeated presentation of an invariant stimulus to the anatomical target of interest, regardless of the underlying neuronal activity, i.e. they are open-loop stimulation paradigms.

Limitations of open-loop neurostimulation

Several key factors severely limit the operational potential of open-loop stimulation paradigms. Neuronal activity patterns evoked as a response to stimuli applied to a few neurons are in fact the result of interaction of the stimulus with uncontrolled ongoing neuronal activity (Arieli et al., 1996; Hasenstaub et al., 2007). One of the first studies to show this was Arieli et al. (1996). The authors found that visually evoked activity levels in the cat visual cortex were dependent on the levels of ongoing activity. They showed that accounting for these fluctuations with a measure derived from the signal strength immediately preceding the stimulus reduced the variability of event related responses across repeated presentations of the physically identical stimulus (trials), and was thus presumably reflective of the instantaneous 'state' of the network. Similar findings were reported in other studies in *in vivo* model systems. In the cat striate cortex, single-cell visually evoked responses showed increased response spike count and reduced latency that was proportional to the increase in membrane potential prior to stimulation (Azouz and Gray, 1999). Kisley and Gerstein demonstrated that in the auditory cortex of anaesthetized rats, click-evoked field potentials and unit responses were modulated by rhythmic ongoing population bursting (Kisley and Gerstein, 1999). Petersen et al. showed that ongoing spontaneous activity in the form of locally synchronous fluctuations in membrane potentials regulate the amplitude and the time-dependent spread of sensory responses in the rodent barrel cortex. Studies *in vitro* have also reported response variations and their dependence on the context of ongoing activity within which the stimulus is presented (Shahaf et al., 2008; Weihberger et al., 2013).

Not only do these studies demonstrate the impact of ongoing activity on properties of evoked responses, but they also suggest that the nature of this interaction may not be a simple additive spillover of the pre-stimulus baseline signal into the response but may involve non-linear modulations. For instance, Kisley and Gerstein (1999) point out that apart from average responses, their variability too changed with changing levels of ongoing activity. In a more abstract and model oriented paradigm, Weihberger et al. (2013) probed the hidden state of an *in vitro* network by stimulating at various latencies relative to synchronous ongoing activity and observed that evoked response lengths fit a non-linear function (saturating exponential) of the latencies.

The complex nature of such network-stimulus interactions are, however, totally discounted in contemporary open-loop strategies. Stimulation parameters, once set, have limited scope for optimization. Where the beneficial effects of stimulation takes longer to appear, as in dystonia or OCD, parameter tuning may not be feasible at all. When stimulation immediately impacts symptoms like in Deep brain stimulation (DBS) for PD, arduous trial-and-error exploration of the parameter space by a trained clinician may be possible, but to a very limited extent.

Being ill-equipped to reform stimulus settings to reflect changes in ongoing activity patters arising from progression of the disorder, cognitive and motor load, mood and concurrent drug therapy, the strategy fails to guarantee safe and sustained outcomes (Hickey and Stacy, 2016; Zhang et al., 2010; Pedrosa et al., 2014). Studies on DBS for PD report that stimulation may eventually be without effect on some symptoms, worsen them, cause disabling side effects or become less efficient with time (Carron et al., 2013; Hariz et al., 1999). Moreover, frequent parameter adjustments have also been linked to improved DBS efficacy (Rosin et al., 2011; Moro et al., 2006). Collectively, these observations indicate that adapting stimulus parameters to the ongoing network state may be key to improving therapeutic outcomes.

Finally, in most clinical applications it would be crucial and desirable to optimize interactions, i.e. extremize a functional of the system's behaviour defined in terms of clinical outcomes, amount of stimulation or energy consumption. Such requirements are beyond the scope of an open-loop paradigm due to lack of feedback and immutability of stimulus settings. Using static settings is therefore an unsuitable strategy for goal-directed interaction with neuronal networks. Ideally, stimuli need to be responsive to the current 'state' of the network and choose settings optimal with respect to pre-defined cost functions as per application demands. Hence, to fully leverage electrical stimulation as a therapeutic, augmentative or research tool, there is a need to re-imagine it in a closed-loop framework.

1.2 Principles of closed-loop control

A closed-loop system is one in which two (or more) dynamical systems are connected together such that each system influences the other. Such systems are ubiquitous in both natural and engineered systems. Closed-loop systems, because of their strong mutual coupling, are very resilient to external disturbances and variations within each system (Åström and Murray, 2010). On the flip side, if applied incorrectly, the paradigm is susceptible to feedback instabilities causing oscillations or even runaway behaviour. It is therefore imperative that control strategies be developed based on a formal understanding of system principles that describe its behaviour.

A control strategy is an algorithm that exploits feedback signals to drive a designated measure (output or state) to a desired behaviour within the constraints of system stability (i.e. bounded disturbances give bounded errors). The central concept involves sensing the operation of the dynamic system, comparing it with desired behaviour, computing safe and appropriate corrective inputs and actuating the system to effect the desired change. The computational step is executed based on mathematical modelling techniques that capture the essential physics of the system and permit the exploration of possible behaviours (Åström and Kumar, 2014).

A control-oriented model is a concise mathematical representation of phenomena of interest in a physical, biological or information system that allows predictions to be made on how it will behave given a set of inputs. There are multiple approaches to modelling dynamical systems: e.g. white-box or internal approach, black-box or external approach and the grey box approach.

The internal approach attempts to describe the system from first principles, i.e. based on detailed mechanistic knowledge of the physical laws governing the system. It results in state models or white box models of the system – a set of coupled differential equations in a set of internal variables – state variables, along with algebraic equations that transform the state variables into system outputs. The 'state' of such a system is the minimal set of variables that fully characterizes the dynamic behaviour of the system and its response to any given set of inputs. The black-box approach, on the other hand, does not presume knowledge of the interior structure but places emphasis on the transfer-characteristics (input-output relationships) of the system. It is particularly useful to study linear time-invariant systems. The grey-box approach is an intermediate technique and employs whatever *a priori* information is available along with input-output data. Both black and grey-box modelling involves selection of model structures followed by estimation of model parameters based on observed data that are then validated (Murray et al., 2003; Le-Yi and Wen-Xiao, 2013; Åström and Kumar, 2014).

As a technology, automatic control based on a closed-loop paradigm has been a key enabler for engineered systems over the past two centuries. Despite tremendous advancements, particularly in the past 50 years, control theory has yet to make substantial contributions to bio-medical – particularly neurotechnological applications (Schiff, 2010). In the next section we examine a few closed-loop control strategies developed for engineered systems and the challenges involved in translating them to the neurotechnological context.

1.3 Control strategies for neurotechnological applications

1.3.1 Event-driven control

This simple closed-loop strategy involves monitoring a pre-determined indicator function in the measured activity (Fig 1.1A). It has recently been explored as a means of introducing feedback in a neurotechnological context. In these studies, a pre-defined indicator signal in the recorded activity triggered delivery of stimulation, though stimulus parameters themselves remained unchanged (Rosin et al., 2011; Little et al., 2013) . The strategy, while simple to implement, is dependent on prior knowledge of the system and is non-adaptive since signal modalities, thresholds and stimulus parameters have to be explicitly defined and are not mutually dependent. Systems of this type are prone to limit cycling – a behaviour in which the steady state error oscillates around zero – since the control action is inactive unless the triggering event occurs (Doebelin, 1985).

Rosin et al. (2011) tested an event-driven stimulation strategy for PD in the MPTP (1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine) primate model of PD. They reported that closed-loop stimulation delivered at the internal segment of the globus pallidum triggered by spikes detected in the primary motor cortex was more efficient in alleviating parkinsonian motor symptoms than the continuous open-loop stimulation paradigm. Following this proof-of-concept study, the first report of event-driven DBS on human patients with PD demonstrated an improvement in symptoms compared to standard DBS, with a simultaneous reduction in stimulation time (Little et al., 2013). In their study, stimuli were triggered when power in the beta frequency band (13-30 Hz) of local field potentials in the chosen reference structure (subthalamic nucleus) – which is known to correlate with motor impairment in PD – exceeded a pre-set threshold.

Studies have shown that tailoring the timing of stimulation to certain phases of tremor-related neuronal oscillations may help selectively decouple the tremor network (Cagnan et al., 2013). Based on this observation, closed-loop phase-locked DBS was recently tested in patients with pathological tremor (Cagnan et al., 2017). By locking stimuli to a particular phase of tremor, they were able to achieve up to 87% tremor suppression in selected patients with essential tremor, using less than half the energy of conventional high frequency stimulation. The real-time estimate of phase served as the 'event' in their study and was derived from peripheral inertial sensors attached to patients' limbs.

These demonstrations are indicative of the potential value addition that the feedback control framework could provide in a neurotherapeutic context. However, more sophisticated control strategies able to interact with the intricacies of network activity are desirable.

1.3.2 Model based control

First-principles modelling

When detailed mathematical models drawn from prior knowledge of the system are available, a model-based approach could help the design and analysis of feedback control strategies. Using control-oriented models specific to the phenomena of interest, appropriate inputs to actuate the system towards desired outputs can be computed. In engineered systems, detailed knowledge of physical laws governing the behaviour of the system to be controlled (plant) often helps devise analytically tractable models of adequate scope. On the other hand, a broadly accepted mathematical theory for the collective activity of populations of neurons does not exist yet (Breakspear, 2017). A mechanistic understanding of the set of internal state-variables or the requisite spatial and temporal scale of influence that uniquely represents such a system at any point in time is often lacking. In addition, the sheer scale of these systems and limitations in technological interfaces with these networks means only an extremely sparse sample of the system is experimentally observable. Consequently, defining the state of a biological network is troublesome.

It is often convenient to think of the veridical 'state 'of a network as an abstraction composed of two components: active and hidden states (Buonomano and Maass, 2009). The former consists of signals the experimenter is able to/chooses to observe at a given point in time while the latter subsumes the remaining signal components along with the rich repertoire of time-dependent properties of the cell and its constituent machinery – e.g. facilitation and depression, ion channel kinetics, Ca^{2+} concentrations in synaptic and cellular compartments – that are shaped by the history of activity/inputs at various time scales. The term 'state', as used in the neuroscientific literature is typically not a well-defined analytical construct. Rather, it is often used in a descriptive sense to denote clearly separable emergent classes/modes of activity or behaviour that are mostly identified in retrospect. For instance, the cortical 'state' is typically described as synchronized or desynchronized. In the former state, average population firing rates in a cortical column fluctuate strongly at a time-scale of $\approx$100 ms or slower whereas the latter is characterized by weak fluctuations in population firing rates (see Harris and Thiele (2011) for review). In another usage, single cells in the neocortex are described as being in distinct 'states' based on intracellular recordings of their membrane potential. The so called UP (resp. DOWN) state represents spontaneously occurring depolarizations (resp. hyperpolarizations) in the membrane potential and thereby the input/output transformations of the neuron (Destexhe et al., 2003).

Due to these limiting factors, control-oriented modelling from first principles i.e. an internal modelling approach, is infeasible for most applications involving biological neuronal networks.

System identification strategies

When the internal model of the plant is unavailable an alternative approach based on input-output based functional characterization could be useful. Such methods have been particularly helpful in the study of a special class of systems: linear time-invariant systems. Biological neuronal networks, being composed of non-linear elements that exhibit sharp threshold phenomena, do not belong to this class of systems. They exhibit non-linear, non-stationary, non-ergodic dynamic properties (Bassett et al., 2011; Bassett et al., 2013; Werner, 2011; Weihberger et al., 2013). Additionally, their dynamics change stochastically and across cognitive conditions or 'states' (Medaglia et al., 2015; Medaglia et al., 2017). These characteristics have been found to persist across scales of organization. Long-term measurements of spike time series from network-embedded single neurons *in vivo* revealed temporally complex dynamics and fractal character (Teich et al., 1997). Because of the inability to separate network effects from the contributions of single neurons to the observed complexity *in vivo*, studies on single neurons isolated from a network were undertaken *in vitro* (Gal et al., 2010). Temporal complexities including critical fluctuations, intermittency and scale-invariant rate statistics were found to persist. Complexities of this nature pose a considerable challenge to widely available system identification based control strategies.

To illustrate the problem, consider the case of focal stimulation in autonomously active generic neuronal networks. Such a stimulus typically elicits multi-phasic responses consisting of a) a fast excitatory component characterized by precise and reliable responses thought to originate from anti-dromic and mono-synaptic activation of neurons local to the stimulation site, b) a low activity period thought to be mediated by inhibitory neurons, c) a delayed excitatory component driven by recurrent poly-synaptic activation (Butovas and Schwarz, 2003; Rowland and Jaeger, 2008; Wagenaar et al., 2004). Across trials, however, the strength and duration of the response as well as the number and distribution of neurons involved is highly variable. This has been demonstrated both *in vivo* (Arieli et al., 1996; Kisley and Gerstein, 1999; Azouz and Gray, 1999; Petersen et al., 2003) and *in vitro* (Shahaf et al., 2008; Weihberger et al., 2013; Gal et al., 2010). As discussed in Section 1.1, studies have attributed this variability to non-linear interactions of the stimulus with various features of ongoing activity (Kisley and Gerstein, 1999; Petersen et al., 2003; He, 2013; Weihberger et al., 2013) and previous stimuli (Gal et al., 2010). Although these influences have been described in a general sense, a consolidated mathematical description reliably predictive of individual stimulation outcomes given ongoing activity remains elusive.

Experimental evidence that repetitive stimulation may lead to interaction between responses points to the non-stationary nature of the system. Studies *in vitro* suggest that this influence may affect responses across several scales of organization, from individual neurons (Gal et al., 2010) to networks (Weihberger et al., 2013; Eytan et al., 2003). Gal et al. (2010) exposed the dynamics of excitability of individual cultured cortical neurons over long time-scales while applying long series of stimulation pulses at various frequencies. They identify distinct phases of neuronal responses and critical stimulation frequencies at which phase transitions occur and that even a single neuron, when observed over a long enough duration, could exhibit a rich repertoire of response patterns. At the network level, Weihberger et al. report a loss of responsiveness and interaction between stimuli when focal stimuli are delivered at intervals less than 10 seconds.

Nevertheless, in applications like brain machine interfaces and prosthetic devices, where performance may be more important than whether the identified model best approximates the true neural system, the system identification approach may be of interest (Grosenick et al., 2015). Where the problem is low dimensional and input-output relations are mathematically well behaved, stable control strategies may be devised without directly modelling the system. This was demonstrated in a Single input-single output (SISO) case by Keren and Marom (2014) in an *in vitro* model system. A target response-probability was explicitly specified *a priori* and was achieved with a hand designed Proportional-Integral (PI) controller by adjusting stimulus amplitudes based on responses elicited by previous stimuli. Using offline studies, they were able to characterize the relationship between response probabilities and stimulation amplitudes and show that the former was monotonically influenced by the latter. By choosing operating conditions where this input-output dependence was nearly linear, they were able to demonstrate stable performance of the PI controller. However, reliable characterization of input-output relationships is not a typical starting point in clinical neuromodulation problems.

1.3.3 Adaptive control

The framework of adaptive control is a promising solution when the parameters of the plant's dynamic model change over long time scales (Landau et al., 2011). It has been applied to a variety of regulation and tracking problems where the objective is for the system output to follow a possibly dynamic set-point or reference trajectory. An adaptive control system can be thought of as having two loops: a normal feedback loop with the process and the controller, and a parameter adjustment loop. The parameter adjustment loop is often the slower loop (Åström and Wittenmark, 2008) that captures slow changes in the system state.

If measurable variables that correlate well with changes in the system dynamics are known *a priori*, or if these changes are predictable, a gain scheduling scheme could be applied. If a reference model connecting process output and input command signal is available, a Model-reference adaptive system (MRAS) may be used to adjust controller parameters such that the error between process and model output is small.

Most widely studied adaptive control schemes assume that the model of the true system is available and that feasible system identification techniques yielding implicit/explicit plant model to the adaptive algorithm exist. As discussed above, these are not typical starting points in neurotechnological applications.

Moreover, regulation and tracking problems, for which these methods were developed, assume prior knowledge of a reference trajectory. However, it is often the case in biology that one may not be able to specify *a priori* explicit desired output patterns, but only overall system objectives. Minimizing activity synchrony in PD and prolonging residence in states that minimize susceptibility to epileptic seizures are examples of such objectives. When the quantitative value of the target response cannot be clearly defined, is intrinsically variable, or where multiple interacting objectives have to be balanced, most adaptive control techniques cannot be directly applied.

1.3.4 Optimal control

In addition to being adaptive, it is desirable from a clinical point of view to optimize closed-loop interactions with respect to pre-defined cost functions. Considerable interest in the field over the past few decades has given rise to a general theory of optimal control. The approach relies on a mathematical model of the system to be controlled, a reference trajectory, a set of admissible inputs and appropriately defined performance measures or cost functions. The control problem is to determine the inputs that generate the desired output and in doing so, extremize the chosen performance measure (Athans and Falb, 2013).

Deriving optimal control policies is equivalent to solving the algebraic Riccati equation for linear systems with quadratic performance indices or the Hamilton-Jacobi-Bellman equation for non-linear systems (Zhao et al., 2015; Lewis et al., 2012; Al-Tamimi et al., 2007). However, a necessary pre-requisite for solving these equations is the availability of accurate models of the underlying system.

Adaptive schemes for optimal control are not theoretically well developed and are almost always based on indirect methods – i.e. control rule is recomputed from an estimate of the process model at each update. Indirect methods are, however, computationally infeasible and are not robust solutions particularly for non-linear optimization problems which involve solving non-linear partial difference equations that are often difficult to solve even if the model is precisely known (Zhao et al., 2015; Wang et al., 2012; Sutton et al., 1992; Sutton and Barto, 1998).

1.4 Our proposal: Reinforcement learning (RL) based autonomous control paradigm

An ideal neurotechnical intervention system would – given a goal – continuously monitor underlying network states, adapt stimulation settings accordingly and interact optimally with respect to pre-defined cost-functions while remaining goal directed. In addition, the system would need to work without reliable pre-defined models of the process dynamics and very often without explicit reference trajectories.

In recent times, RL – a computationally inexpensive solution based on methods of online dynamic programming – has been proposed as a tool capable of handling problems of this nature. It offers a direct approach to adaptive optimal control. Reinforcement learning involves an active decision making agent interacting with its environment and learning to map its perceived state to actions that in turn affect the state – so as to maximize a numerical reward signal. These methods are direct in that they are capable of adaptively converging to optimal control policies without using a model of the system or conducting an explicit search over possible sequences of future states and actions (Sutton et al., 1992; Sutton and Barto, 1998).

In this thesis we propose to establish and evaluate an RL based closed-loop paradigm to autonomously choose optimal control policies, without *a priori* models of the system or its interactions with electrical stimulation, in biological neuronal networks (Fig 1.1B).

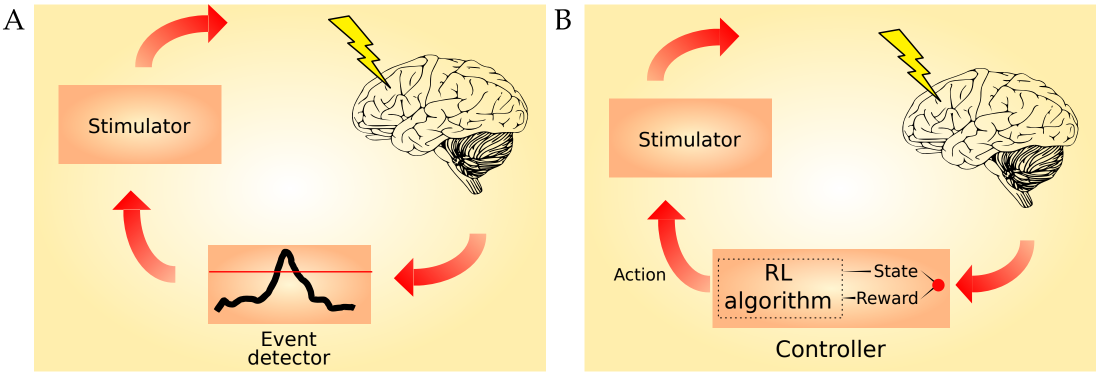

Figure 1.1. *(A) Schematic of a state-of-the-art event-driven scheme. Chosen signal features trigger stimuli when a pre-determined threshold value is crossed. Stimulus settings are typically fixed. (B) Schematic of the proposed RL based autonomous paradigm. Based on 'state' information obtained from activity in the network and an appropriate reward function, the algorithm learns to choose actions (stimulus settings) that maximize the reward signal.*

Challenges for the proposed RL based control paradigm

RL offers a computational approach to understand and automate goal-directed learning and decision-making. Unlike other computational approaches, it emphasizes learning of an agent from direct interaction with its environment, without relying on exemplary supervision or complete models of the environment (Sutton et al., 1992; Sutton and Barto, 1998). In this approach the tendency to choose an action is strengthened if it results in an improvement in the state of affairs as determined in an unambiguous manner. Extending this idea to allow action selection to depend on state information introduces aspects of feedback control, pattern recognition and associative learning.

The complex adaptive environment that plastic neuronal networks present poses several novel challenges for current RL algorithms. First, the task has to be defined in terms of a Markov decision process (MDP). The definition of the 'state' in this context should be problem relevant and should ideally capture sufficient information to solve the task at hand. Given the richness of activity patterns exhibited by neuronal networks, the state encoding is likely spatially and temporally distributed – i.e. high dimensional and history dependent – which not all algorithms gracefully scale to. Additionally, the action space to explore, i.e. stimulus modalities like location(s), timing, intensity, frequency, pulse-width etc. are also very large. To further add to problem complexity, every network being distinct in its properties and activity patterns limits generalizability across learning instances. Lastly, measured activity is noisy, highly variable and fluctuates over a range of time scales, all of which are likely to impact the efficiency of traditional RL algorithms.

1.5 Objectives of the thesis

The main goal of this thesis was to establish and evaluate paradigms and techniques to autonomously control activity dynamics in neuronal networks by electrical stimulation towards desired patterns of activity. We used techniques of reinforcement learning to identify optimal stimulation patterns and conditions in a novel autonomous closed-loop scheme. To develop the concepts and algorithms and explore their feasibility, we used generic neuronal networks *in vitro* on microelectrode arrays as a model system. Specifically, we address the following questions:

1. How to capture the interaction of ongoing network activity, electrical stimulation and evoked responses in a quantifiable 'state' to formulate a well-posed control problem for an RL controller?

2. How to develop algorithms to learn optimal stimulation settings?

3. How to develop algorithms to dynamically adapt to temporal inhomogeneities in network-stimulus interactions?

4. How to evaluate autonomously learned solutions for optimality, dynamic stability and robust performance across networks?

1.6 Generic neuronal networks as a model system

The dynamics of neuronal activity *in vivo* depend on a multitude of factors including and not limited to uncontrolled modulations by other brain regions and specificity of the anatomy and connectivity of the region of interest. Biological complexity in this scale makes it difficult to extract a consistent understanding of signal relationships between the network and an external stimulus – a crucial step in developing feedback control techniques (Kermany et al., 2010; Wallach, 2013). Therefore, in order to develop the concept and RL algorithms, it is imperative that we work with an appropriate model system that – while capturing the structure of the challenges and preserving the mechanistic basis thought to underlie these challenges – offers a stable, controlled and accessible environment.

Large-scale neuronal networks grown on substrate integrated microelectrode arrays are an appropriate model system in that they are easily accessible generic neuronal networks that can be maintained in a controlled environment, exhibit spontaneous activity known to influence the network's interaction with external stimuli, and are known to operate in distinct network modes across a wide-range of time scales (Wagenaar et al., 2006). Previous studies provide a partial understanding of the dynamics in such networks and of the rules governing their interaction with electrical stimuli (Weihberger et al., 2013), which allow to quantitatively evaluate solutions found by the RL based controller. Furthermore, these networks preserve many of the challenges current RL algorithms would face in a neurotechnological context such as high-dimensional state spaces, continuous action spaces and non-stationary activity dynamics, etc. Since these networks retain the richness of cellular level processes and networked neurophysiological mechanisms characteristic of neuronal ensembles, concepts emerging from such co-adaptive interaction schemes are expected to be generalizable.

1.7 Structure of the thesis

The thesis has been structured as follows.

Chapter 2 describes the model system, the technical details of the experimental set-up and the foundations of the closed-loop learning based system. We also present the details of experimental procedures. Finally, techniques employed for the analysis of multichannel extracellular spike data are discussed in detail.

In **Chapter 3**, we discuss the question of how to develop and evaluate techniques to autonomously optimize stimulation policies to interact in a goal-directed manner with biological neuronal networks. To this end, we identified a simple extremum seeking trade-off problem that arises as the result of the interplay of ongoing and evoked activity patterns in neuronal networks. Using prior studies on such networks and numerical methods, we show that the problem so formulated is well-defined, supports a unique solution throughout the span of parameter values observed experimentally, and had a relatively stationary optimum over the duration of the experiment. Using open loop stimulation, model based predictions could be made regarding the optimal stimulus timing. An RL controller was set to find this optimum autonomously by interacting with the network in a closed-loop setting. The quality of the learned solutions were positively evaluated based on open-loop predictions. These results demonstrate the capacity of RL based controllers to autonomously exploit underlying quantitative relationships to choose optimal actions without *a priori* knowledge of activity dynamics in the network.

Chapter 4 examines the long-term dynamics of network-stimulus interactions. We asked if signatures of temporal fluctuations and drifts in the quantitative rules governing stimulus-response relations were present in our model system. Residual analysis relative to known non-linear models quantifying such relations indeed pointed to the presence of slow fluctuations in the tens of minutes to hour scale modulating the interaction. Assigning an RL agent a static goal against such a non-stationary dynamic floor could offer a potent benchmark problem to develop algorithms for dynamic adaptive control with biological neuronal networks. We discuss the formulation of such a control problem, namely, to clamp response strengths over trials to a pre-defined value.

In **Chapter 5**, we extend the framework to focus on the control problem formulated in the previous chapter, i.e. dynamic optimization of learned policies in the absence of *a priori* information and models governing system dynamics. The technical framework developed in the previous chapter was augmented to handle high-dimensional and history-dependent states and actions. We attempted to dynamically clamp response strengths using enriched state information from the temporal history of activity. Our experiments brought to focus some of the important caveats attendant with such co-adaptive paradigms. Performance of the paradigm varied widely across networks. Dynamic instabilities were found to stochastically arise from network mode switches, sharp non-linearities in stimulus-response relations, high-dimensionality of the action set and learning delays. Although we were able to surmount these challenges in some of our networks, our results bring to the fore pertinent questions that need to be formally dealt with before such technologies can safely be translated to the clinical realm.

The main achievement of this thesis is that we translated some of the major challenges for neuromodulation in the clinical domain to an RL based autonomous stimulation paradigm in a controlled setting. We developed the technical foundations of such a closed-loop framework and demonstrated its viability using generalizable toy problems that captured some of the most challenging elements of a clinical neuromodulation problem. We tested the paradigm on a multi-objective optimization problem and a dynamic adaptive problem with little *a priori* knowledge made available to the agent. Using phenomenological models describing network-stimulus interactions, we showed that the quality of autonomously learned solutions could be quantitatively evaluated. The study also identified specific challenges that need formal treatment for the mature development of a safe and stable technical framework that could translate to effective clinical solutions in the future.

Chapter 2

Materials and Methods

[*] In order to develop our concepts and algorithms, we used networks of cortical neurons grown in cell cultures. While being generic and independent of specific functions and/or modalities, these networks preserve the biophysical complexity of the neuronal ensemble and relevant challenges an autonomous controller would face in a more complex context.

2.1 Cell culture preparation

For a detailed description of the steps involved in preparing and maintaining these cultures on microelectrode arrays, please see Kandler (2011), Okujeni et al. (2017) and Appendix A. In brief, frontal cortical tissue was dissected from newborn Wistar rats (obtained from the breeding facilities of the University of Freiburg) after decapitation, enzymatically dissociated, and cultured on polyethyleneimine (PEI)-coated microelectrode arrays (MEAs) from Multi Channel Systems (MCS), Reutlingen, Germany. The culture medium (1 mL) consisted of minimum essential medium (MEM) supplemented with 5% heat-inactivated horse serum, 0.5 mM L-glutamine, and 20 mM glucose (all compounds from Gibco Invitrogen, Life Technologies, Grand Island, NY). Cultures were stored in a humidified atmosphere at 37 °C and 5% CO_2 – 95% air. Medium was partially replaced twice a week. Neuronal density after the first day(s) *in vitro* (DIV)

[*] Portions of Sections 2.1, 2.2 and 2.4.1 of this chapter were included in the 'Materials and Methods' section of our published article: Kumar SS, Wülfing J, Okujeni S, Boedecker J, Riedmiller M, Egert U (2016) Autonomous optimization of targeted stimulation of neuronal networks. *PLoS Comput Biol* 12(8): e1005054.

ranged between 1500 and 4000 neurons/mm^2. The final density after 21 DIV settled at 1500–2000 neurons/mm^2, independent of the initial density. At the time of recording, network size thus amounted to $5 - 6 \times 10^5$ neurons. MEAs were occassionally plasma cleaned ($\approx$ 20 min, 40 kHz, 100 W; Femto plasma cleaner, model: Femto A, Diener Electronics, Nagold, Germany). Animal treatment was according to the Freiburg University (Freiburg, Germany) and German guidelines on the use of animals in research. The protocol was approved by the Regierungspräsidium Freiburg and the BioMed Zentrum, University Clinic Freiburg (permit nos. X-12/08D and X-15/01H). All cell cultures were prepared by Ute Riede and Samora Okujeni.

2.2 Electrophysiology

The activity of neuronal cultures was recorded and after 14 DIV to test for culture viability and general activity levels. Cultures with very few active channels and temporally sparse activity were discarded at this stage. Experiments were performed between 19 and 35 DIV. MEAs used for the experiments had 60 electrodes and a pitch of 500 µm (rectangular 6x10 grid). Electrodes were 30 µm in diameter, made out of titanium nitride (TiN) and initially had an impedance between 30 – 50 kΩ. Electrode tracks were made of Ti or or indium tin oxide (ITO) and insulated with silicon nitride (MEA Manual, 2017). Contact pads were made of either TiN or ITO. The MEA versions used were 60MEA500/30iR-Ti and 60MEA500/30iR-ITO. One channel was used as an internal reference and was thus unavailable for recording.

The MEA was placed into the pre-amplifier (MEA-1060-Inv-BC, MCS) connected to a filter amplifier (FA60S-BC, MCS) placed outside the incubator. The setting had a gain of 1100 and passband of 1 – 3500 Hz. Data was acquired and A/D converted with MC_card (MCS) at 25 kHz and a 12 bit resolution with a PC installation running Ubuntu Linux 10.04 with kernel version `2.6.32-38-generic-pae`. The kernel module for MC_Card under Linux was provided by Thomas DeMarse.

Neuronal activity was recorded inside a dry incubator (CB 210, Binder, Tuttlingen, Germany) at 37°C and 5% CO_2 – 95% air. Online spike detection was done with MEABench (versions 1.1.4 and 1.2.5) (Wagenaar et al., 2005) at six to eightfold root mean square noise level for spike threshold. It enabled fast and flexible online intervention by providing direct access to spike and raw data streams during recording. MEABench 1.1.4 was used for offline data acquisition while version 1.2.5 was used for online experiments.

Custom changes were made to MEABench 1.1.4 used for offline experiments. Spike cut-outs were extended to -2 ms to +3 ms relative to spike time, an absolute dead time of 2 ms after each detected spike was enforced and online access to spike and raw data was made possible. All recordings were performed with a gain factor of 2, i.e. a signal voltage range of $-683\,\mu\text{V}$. Signals were software-filtered between 150 – 2500 Hz. Online spike detection was performed on filtered data with a threshold crossing criterion of 6 – 8 fold estimated root mean square noise level. Spike cut-outs and time stamps were saved on hard disk for later analysis. Spike amplitudes varied strongly, e.g. due to varying position of neurons to the recording site and quality of the MEA electrodes. Typical spike amplitudes ranged between $\approx 20\,\mu\text{V}$ and $100\,\mu\text{V}$. Noise level under good recording conditions was less than $\pm 5\,\mu\text{V}$.

For closed-loop experiments, MEABench 1.2.5 was used. Its *Neurosock* server module was used to perform online analysis in another dedicated PC connected to the data acquisition via ethernet. The closed-loop architecture was realized by interfacing MEABench with the closed-loop control software, CLS^2 (Closed-loop Simulation System), an open source framework suitable for testing reinforcement learning controllers. It is developed and maintained by the Machine Learning Lab, University of Freiburg.

2.3 Electrical stimulation

A stimulus generator (STG2004, MCS) was used to deliver monophasic pulsatile stimuli to the MEA. Two stimulus inputs were available at the MEA pre-amplifier. Stimuli could be routed to an electrode of choice via custom written applications. Stimulation parameters were uploaded with custom-written C/C++ applications via the USB port. Setting and switching stimulation sites and grounding defective or noisy electrodes was achieved via the COM port. All custom-made C/C++ applications were obtained by modifying Visual Basic scripts provided by MCS. The Linux kernel module for

STG2004 was also provided by MCS.

Electrical stimuli consisted of single monophasic negative going pulse 400 μs wide and 0.5 – 1 V in amplitude. Monophasic negative pulse shapes were preferred to prevent oxidation of the TiN electrodes that would otherwise increase their impedance. The amplifier's integrated blanking circuit was used to suppress stimulus artefacts. This involved the following steps. All electrodes were briefly disconnected from the amplifier when the TTL pulse signalling a stimulus was received. The stimulation electrode alone was connected to the stimulus input between stimulus start and 400 μs after its end. An additional waiting period of 3 ms was set before the stimulation electrode would be disconnected from the stimulus input and the remaining electrodes reconnected to the amplifier. This reduced cross-talk among stimulation and recording electrodes.

To select MEA electrodes to serve as sites of stimulation (input) and of evaluation of responses (output) for closed-loop studies, we analysed spontaneous spike activity. As candidate stimulation electrodes (SEs) we selected sites that were more likely to participate early in spontaneous bursts (SBs) (Weihberger et al., 2013). This procedure identified the so-called "major burst leaders" (Eytan and Marom, 2006; Ham et al., 2008). Stimulation sites were chosen between rank 1 – 10. Periodic stimuli were delivered at these sites cyclically with an inter-stimulus interval (IstimI) of 10 s.

The responses were analysed to assess network responsiveness and channel-wise response properties. Response strength was typically defined as the count of spikes detected in a 500 ms post-stimulus window. Stimulus latencies – the periods of inactivity at a channel prior to each stimulus – were also extracted.

Peri-stimulus time histograms (PSTHs) and pairwise response strengths vs. latency relationships were used for qualitative assessments based on which a final SE and recording electrode (RE) pairs for closed-loop experiments were selected (see Section 2.5.3 and Appendix B.2 for further details). The chosen REs were typically found to be sites with responses consisting of both early ($\leq$ 15 ms) and late ($\geq$ 25 ms) components.

Apart from periodic stimulation for open-loop characterisation, a 'fixed latency' paradigm was also applied occasionally to control the timing of stimuli relative to ongoing SBs. In this case, stimulation was triggered whenever a pre-defined period elapsed without spikes detected at a selected RE. A minimal IstimI of 10 s was imposed to prevent network over-excitation.

2.4 Closed-loop experimental paradigm

Networks of dissociated neurons *in vitro* exhibit spontaneous activity characterized by intermittent network-wide synchronous bursts separated by periods of reduced activity. Inter-burst intervals (IBIs) in these networks fit an approximate lognormal distribution. Stimulating the network also evoked bursts of action potentials (response). The length of these responses at a chosen recording electrode can be modulated by the stimulus latencies relative to the SB at that channel. Their relationship was shown by (Weihberger et al., 2013) to fit a saturating exponential model. Further such stimulus-response relations when studied over long time-scales exhibited temporal fluctuations and drifts. Based on these observations, we formulated two closed-loop problems to use as a platform to explore the concepts and develop algorithms to realize an autonomous learning based controller.

Closed-loop learning problems were identified by SSK[5] in offline experiments. They were formalized into a framework approachable by reinforcement learning algorithms by JW[6]. The choice, implementation and extension of RL algorithms and parameter settings was done by JW[6]. The technical framework of the closed-loop system was designed by SSK[5] and JW[6]. The experiments were performed by SSK[5]. All analyses of the acquired data and quality assessments of the learned solutions were performed by SSK[5] and the results interpreted by SSK[5] and JW[6].

2.4.1 Optimization of Targeted Stimulation

In the first part (Chapter 3), we consider the problem of autonomously optimizing stimulation policies to interact in a goal-directed manner with biological neuronal

[5] Sreedhar Saseendran Kumar, Laboratory of Biomicrotechnology, IMTEK, University of Freiburg

[6] Jan Wülfing, Machine Learning Lab, University of Freiburg

networks.

Trade-off problem The optimization problem was defined as the following: what is the optimal stimulus latency relative to the end of the previous SB at a selected RE that maximizes response strengths evoked at that site per SB? To illustrate the problem, consider the following opposing strategies: A) Choosing a long latency: Based on the saturating recovery model, longer latencies would elicit longer responses. However, such a strategy would prove futile in the long run; long latencies are prone to interruptions by succeeding SBs and opportunities to stimulate will be forfeited. This would lower the count of evoked spikes per SB. B) Choosing short latencies would ensure that stimuli are delivered more often, but at the cost of evoking shorter responses. Optimization involves finding the trade-off between these opposing strategies. We asked that an RL based controller autonomously find the optimal time for stimulation to balance this trade-off for individual biological networks based only on the activity at the RE.

Experimental procedure Experiments were performed on 20 networks between DIV 19 and 35 ('network' denotes a culture at a specific point in time). Each experiment began with recording one hour of spontaneous activity, from which bursts were detected offline. A statistical model of SB occurrence was estimated by fitting a lognormal function to the IBI distribution to extract the location and scale parameters (μ and σ respectively). The task was formalized as a Markov Decision Process (MDP) to learn a controller with RL (see Appendix B for details). Each state of the MDP was defined as the period of latency subsequent to an ongoing SB event (see Fig 2.1A-B). Actions included the choice of wait or stimulate at each state. A detailed description is included in Appendix B.

Response strength Following the choice of SE and RE we studied the dependence of response strengths on periods of latencies preceding stimuli for each network. The number of spikes at the recording channel in a 500 ms window following a stimulus was typically defined as the response strength (RS). In few networks response bursts at the recording channel were found to extend beyond this window. In these cases, window widths were heuristically chosen to envelop most response bursts. The dependence of RS on stimulus latencies was modelled by a saturating exponential function of the form $A(1 - e^{-\lambda t}) + B$ (see Weihberger et al. (2013); Kumar et al. (2016)). The model captures the dynamics of recovery of post-burst network excitability with parameters A representing the gain of the network, B, the excitability threshold for SB termination and λ, the time constant of the recovery. The model fits to data was used to estimate these parameters corresponding to each network.

Closed-loop stimulation Closed-loop episodic learning sessions were performed using RE and SE positions identified as above. The controller was designed to learn in episodes that commenced at the termination of each SB (Fig 2.1A–B). The closed-loop architecture was realized by interfacing the data acquisition software (MEABench) with the closed-loop control software, CLS^2 (Fig 2.1C). Learning sessions proceeded in alternating training and testing rounds. During training, the controller explored the state-action space and learn a control law using the RL algorithm described in the following section, while during testing it always behaved optimally based on the knowledge hitherto acquired. Subsequent to the closed-loop session, spontaneous activity was recorded for one hour to check for non-stationarity in the IBI distribution.

Learning algorithm As a learning algorithm, we used online Q-learning with a tabular representation of the Q-function (Watkins and Dayan, 1992). Q-learning allowed us to learn a Q-function without having a model of the system dynamics, which in general is not available when dealing with biological systems. Secondly, since the state space for the control task at hand could be defined as a single discrete variable, a tabular representation of the Q-function was applicable, which is a prerequisite for guaranteed convergence (Watkins and Dayan, 1992). A tabular representation of the Q-function is a suitable choice as long as the biological system can be described by low-dimensional discretized states. The formulation of the online learning problem from the acquired data and the software tools used to interact with the network are summarized in Fig 2.1 and are described in Appendix B.

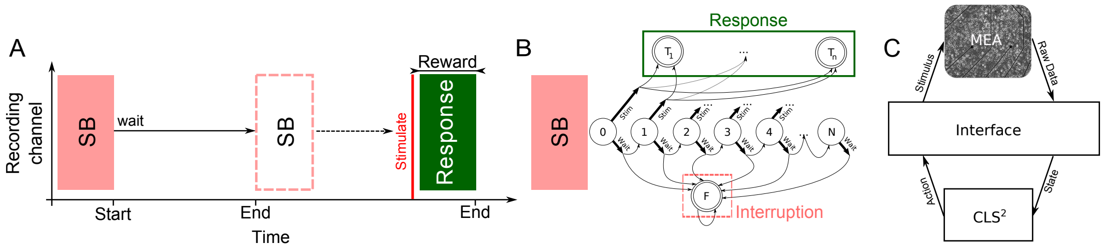

Figure 2.1. **Stimulation trials and the closed-loop architecture** *(A) A trial started with the end of an SB. The trial was terminated either by the next SB (dotted box) or a stimulation. In our paradigm, reward was defined as the number of spikes in the response. Interruptions by SBs led to neutral rewards (punishment). (B) The time within each trial was discretized into 0.5 s steps, corresponding to states* $1, \ldots, N$*. At each state, the controller could choose between two actions: to wait or to stimulate. A 'stimulate' action led to one of the terminal states* T_i*, with* i *indicating the strength of the response. Terminal state F was reached if the trial was interrupted by ongoing activity. (C) Schematic visualization of the closed-loop architecture. Figure reproduced from (Kumar et al., 2016).*

2.4.2 Autonomous adaptive control of responses

In Chapter 5, we consider the problem of clamping response strengths to pre-defined levels by continuously adapting to ongoing trends in stimulus response relations. The problem statement captures the essential structure of a dynamic adaptive problem in the context of a general neurotechnological application.

Experimental procedure Experiments were performed on 40 networks between DIV 19 and 35. Similar to the optimization problem, experiments began with one hour of spontaneous activity recording, and subsequent selection of candidate SEs. Open loop periodic stimuli were delivered at these sites and data analysed to identify viable pairs of SEs and REs (see Section 2.5.3 and Appendix B.2 for further details).

The recovery function at these REs were visually assessed and a target response strength defined such that it belonged to the first quartile of the ranked observed response strengths.

The task was formalized as a Markov Decision Process (MDP) to learn a controller with RL (see Appendix B for details). For this problem, each state of the MDP was defined as a vector of strengths (i.e. spike counts) of spontaneously occurring and evoked events prior to each trial. The number of events of each kind, included from the history of activity was typically set to 2 (4 and 5 were used in some cases). The time to the previous stimuli was added as an additional dimension in a few sessions (see

Fig B.1 for a schematic illustration). This high-dimensional state vector helped capture the network state in the temporal neighbourhood of the stimulus.

Compared to earlier experiments, a larger action set was available to the controller. From a visual assessment of recovery function at the RE, the approximate stimulus latency necessary to achieve target response strength was estimated. An interval spanning this latency was selected, discretized and provided to the controller. For details see Appendix B. Further, a choice of multiple stimulation sites was also provided. Typically, stimulus amplitudes were set to 700 mV. In a few sessions, a choice of multiple stimulus amplitudes were also included in the action set (see Appendix B for details).

During online learning, the controller explored various actions. Absolute errors between achieved and target response strengths were used as punishments for the controller. Thus minimizing punishments would improve the controller's performance toward target levels.

Learning algorithm We used online Q-learning as the algorithm for the paradigm. Unlike in the optimization problem, state-action space for the adaptive control problem had multiple dimensions. Thus a tabular representation of the Q-function was not advisable due to the memory demand growing exponentially with increasing dimensionality.

To cope with this problem, we switched to an approximate algorithm based on Least-Squares Policy Iteration (LSPI) that approximated the Q-function as a linear combination of the state features (Lagoudakis and Parr, 2003). To address challenges stemming from sharp non-linearities in stimulus-response relationships and higher action cardinalities, we switched to Neural Fitted Q-iteration (NFQ) algorithm (Riedmiller, 2005). In NFQ, the Q-function is approximated with an artificial neural network (ANN) to alleviate the memory problem. A further advantage of using ANNs for approximating the Q-function are their generalization capabilities; that is, we would expect approximate Q-value predictions by ANNs to give reasonable estimates not only for observed but also for unseen states. Details of the learning algorithm and the approximate methods we used are described in Appendix B.

2.5 Data analysis

Data analysis was performed with Matlab®(versions R2013b – R2017b, The MathWorks, Natick, MA, USA) with our own scripts. Data were loaded into Matlab using modified versions of scripts provided along with MEABench. Typically data consisted of extracellular spike timing and channel number. Spike sorting was not performed. When electrical stimulation was delivered, all spikes up to 2 ms after each stimulus were removed to avoid potential artefacts. All remaining spikes in open-loop random stimulation, closed-loop fixed-delay stimulation and spontaneous recordings were put through a spike cleaning procedure described below (see Section 2.5.1). Since such shape based processing was unavailable to the controller during online session, this process was skipped for the analysis of closed-loop learning experiments.

2.5.1 Spike cleaning procedure

A shape based custom post-hoc cleaning routine was written to isolate potentially spurious spikes (i.e. cases when spike detection threshold were crossed due to slower components or drifts in the extracellular voltage traces), from the train. The routine was run only on data from offline (open-loop) experiments. During online learning, since the spike train the controller received was not cleaned, the procedure was not applied also for post-hoc analyses.

Context traces (-2 ms to 3 ms with the peak at $t = 0$) of the recorded spikes were put through a series of tests. The basic rationale was to verify that voltage trace was sufficiently cuspidate. In case they were shallow, we tested if intermediate peaks existed within the shallow region. These would then be considered indicative of multi-unit activity and accepted as a spike. Otherwise, such shapes could indicate threshold crossings arising from slower components or drifts in the extracellular voltage traces. They could be of biological or technical origin, but should in either case not be considered a spike. The algorithm and test criteria are detailed in Algorithm 1. For spikes with positive peaks, the same algorithm with appropriately reversed operators was applied. A sample context that passed the tests is shown in Fig 2.2. Custom routines were obtained by modifying context cleaning scripts provided with MEABench 1.1.4.

Algorithm 1 Algorithm for cleaning spikes. Variables: $v(t)$ - voltage as a function of time, v_p - peak voltage (in μV), $zone_1 : 200\,\mu s < |t| \leq 1000\,\mu s$, $zone_2 : 500\,\mu s < |t| \leq 1000\,\mu s$, $t_{breach1}$ - time such that $v(t_{breach1} = 0.9 \cdot v_p)$, $t_{breach2}$ - time such that $v(t_{breach2} = 0.5 \cdot v_p)$

```
loop
    if $t_{breach1} \in zone_1$ then
        if $v(t) \leq (thresh - 4) \cdot rms_{noise} \quad \forall |t| \leq t_{breach1}$ then
            $reject \leftarrow spike$
        else
            $retain \leftarrow spike$
        end if
    else if $t_{breach2} \in zone_2$ then
        if $v(t) \leq (thresh - 1) \cdot rms_{noise} \quad \forall |t| \leq t_{breach2}$ then
            $reject \leftarrow spike$
        else
            $retain \leftarrow spike$
        end if
    end if
end loop
```

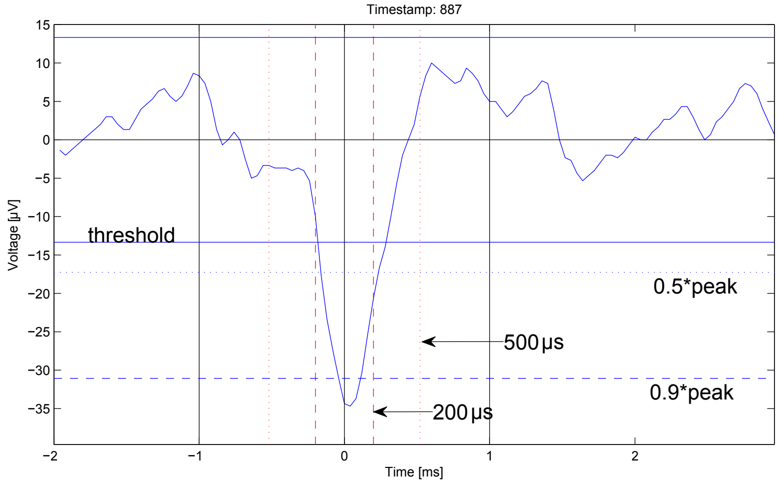

Figure 2.2. **An extracellular spike shown overlaid with the zone limits used for the testing procedure**. *Blue solid lines indicate the voltage threshold for spike detection in that channel (positive and negative). Black solid lines denote the outer limits of both zones 1 and 2. Red dashed lines (*200 µs*) and dotted lines (*500 µs*) indicate the inner limits of zones 1 and 2 respectively. Blue dashed and dotted lines indicate the 90% and 50% peak voltages. If a spike is not pointy enough and violates the zone 1 or 2 tests, crossing these secondary thresholds will prevent its rejection.*

To quantitatively validate the cleaning procedure, isolation scores were computed (Joshua et al., 2007). Isolation score measures the overlap between rejected (non-spike) and accepted spike clusters in high-dimensional space (script courtesy Ioannis Vlachos). To verify the effectiveness of the cleaning procedure, 10 batches of 1000 spikes each

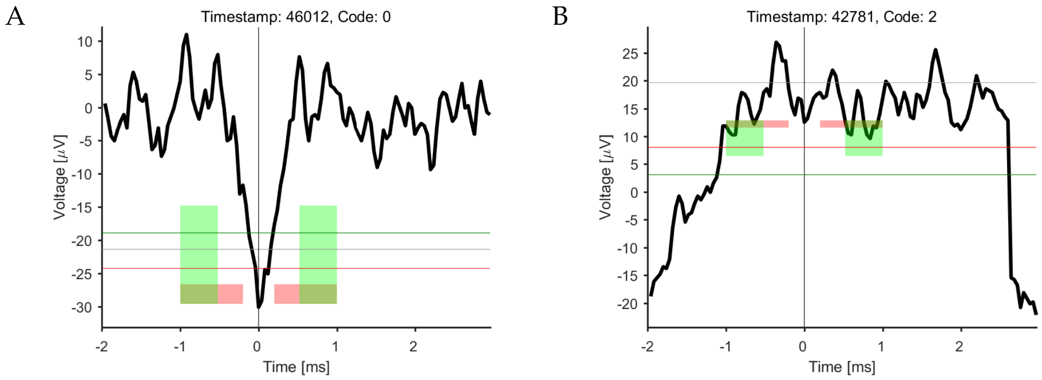

Figure 2.3. Example of a spike accepted (A) and rejected (B) by the algorithm. *Extracellular contexts are shown over a 5 ms period. Red and green patches denote zones 1 and 2 respectively. Lines show the respective secondary thresholds. Grey line denotes the threshold. The code numbers in the panel titles indicate processing status. Code 0 in panel (A) meant that the spike was accepted while code 2 meant that a zone 2 violation was the reason for rejecting the spike in panel (B).*

were randomly selected from the raw spike train to form the spike and noise clusters (before cleaning). This was repeated after the cleaning procedure, but the batches were selected from the accepted and rejected spike sets (after cleaning). The mean and standard deviation of the isolation scores over the 10 batches are shown in Fig 2.4 for an hour long spontaneous activity recording. In every case, isolation scores increased significantly suggesting that overall, the shapes of rejected spikes were significantly distinct from those of accepted spikes.

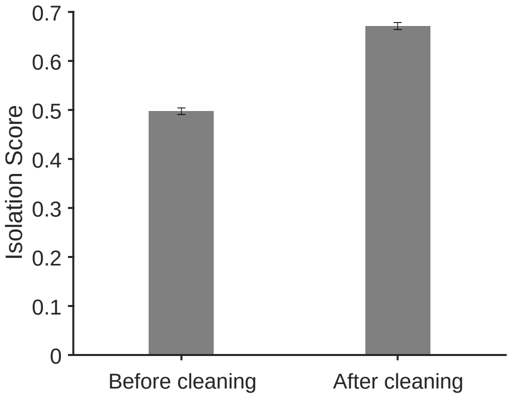

Figure 2.4. **Isolation scores before and after cleaning** *Isolation scores averaged over 10 batches of 1000 spikes each randomly selected from the spike train ($\mu \pm \sigma$). Spikes were first selected from the raw train to form the spike and noise clusters (before cleaning). After the cleaning procedure, batches were selected from the accepted and rejected spike sets (after cleaning). Increased isolation scores indicate that rejected spike shapes were significantly distinct from accepted ones.*

For data sets involving stimulation at multiple locations, switching an electrode from stimulation back to recording and another from recording to stimulating created highly synchronous artefacts and were identified and removed by an additional script. Periods where more than half the channels were synchronized within a 40 μs window were detected and the corresponding spikes removed (courtesy Samora Okujeni).

2.5.2 Spontaneous activity

Offline burst detection was performed for spontaneous data using the following algorithm: For spikes recorded from each electrode: a) inter-spike interval (ISI) had to be $\leq$ 100 ms, b) an interval $\leq$ 200 ms was allowed at the end of a burst and defined the minimal IBI, and c) the minimum number of spikes in a burst was set to three. Furthermore, at least three recording sites had to have burst onsets within 100 ms, and only one larger onset interval $\leq$ 200 ms was allowed (Weihberger et al., 2013).

For online burst detection at a single chosen channel, an individual ISI threshold was defined for each network based on spontaneous activity at the channel of interest prior to the closed-loop session (see Section 2.5.3 and Fig 2.5B). The ISI distribution of spontaneous activity was typically bimodal, with a strong first peak corresponding to ISI within SBs and a second peak for the intervals between bursts. The minimum between the intra- and inter-burst intervals was chosen as the threshold. The minimum number of spikes in a burst was set to three.

For the trade-off problem, parameters extracted from the fitting procedures were used to compute t^*, the open-loop parametric estimate of the optimal latency (Eq 3.1–3.5).

To compare the predicted and realized improvement in stimulation efficacy after learning we estimated the stimulation efficacy of a strategy using random stimulation latencies taken from the objective function as the baseline model in Fig 3.8E and Fig 3.9F. The efficacy of this strategy corresponds to the mean of the objective function of each network.

2.5.3 Response strength analysis

Response strength was defined as the count of spikes detected in a pre-defined post-stimulus window. This window width was typically set to 500 ms. However, if, on visual inspection, response bursts consistently lasted longer across trials, the window was heuristically increased to contain them.

PSTHs help analytically assess the efficacy of stimulation across trials. Channels that contained less than 1% of the overall spike count in the recording were excluded from further analysis. PSTHs were computed for all remaining SE-RE pairs. For each channel, spikes 50 ms before stimulus delivery till the outer edge of the response window were binned into 10 ms bins.

For all pairs mentioned above, recovery functions were fitted. The recovery function represents the mathematical relationship between response strengths and the period of prior latency at the given RE. This relationship was modelled by a saturating exponential function of the form $A(1 - e^{-\lambda t}) + B$ (see Weihberger et al. (2013); Kumar et al. (2016)).

For the fitting, outliers were visually identified and excluded. Note that statistically correct identification of outliers was not critical for the task. Latencies between 0.1 s and 10 s were binned in 0.33 s steps and mean response strengths in each bin calculated. Model fits between binned latencies and response strengths were made using the Levenberg-Marquardt non-linear least squares algorithm (Seber and Wild, 2003)(Fig 2.5A). The goodness of fit was assessed using the adjusted co-efficient of determination (R^2_{adj}), calculated as follows:

$$R^2_{adj} = 1 - \frac{SS_{res}/df_{res}}{SS_{tot}/df_{tot}} \tag{2.1}$$

If y_i is the mean response strength at latency bin i, and $\hat{y}_i$, the response strength predicted from the model, N, the number of stimuli, $\bar{y} = \frac{1}{N}\sum_i^N y_i$, the average of the y_is and p the number of model parameters (3 in our case), the expressions for SS_{res},

SS_{tot}, df_{res} and df_{tot} can be written out as follows:

$$SS_{res} = \sum_i (y_i - \hat{y})^2$$

$$SS_{tot} = \sum_i (y_i - \bar{y})^2$$

$$df_{res} = N - p - 1$$

$$df_{tot} = N - 1$$

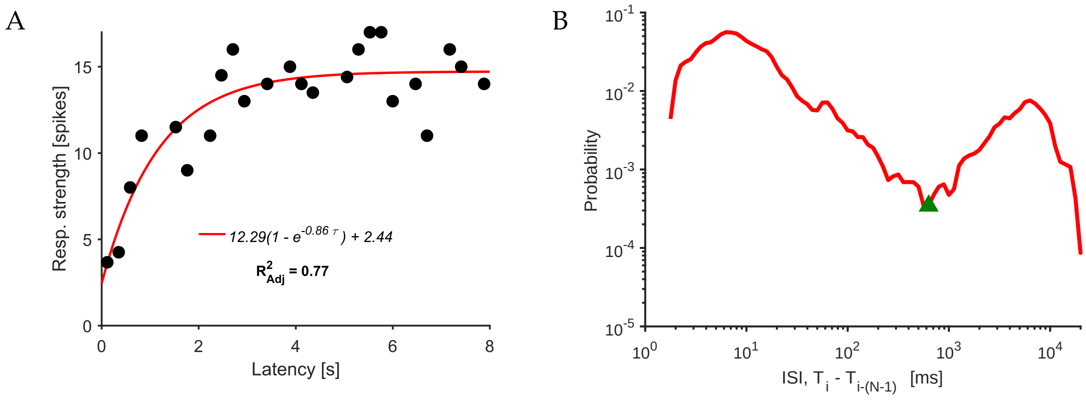

Figure 2.5. **Model fitting and choice of burst detection threshold** *(A) Model fit between binned latencies and bin-wise mean response strengths for an example SE-RE pair. Latencies were binned in steps of* 0.33 s. *The fitted model (red), the corresponding parameters and the goodness of fit (R^2_{adj}) are shown. (B) The distribution of ISIs for the RE in (A) during an hour long recording of spontaneous activity. The valley of the distribution* (630 ms; *green triangle) was chosen as the threshold for online burst detection at the RE.*

We identified a set of SE-RE pairs, where the model fits were good ($R^2_{adj} > 0.5$). A final selection was made after checking for consistent participation of each RE in SB events. When many feasible pairs were available, the choice was arbitrarily made. For the optimization experiments, a single SE-RE pair was chosen per session. For the adaptive control experiments, a single RE and multiple SEs were chosen. Additionally, a target response strength was also defined for these experiments (see Appendix B for details). Spontaneous spiking activity in the chosen RE was analysed to choose a suitable ISI threshold for online burst detection (see Section 2.5.2). From a visualization of the bimodal probability distribution of the ISIs at the RE in a spontaneous activity recording, the valley was chosen as the online burst detection threshold (Fig 2.5B).

Autonomous Optimization of Targeted Stimulation of Neuronal Networks

The following chapter (excluding Figs 3.4, 3.12–3.15 and Sections 3.4, 3.5) has been published as a peer-reviewed research article titled "Autonomous optimization of targeted stimulation of neuronal networks" in the journal *PLoS Computational Biology* (2016, Volume 12(8): e1005054). Authors: Sreedhar Saseendran Kumar (SSK), Jan Wülfing (JW), Samora Okujeni (SO), Joschka Boedecker (JB), Martin Riedmiller (MR) and Ulrich Egert (UE).

Contributions to the article:

- **SSK**: Developed the theoretical model; Performed experiments to characterize networks and predict optimal policies and closed-loop experiments; Data analyses; Quality analysis; Literature review.
- **SSK and JW**: Conceived the trade-off problem statement; Designed the experiment and implemented the technical framework for closed-loop experiments; Discussion of results and interpretations.
- **SSK and UE**: Discussed results and interpretations; Wrote manuscript.
- **JW**: Formalized the trade-off problem for RL algorithms; Choice, implementation and extension of RL algorithm and parameter settings.
- **JW and JB**: Wrote sub-section, "Reinforcement Learning"; Contributed to writing the "Discussion" section; Proof-read the manuscript.

Chapter 3

Autonomous Optimization of Targeted Stimulation of Neuronal Networks

In this chapter, we take up the question of how to develop and evaluate techniques to autonomously optimize stimulation policies to interact in a goal-directed manner with biological neuronal networks. The specific challenges we had to address here are: (1) how to capture the interaction of ongoing network activity, electrical stimulation and evoked responses in a quantifiable 'state' to formulate a well-posed control problem for a Reinforcement learning (RL) controller; (2) how to develop appropriate algorithms to learn optimal stimulation settings; (3) how to evaluate the quality of solutions autonomously learned?

To this end, we identified a toy extremum seeking trade-off problem that, as we show, emerges from the interplay of ongoing and stimulus evoked activity patterns in generic neuronal networks *in vitro.* Drawing on prior studies on such networks and numerical methods, we show that the problem so formulated is well-defined, supports a unique solution throughout the span of parameter values observed experimentally, and had a relatively stationary optimum over the experiment duration – a feature that ensured we could validate the learned solutions. Using open-loop stimulation on the same networks, we fit mathematical models to predict optimal stimulus policies for each case.

Armed with a well posed problem, we asked if and how RL could be employed to autonomously learn optimal control policies while interacting with such networks. Network-wise model based predictions of the optimal control policies were used to evaluate the quality of the solutions autonomously learned.

We first describe the properties of observed ongoing activity, the response of these networks to external stimuli and the trade-off problem emerging from their interplay. In the next sections we develop a numerical model of these activity components and study their interaction and make network-wise predictions of optimal stimulation policies. We then present results of the closed-loop learning sessions and how they compare with open-loop model based predictions. We conclude by exploring the consequences of bi-directionality in our paradigm and ask what a goal-directed machine reveals about the underlying biological network.

3.1 Properties of spontaneous network activity and response to electrical stimulation

Neuronal networks cultured on microelectrode arrays (MEAs) display spontaneous activity that consists of synchronized network-wide spontaneous bursts (SBs) separated by periods of inactivity. Burst-lengths ranged between hundreds of milliseconds to few seconds. SBs were detected using an algorithm that combined an inter-spike-interval threshold and the number of simultaneously active sites (Fig 3.1A). Inter-burst intervals (IBIs) were approximately lognormal distributed (Fig 3.1B). Fitting yielded the location and scale parameters (μ and σ) of the corresponding lognormal distribution. The cumulative of this distribution was used to estimate the probability of another SB occurring given the period of inactivity that elapsed – or what we term the 'probability of interruption' following an SB (Fig 3.1B, red line).

Stimulating a network at a channel evoked a burst of activity at others. For our experiments, we selected one stimulating and recording channel each. Weihberger et al. showed that the greater the duration of network inactivity, the longer the responses at a chosen site will be, according to a saturating exponential model. In order to verify this relationship and extract the parameters of the corresponding model, stimuli were delivered at random latencies relative to the previous SB (open-loop stimulation).

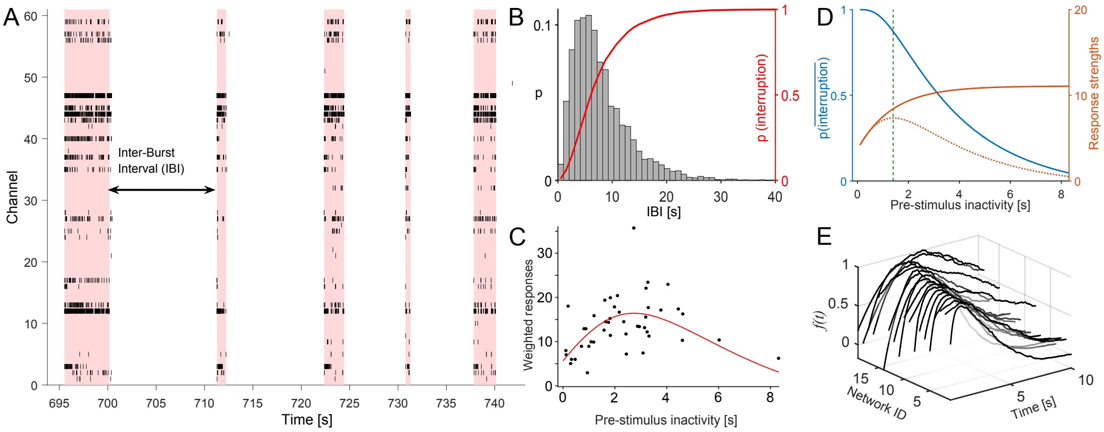

Figure 3.1. *Identification of network specific objective functions. (A) Networks of dissociated neurons in vitro exhibit activity characterized by intermittent network-wide SBs separated by periods of reduced activity (raster plot for 60 channels in a DIV 27 network). The shading marks the limits of individual SBs as detected by the burst-detection algorithm. (B) The distribution of IBIs is approximately lognormal. The histogram shows the IBI distribution for the network in (A). The cumulative of this distribution (red) is predictive of the probability of being interrupted by ongoing activity given the elapsed period of inactivity, i.e. the current state s_t. (C) Such a distribution was used to weight response strengths so that each dot represents the mean response strengths that can be evoked over a set of trials, including those that did not lead to stimulation, for a given stimulation latency. The fit predicts the objective function of the optimization problem. The example shows the data for the network shown in Fig 3.2C. The curve reveals a quasiconcave dependency, a unique global maximum and an optimal latency of ≈ 2.5 s in this network. (D) Fits to the probability of avoiding an interruption (blue), response strengths prediction (orange), and the resulting weighted response curve (orange, dotted) shown for another network. An optimal latency of ≈ 1.5 s emerges in this case. (E) All predicted objective functions for each of the 20 networks studied were quasiconcave and unique choices of optimal stimulus latencies were available. The objective functions were normalized to peak magnitude (Kumar et al., 2016).*

Fig 3.2A shows responses at the recording channel to 50 such trials in an example network. Responses typically consisted of an early ($\leq$ 15 ms post-stimulus) and late ($>$ 25 ms post-stimulus) component (Fig 3.2B). The early component, presumably reflecting responses to anti-dromic stimulation, was characterized by temporally precise and reliable responses while the late component, likely reflecting responses to orthodromic, transsynaptic activation, was both variable and unreliable (higher and lower probabilities respectively in Fig 3.2B).

A least square fit of the response strengths to a saturating exponential model with stimulus latency as the independent variable was carried out. The fitting function was of the form $A(1 - e^{-\lambda t}) + B$ (in red in Fig 3.2C). We then weighted all response strengths with the probability of being able to deliver a stimulus at the corresponding latencies, without being interrupted by ongoing activity. The weighted response strength curve (objective function) thus provides an estimate of the average number of response spikes

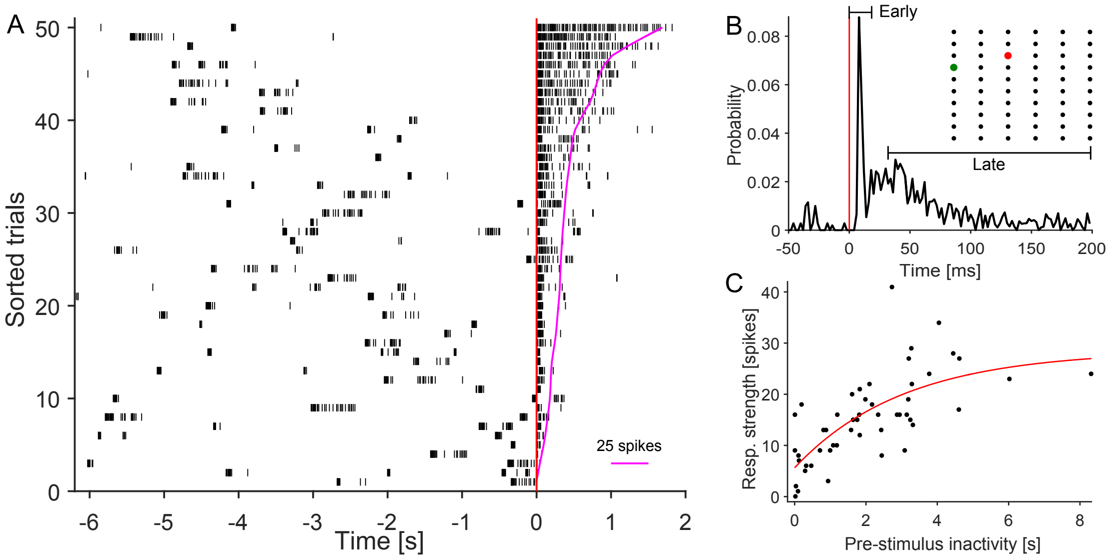

Figure 3.2. *Stimulating the network at an electrode evokes a burst of activity. Response strengths were dependent on the period of inactivity preceding the stimulus. (A) Raster plot shows responses at a recording channel (green in (B)-inset) to 50 stimuli delivered at a single channel (red in (B)-inset). Stimuli were delivered at random latencies relative to the previous SB. Trials were aligned to the time of stimulation (red line) and sorted by the count of spikes within the designated response window (see magenta overlay). A response window of 2 s was chosen for this network. The diagram exposes the relationship of response strengths to the period of prior inactivity. (B) Responses typically consisted of an early ($\leq$ 15 ms post-stimulus) and late ($\geq$ 25 ms post-stimulus) component. (inset) Schematic of a MEA with the chosen stimulation (red) and recording channel (green) marked. (C) The relationship between response strengths and periods of prior inactivity can be captured in a saturating exponential model of the form $A(1 - e^{-\lambda t}) + B$. Our model is similar to the response duration model proposed in Weihberger et al. (2013) (Figure modified from Kumar et al. (2016)).*

that can be evoked for each SB cycle, i.e. stimulation trial (Fig 3.1C, D). A solution that maximizes this estimate is therefore the optimal solution to the proposed trade-off problem, namely, to find the stimulus latency that maximizes the number of response spikes per SB.

We observed that a unique optimal stimulus latency exists for each of the 20 networks we studied (Fig 3.1E). The optimal latency emerges as the result of interaction of processes underlying ongoing and stimulus evoked activity dynamics of the network. Insights from previous studies (Weihberger et al., 2013) on such networks allowed us to parametrize each network based on recorded data. These parametric models were used to predict the network-specific optimal latencies offline before the RL controller was allowed to explore the problem in closed-loop setting. Note that the model was not used to instruct the controller.

3.2 Dependency of optimal stimulus latencies on properties of network activity

To understand the emergence of the optimal stimulus latencies from interacting biological processes and visualize the nature of the input-output relations and their relationship with the underlying parameter space, we considered simplified phenomenological models of each of the major contributing processes. Input, in the context of this problem refers to the period of inactivity/latency after which a stimulus is delivered, and output – the average number of response spikes evoked for every SB – the response feature of interest. The recovery of post-burst network excitability was modelled as a saturating exponential function (Eq 3.1). A statistical model of the temporal occurrence of SB events was considered (Eq 3.2). The corresponding model parameters were extracted from spontaneous and evoked activity recorded from each network.

$$R(t) = A(1 - e^{-\lambda\, t}) + B, \tag{3.1}$$

$$IBI(t) = \frac{1}{t\sigma\sqrt{2\pi}} e^{-\frac{(ln\, t - \mu)^2}{2\sigma^2}}, \tag{3.2}$$

$$\bar{I}(t) = 1 - \Phi\left(\frac{ln\, t - \mu}{\sigma}\right), \tag{3.3}$$

$$\text{where } \Phi(x) = \frac{1}{2\pi} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt,$$

$$f(t) = \bar{I}(t) \cdot R(t), \tag{3.4}$$

$$t^* = \underset{t}{\operatorname{argmax}}\, f(t). \tag{3.5}$$

$R(t)$ and $IBI(t)$ are the response strengths and the IBI respectively, modeled as a function of the period of inactivity, t (input). $\bar{I}(t)$ is the computed probability of avoiding an interruption, given a period of inactivity, t, and $f(t)$ the appropriately weighted response strength model – the objective function (the input-output relationship). $f(t)|t$ then gives the stimulus efficacy for repeated stimulation at latency t. The optimal latency t^* is the maximizer of this function.

In order to visualize the dependence of the input-output relations on the contributing parameters, we numerically computed objective functions and the corresponding t^*, while varying one or more parameters and holding the remaining constant. Initially, A was allowed to vary while parameters B, λ, μ, σ were held constant. Fig 3.3A–B shows the family of recovery functions considered and the corresponding family of objective functions. In general, all objective functions shared the property of being quasiconcave and permitted a unique maximum. These maxima (marked as dots) were the desired outputs and the corresponding stimulus latencies t^*, the desired optimal latency. The desired output – or equivalently the desired latency – increased non-linearly with A (Fig 3.3B). When, B was allowed to vary , holding parameters A, λ, μ, σ constant, the quasiconcavity of the family of objective functions was preserved (Fig 3.4A–B). Desired outputs increased and optimal latencies decreased with higher values of B (Fig 3.4C–D).

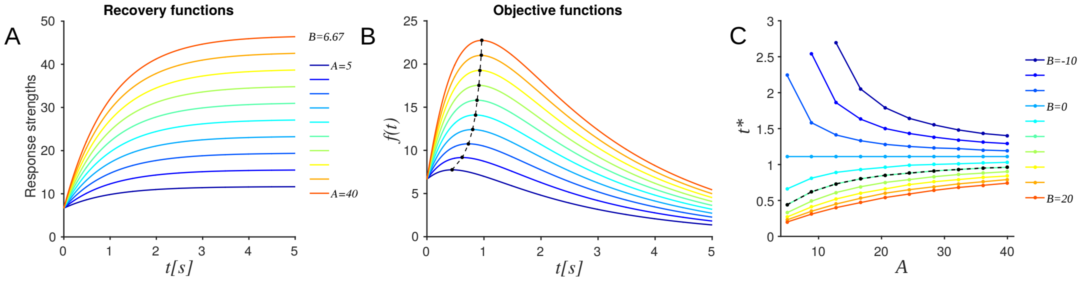

Figure 3.3. *Dependence of objective function and optimal latency on gain (parameter A). In all panels the parameters λ, μ and σ were set to 6.67, 1, 0.6 and 1, respectively. (A) Changes of response strength with the gain A of the response strength model within the range observed experimentally ($5 \le A \le 40$, B = 6.67; t: stimulus latency) (B) The optimal latencies t^* (dots), i.e. the maxima of the objective function $f(t)$ increased non-linearly with increasing A (dashed line). Colour code as in panel* A *(B = 6.67). (C) Changes of optimal timing t^* as a function of gain A and y-intercept B within the range observed experimentally (-10 $\le B \le$20). B influences the relationship of t^* with A and was trivial at B = 0. Black dots and dashed line indicate the case $B = 6.67$ shown in (B). Note that $A + B > 0$ was imposed to ensure that the maximal responses were strictly positive. (Kumar et al., 2016)*

Within the parameter range observed for A (15.5 ± 9.3) and B (4 ± 5.8) in our networks, the nature of the objective function family was preserved; a unique optimal latency existed, and monotonically increased or decreased non-linearly with A, depending on the value of B (Fig 3.3C). Fig 3.5A summarizes the dependence of t^* on the $A - B$ plane. Each colour coded plane corresponds to a different value of the time constant λ. λ was allowed to vary in the range observed experimentally ($0.2 \le \lambda \le 1.2$).

Next, we varied the location and scale parameters, μ and σ respectively – see Eq 3.2 of the IBI distribution. The corresponding input-output relations were still

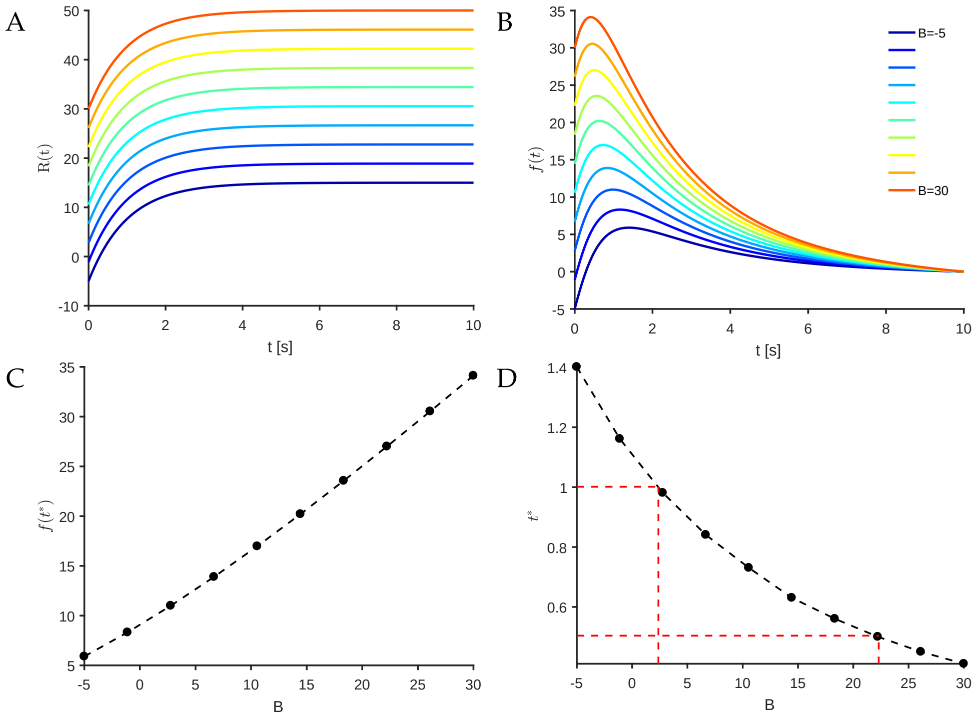

Figure 3.4. *Dependence of objective function and optimal latency on offset (parameter B). (A) Changes in the recovery function model $R(t)$ with the y-intercept B varying from -5 to 30. In these panels, parameters A, λ, μ and σ were set to 20, 1, 0.6 and 1, respectively. t: stimulus latency). (B) Shows the corresponding objective functions; $f(t)$: stimulus efficacy (C) $f(t^*)$, the optimal value of stimulus efficacy, increases with parameter B as seen in panel (B). (D) The corresponding optimal stimulus latency decreases with increasing values of B. This trend is also captured in a vertical slice of data points in Fig 3.3C. A Δt of 0.5 s was chosen as the discrete time step for our closed-loop learning experiments. Red dashed lines indicate the change in B (abscissae) that would be necessary for optimal stimulus latencies to change by 0.5 s (ordinates). Temporal modulations in parameter values though present in our networks are unlikely to exceed this range. Thus, given our discretization the trade-off problem is reasonably stationary.*

quasiconcave, thus ensuring the existence of a unique maximum. Optimal latency depended almost linearly on μ (Fig 3.5D). Fig 3.5E illustrates how optimal latency is modulated in the $A - B - \mu$ space for $\lambda = 1$. The scale parameter σ, however, had no significant effects on the shapes of the objective functions and hence the corresponding optimal latencies (Fig 3.6B). Based on these analyses we predicted optimal stimulus latencies for each network.

3.3 RL based strategy to learn optimal latencies

Closed-loop learning sessions consisted of alternating pairs of training and testing rounds. During training rounds, the controller explored the state-action space and

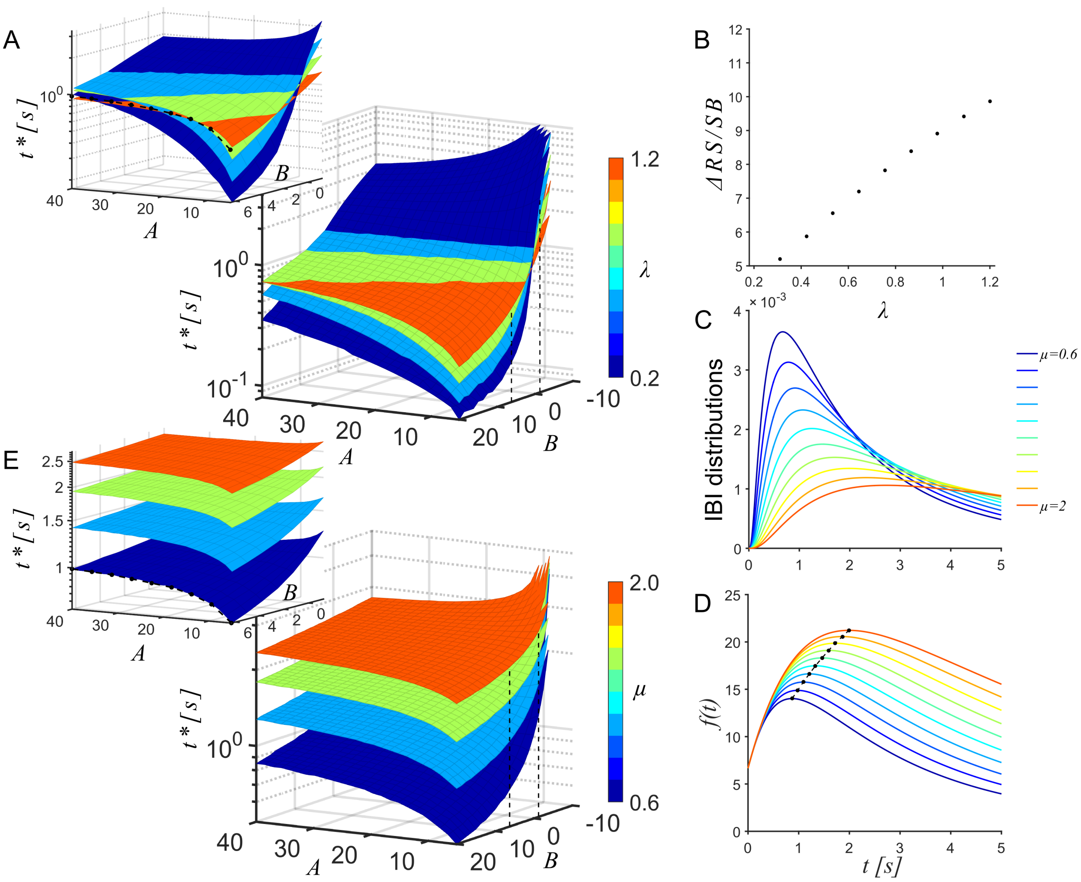

Figure 3.5. *Dependence of the optimal latency on properties of the network's activity dynamics. (A) Dependence of the optimal stimulus latency t^* on the $A - B$ plane. Each plane corresponds to a different value of the time constant λ of the recovery function within the range observed experimentally (0.2 $\leq \lambda \leq$ 1.2). (inset) Zoom-in to $-2 \leq B \leq 6.67$ to reveal the monotonic rise of t^* (dots and dashed line) that corresponds to the case described in Fig 3.3B ($\lambda = 1$). (B) Dependence of the gain in stimulation efficacy by using t^* over random stimulation latencies on the time constant λ of the recovery function. μ, A, B, and σ were set to 0.6, 20, 6.67, and 1 respectively. (C) IBI distributions for the range of values observed experimentally of the location parameter μ (0.6 $\leq \mu \leq$ 2) for A, B, λ, σ set to 20, 6.67, 1 and 1 respectively. (D) The family of objective functions corresponding to the IBI distributions in (C) shows the near linear relationship of the optimal latencies with μ (dots and dashed line) (A, B, λ, σ were 20, 6.67, 1 and 1 respectively; colours as in (C)). (E) Summary of the dependence of the optimal stimulus latency on the A–B–μ space for λ = 1. Each plane corresponds to a different value of the location parameter μ of the IBI distribution. (inset) Zoom-in to $-2 \leq B \leq 6.67$) to reveal the rise of t^* (dots and dashed line) that corresponds to the case described in Fig 3.3B ($\lambda = 1, \mu = 0.6$) (Kumar et al., 2016).*

updated its action-value function estimates, while in a testing round, it always chose an optimal policy based on the knowledge hitherto acquired. The time taken to run through with the experiment varied across networks, but was typically around 3-5 hours, covering ≈1000 SBs. This variability was due to differences in the average burst rate between networks. The latency chosen by the algorithm during the final testing session was considered the learned latency. To test the stability of the learned

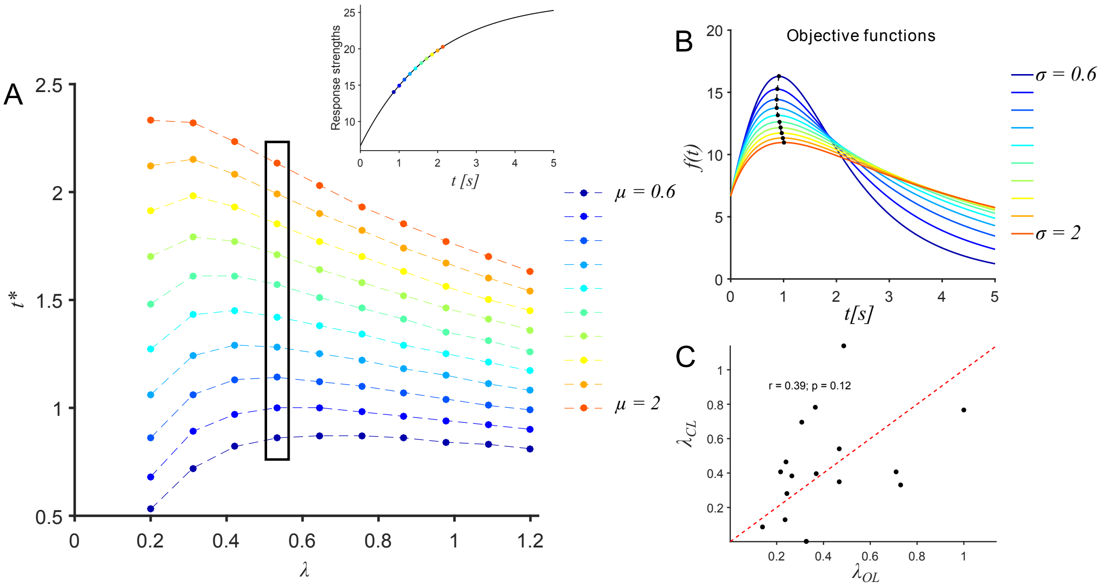

Figure 3.6. *Dependence of the optimal stimulation latency on the slope of the recovery function and the location and scale parameters of the IBI distributions (A) t* depends on the shape of the recovery function. t* shifts to later times with increasing recovery slope (λ increases) when average inter-burst intervals μ are short, i.e. spontaneous activity is high and the probability for interruption is high. In low activity regimes, however, the probability of interruption is low, hence t* is late and increasing the slope will lead to a decrease of the stimulus efficacy with increasing latencies since increasing interruption probability then outweighs the gain in spikes/stimulus. Because of the saturation of recovery changes in the probability for interruptions have a dominating influence on t*. (inset) t* shifts to later latencies with increasing μ for a given λ (boxed). A, B and σ were set to 20, 6.67, 1 respectively. (B) Scale parameter, σ of the IBI distribution had little impact on the optimal stimulation latency. A, B and λ and μ were set to 20, 6.67, 1 and 0.6 respectively. (C) Across networks, values of λ recovered from fits to closed-loop data were weakly correlated with open-loop estimates (Kumar et al., 2016).*

latency some of the sessions were run with up to 3000 SB in further training and testing rounds. Fig 3.7 illustrates a typical session in an example network. In this case, learning proceeded in three pairs of 200 training and 50 testing trials each. Note that a trial in our paradigm refers to the period between SBs where stimulation can potentially be delivered. Each trial is therefore initiated by ongoing activity (SB termination) and not by stimuli. Some of the trials were interrupted by ongoing activity, resulting in stimulus counts less than the planned number.

To analyse the closed-loop sessions, we first looked at the model parameters A, B and λ extracted from the recovery function and compared values predicted from open-loop sessions with those recovered from fits to the closed-loop data. Note that in this paradigm responses are available only at fixed latency intervals corresponding to the state definition (Fig 3.8A). The gain A of the network showed a strong positive

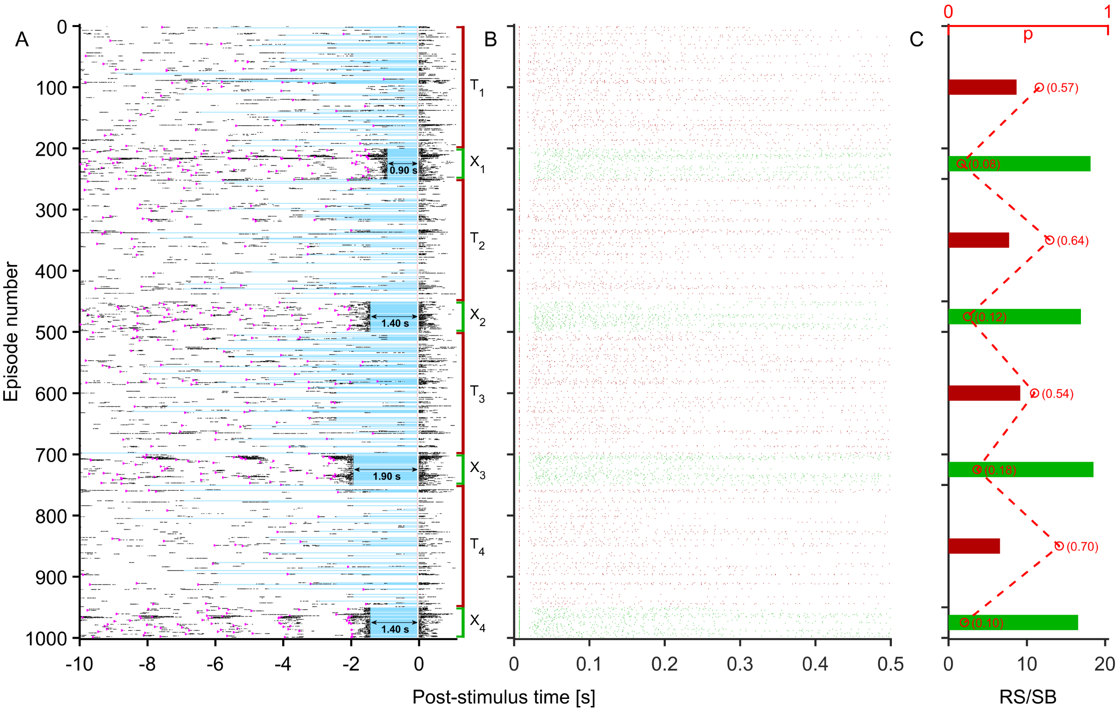

Figure 3.7. *A closed-loop learning session in an example network. The session consisted of 1000 trials (200 training (T_i, red), 50 testing (X_i, green) trials and 4 such pairs) (A) Raster diagram showing the activity at the recording channel around the time of stimulation. Trials interrupted by ongoing activity are left empty at $t > 0$ s. The spikes of the interrupting SB were removed in (A) and (B) for clarity. Successful stimuli evoked responses at $t > 0$ s. Blue lines mark the period of latency prior to the stimulus at $t = 0$ s Magenta triangles indicate stimuli delivered in preceding trials. Within training rounds, the controller was free to explore the state space. Note that these rounds are in closed-loop mode but with a random sequence of stimulation latencies. The strategy in this example was non-greedy. During testing rounds the hitherto best policy was chosen. After the final round, a latency of ≈ 1.4 s was learned in this example. (B) Zoom-in on responses evoked throughout the session. Interrupted trials without stimulation appear as empty rows. In this example all stimuli elicited responses. (C) Stimulus efficacy estimated as the response strength per SB (response strength (RS)/SB) computed over each of the training/testing rounds. RS/SB improved considerably during testing compared to the training rounds. The fraction, p, of trials interrupted in each round is shown as red circles and numerically. The dashed line was added for clarity (Kumar et al., 2016).*

correlation to the open-loop ones (r=0.91, p <10 $^{-5}$, n=15 networks, Fig 3.8A, B), indicating relative stationarity of the quantitative relationship and its accessibility for the controller. Parameter B, which can be interpreted as the excitability threshold for SB termination, too showed positive correlation (r=0.66, p=0.003, n=18 networks, Fig 3.8C) but weaker than gain A, suggesting that SB termination may depend on additional factors not captured by the model. Parameter λ showed a still weaker correlation (Fig 3.6C).

We then compared the learned stimulus latencies with those predicted from open-loop sessions. Overall, stimulus latencies learned by the controller showed a strong

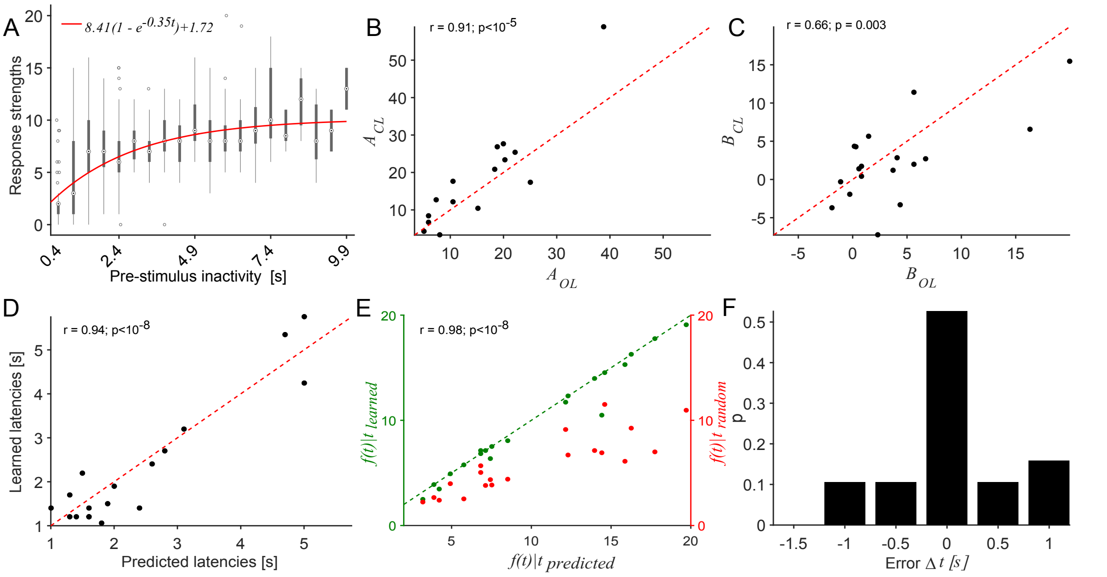

Figure 3.8. *Comparison of open-loop predictions with autonomously learned strategies. (A) Dependence of response strengths on pre-stimulus inactivities in data during a closed-loop session in an example network. Each box shows the statistics of response strengths recorded at one discrete state. The central measures are median and the edges with 25th and 75th percentiles. Whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually. The fit (red) was made to the medians. The minimal latency for burst termination was 0.4 s in this example, which was thus the earliest state available for stimulation. (B) Across networks, closed-loop estimates of the gain A correlated strongly with open-loop estimates (r=0.91, p<10*$^{-5}$*, n=15 networks), indicating that A was mostly stable during the experiments. (C) Similarly, closed-loop estimates of B were in agreement with open-loop ones (r=0.66, p=0.003, n=18 networks), although to a lesser degree. (D) Across networks, learned stimulus latencies show a positive correlation with predicted optimal values (r=0.94, p<10*$^{-8}$*, n=17 networks). (E) In spite of some variability in B-D the magnitudes of the modelled objective functions for predicted and learned latencies matched closely (green dots), indicating that the network/stimulator system was performing at a near optimal regime, regardless of slight discrepancies in the latencies. Exact optima were likely unreachable owing to the coarse discretization (0.5 s) of states. Red dots denote the corresponding magnitudes at* t_{rand} *for a strategy delivering stimuli at random latencies estimated as the mean of the objective function. (F) The distribution of errors between learned and predicted latencies is centered around the predicted optimum and confined to within 2 discrete steps from it (Kumar et al., 2016).*

positive correlation with the optimal latencies estimated from model predictions based on open-loop experiments (r=0.94, p<10^{-8}, n=17 networks, Fig 3.8D). Nevertheless, in some networks learned latencies differed from predicted ones, as is visible in their distances to the diagonal in Fig 3.8D. Next we compared stimulation efficacy, estimated as the response strength per trial, corresponding to learned and estimated latencies. The objective function, $f(t)$ (Eq 3.4) calculated specifically for individual networks was used to estimate the maximal stimulation efficacy $f(t)|t$ achievable with the predicted optimal latency vs. the one learned for a given network. Values of this measure were in strong agreement (Fig 3.8E), indicating that the control goal was achieved despite errors in predicted latencies (Fig 3.8D). One possible source of errors could be the discretization of the controller's state space into 0.5 s steps. Indeed, the error distribution showed that 74% of the networks studied fell within ±0.5 s around the optimum (Fig 3.8F).

Finally, the performance of the controller was evaluated with respect to the defined goal: to maximize stimulation efficacy measured as the total number of response spikes evoked for every detected SB in the network. A session-by-session analysis showed that in 94.2% of the sessions (n=52 sessions with non-greedy training, 11 networks), the percentage of interrupted events per session diminished post learning (Fig 3.9A). While the number of spikes in a response did not significantly change across sessions (Fig 3.9B) the standard deviation across stimuli in a session decreased (Fig 3.9C; p=0.01, two-sample t-test). Concurrently, in 90% of the cases (n=52 sessions, 11 networks), stimulus efficacy had increased after learning, supporting the effectiveness of the learning algorithm.

Model parameters were derived from fits to noisy experimental data, i.e. IBIs during spontaneous activity and response strengths during open-loop stimulation. Predictions made from these models were therefore only as good as the fits. Comparison of optimal stimulus efficacies predicted from our models with those achieved during the final closed-loop testing sessions showed that achieved efficacies were within the 99% confidence interval for the models fitted for each network (Fig 3.9D). Achieved stimulus efficacies fell within the interval in 8 of 11 networks studied.

Learning clearly improved performance in each network (p<0.001, two-sample Kolmogorov-Smirnov test). The amount of improvement, however, varied across networks (Fig 3.9E). To compare performance across networks, we captured each network on a normalized response-per-stimulus vs. interruption probability plane (Fig 3.9F). Each network is shown before and after learning. Only the last pairs of sessions were used for this plot (n=11 networks). The distribution shows a clear separation of the mixed-mode performance before and after learning, indicating the improvement of stimulation efficacy. The improvement was almost exclusively due to a reduction in interruption probability (Fig 3.10 and Fig 3.11). This, however, also says

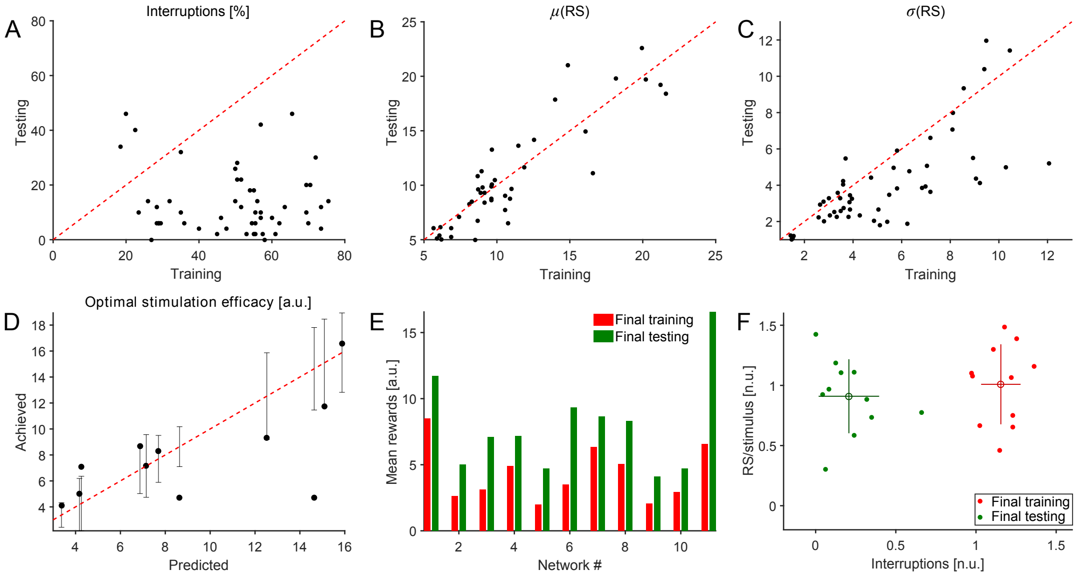

Figure 3.9. *Performance evaluation of the controller. (A) The percentage of interrupted trials during training (x-axis) and testing (y-axis) sessions (n = 52 pairs across 11 networks). This percentage decreased sharply after learning in 94.2% of the recorded sessions. (B) The mean RS evoked per stimulus was, however, preserved in both sessions. (C) The variability in RS per stimulus decreased significantly (p=0.01, two-sample t-test). (D) Comparison of the optimal stimulus efficacies predicted from our models with the efficacies achieved during the final closed-loop testing sessions. Vertical bars represent 99% confidence intervals corresponding to the models fitted for each network. Achieved values fall within the interval in 8/11 networks studied. (E) Mean rewards were calculated over trials in the final training and testing rounds to compare the controller's performance. After learning, mean rewards increased in each network, which is indicative of the improvement in stimulation efficacy. The rewards across the sequence of trials in each round were drawn from distinct distributions in every network (p<0.002, two-sample Kolmogorov-Smirnov test). The individual distributions are shown in Fig 3.10 and Fig 3.11. (F) Summary of learning across networks on a normalized RS/stimulus vs. interruption probability plane (11 networks). Only final training and testing rounds were considered. Normalization for interruptions was performed relative to the model-based estimate of interruption probabilities, corresponding to stimulation at random latencies for each network. The RS/stimulus measure was similarly normalized to the model-based estimates of the efficacy assuming a random stimulation strategy. The improvement in performance clearly separates the data points in the plane. Of the two modalities that contribute to stimulus efficacy, the improvement was dominated by reduction of interruption probabilities (Kumar et al., 2016).*

that the controller learns to avoid losing in response magnitude by not further reducing the interruption probability, i.e. it balances the trade-off.

3.4 'Learning' from the machine

An attractive element of autonomous learning approaches is its potential to exploit patterns and quantitative relationships not readily accessible to conventional analyses. Neuronal networks, being composed of noisy non-linear elements and prone to temporal non-stationarities, are extremely onerous to characterize in terms of input-output

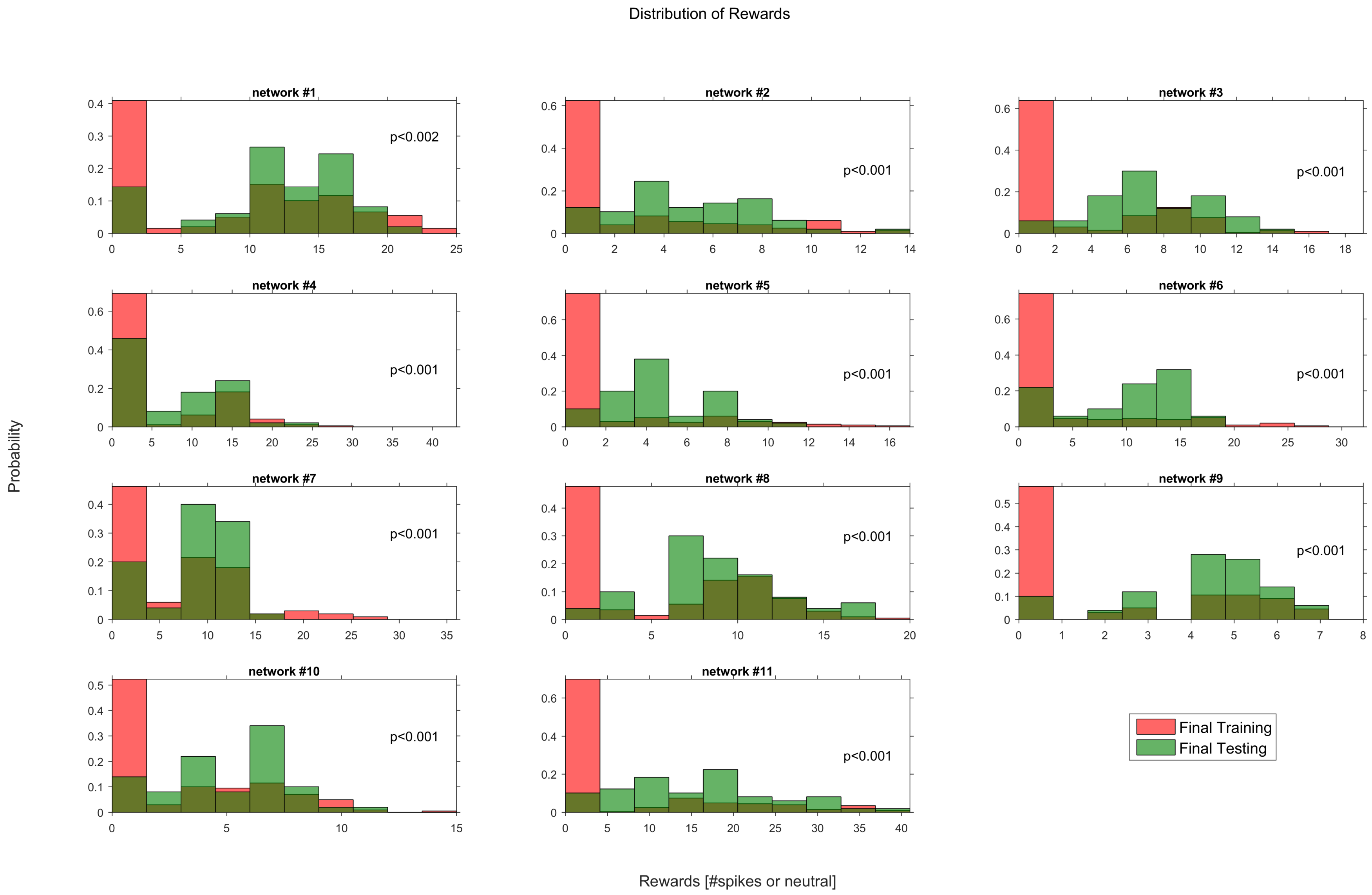

Figure 3.10. *Reward probability distributions for all networks. In each training trial the controller received a reward according to the number of spikes elicited by the stimulus. In trials interrupted by SBs this resulted in neutral reward* (-10^{-3})*, pooled with trials eliciting* 0 *spikes in the histograms. After learning, the probability for very high rewards was reduced but this was outweighed by the lower frequency of* 0 *and neutral rewards (Kumar et al., 2016).*

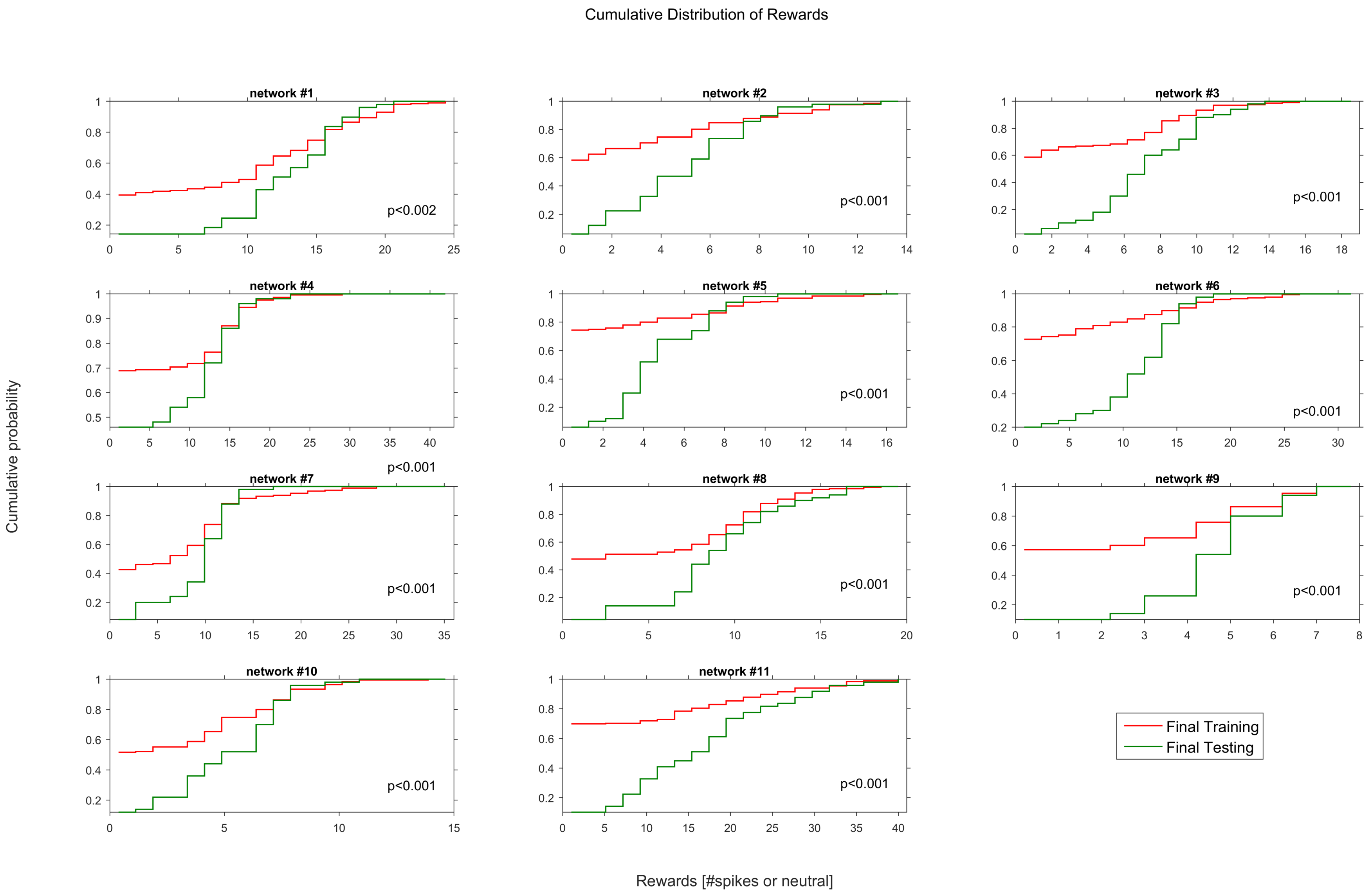

Figure 3.11. *Empirical cumulative distribution of rewards for all networks. The Empirical cumulative distribution function (ECDF) of the rewards clearly shows that the improvements by learning were dominated by reduced probabilities to receive* 0 *or neutral rewards (Kumar et al., 2016).*

relations. In this context, the ability to selectively clamp defined activity features offers an opportunity to unravel the functional complexity of the system. It might be possible to better understand the network from exploring the conditions under which goal-directed behaviour was sustained.

In the context of our toy problem, the nature of the underlying quantitative relation of stimulation and ongoing activity was known from previous studies (Weihberger et al. (2013) and its extensions described above). Thus the inverse problem translates to whether properties of this relation can be salvaged from the learned machine. Our experiments proceeded based on a Q-learning algorithm. The Q-function, though not a model of the input-output relationship per se, should be able to capture aspects of the underlying quantitative relationship because of the manner in which rewards and punishments affect it.

To test this proposition, we asked if starting from the learned Q-function one could infer the quantitative stimulus-response relationships underlying each network. Fig 3.12 shows the Q-table from an example network. The y-axis represents the Q-value associated with choosing a particular action (wait/stimulate coloured green/red) at a given state. The Q-values corresponding to stimulating at each state seems to approximate the exponential model as shown by the least square fit to the red data points (red line).

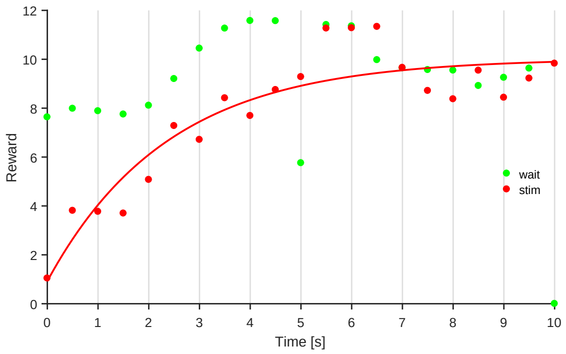

Figure 3.12. *Example Q-table after learning. Each dot represents values of the Q-function corresponding to each state (time) and action (wait/stimulate coloured green/red respectively). Q-values for stimulation agreed with the known model relating response strengths to pre-stimulus latencies.*

We extracted the model parameters resulting from fits to such Q-values across networks and found that they were strongly correlated to those from model fits made to response strength series from the closed-loop session (see Fig 3.13). This suggests that it might indeed be a valid proposition to learn from the machine.

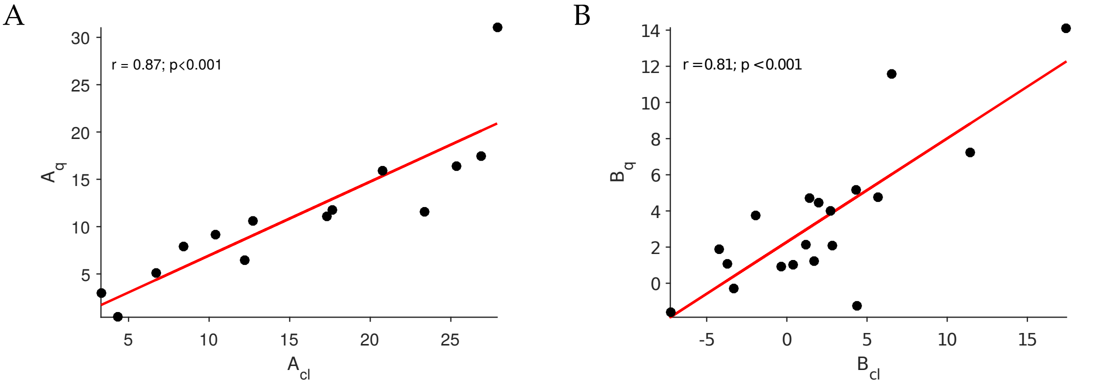

Figure 3.13. *Comparison of model fits to the Q-table with closed-loop models. Parameters A (panel A) and B (panel B), extracted from exponential model fits to Q-values (subscript 'q' in figure) were strongly correlated to those from fits to response strengths during the closed-loop session (subscript 'cl' in figure).*

Model parameters extracted from the controller were also correlated to those obtained from open-loop data recorded prior to the learning session (Fig 3.14). Interestingly, these correlations were considerably weaker than with parameters closed-loop data. The observation is suggestive of temporal trends in the recovery function over long time-scales.

This observation seemingly contradicts our earlier assumption of system stationarity that was necessary to evaluate the quality of solutions learned autonomously. However, our specific argument was that by coarse graining the state-space discretization of our controller, the objective function and hence the optimal solution remained reasonably invariant under the range of temporal modulations observed in our model parameters and that therefore the optimization problem itself could be assumed relatively stationary (see Fig 3.4, 3.8F). In the next chapter we characterize temporal fluctuations, drifts and inhomogeneities that underlie network-stimulus interactions.

3.5 Drawbacks of session-wise learning

In these experiments the closed-loop framework was designed to learn session-wise. A set of trials were exclusively dedicated to explore the action space (training session). In the next session, the controller executed an optimal policy based on its experiences until then. No further learning occurred during this period. While such a strategy contributed to the straightforward separation of performance measures before and after

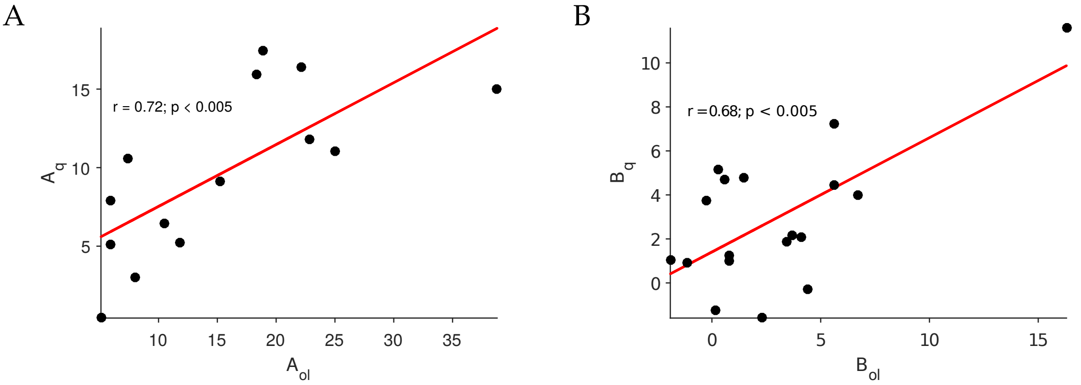

Figure 3.14. *Comparison of model fits to the Q-table with open-loop models. Parameters A (panel A) and B (panel B), extracted from exponential model fits to Q-values (subscript 'q' in figure) were correlated to those from fits to made to open-loop data collected prior to the each closed-loop session (subscript 'ol' in figure).*

learning in this study, it may not be ideal for general problems involving interactions with biological neuronal networks.

Effects illustrating why this may be the case were observed during some of our sessions where the nature of interactions was distinctly different across adjacent training and testing sessions. After initial an training session, a low value of stimulus latency was incorrectly learned presumably because of the small number of trials. repetitively delivered to the network, resulting in increased response delays (see Fig 3.15). Our feedback mechanism failed to capture this unintended consequence, not just because it was in the testing phase, but also since delays were not a state feature that the controller was designed to measure.

Note that such effects may not be relevant to the outcome of this particular study, since it can be shown that response strengths do not depend on delays leaving our objective function invariant. But the principle, when extended to a yet unconsidered dependent feature, could be consequential in a more general application context (see Section 6.1.5 for further discussion).

Summary

- We identified a toy trade-off problem emerging as the interplay of ongoing and evoked activity.
- We developed and evaluated algorithms to autonomously optimize stimulation

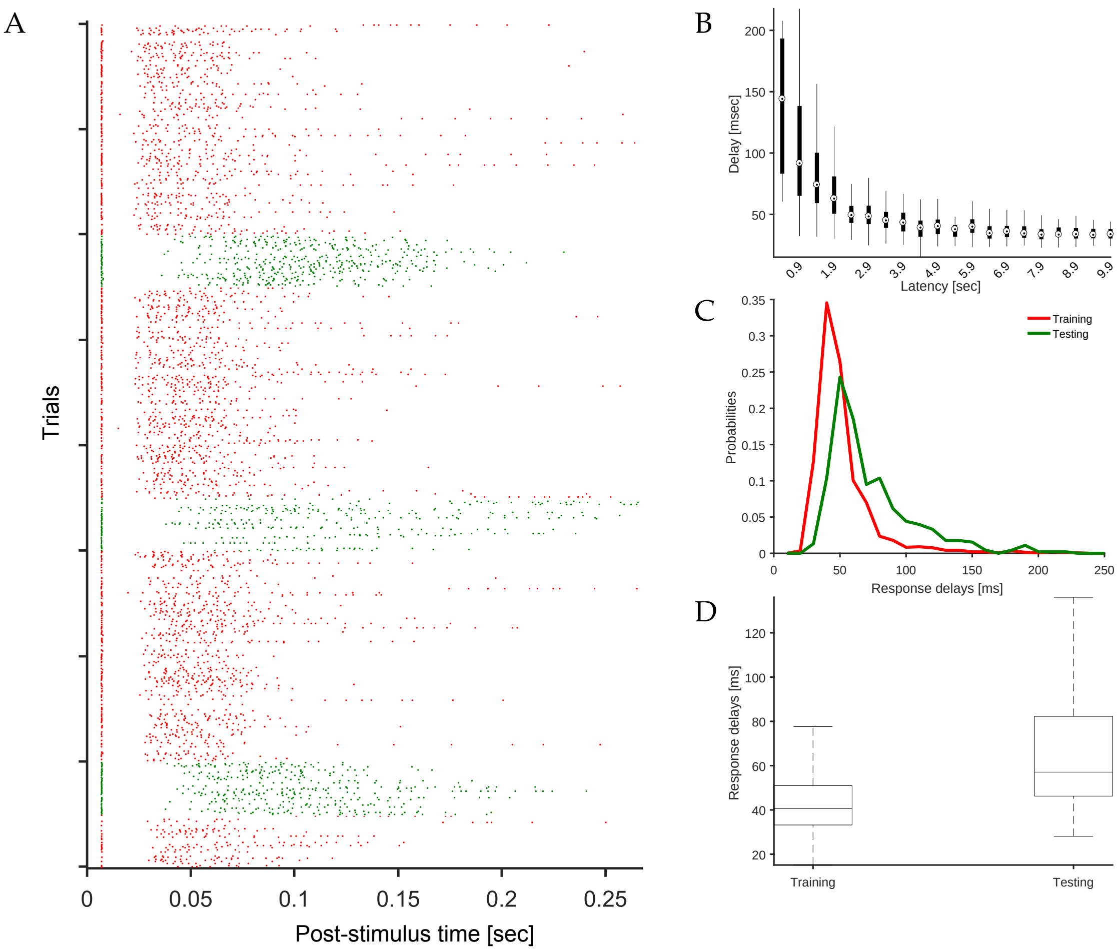

Figure 3.15. *Response delays observed during training (green) and testing (red) sessions were distinctly different. (A) The raster shows hundreds of trials of the closed-loop trade-off problem in an example network. Stimuli were delivered at 0 s. (B) The relationship between response delays, defined as the time to the second phase of the response and stimulus latencies. (C) The distribution of response delays during training vs. testing shows they were drawn from distinct distributions (Kolmogorov-Smirnov test, p<0.001) (D) Response delays during training and testing: each box shows the median, 25th and 75th percentiles; whiskers indicate extreme data points.*

policies for goal-directed interaction with biological neuronal networks (BNNs).

- We numerically verified that the problem was well-defined and had a unique solution for each network and predicted optimal stimulus policies for each network.
- Policies learned autonomously were in good agreement with those predicted.
- We demonstrate that RL based techniques may indeed be feasible to exploit underlying network-stimulus relationships to find optimal policies.
- Our proof-of-principle study is the first demonstration of the potential of artificial agents to interact optimally with biological neuronal networks.

Chapter 4

Temporal Inhomogeneities in Stimulus-Response Relations

In Chapter 3, we presented proof-of-principle of an autonomous RL based controller capable of interacting optimally with respect to a pre-defined goal with generic neuronal networks *in vitro*. To validate autonomously learned policies, we relied on predictions from open-loop data acquired at a different point in time. Stationarity of system dynamics was thus a necessary assumption to be able to evaluate the quality of solutions learned autonomously. By coarse graining the state-space discretization of our controller, we argued there that the objective function and hence the optimal solution remained reasonably invariant under the range of temporal modulations observed in our model parameters and that therefore the optimization problem itself could be assumed relatively stationary.

However, spatio-temporal fluctuations in neuronal activity dynamics are ubiquitous in the brain and have long been described (Cabral et al., 2014, and references therein). Distinct spatial, temporal and spectral patterns in network activity are thought to reflect the underlying functional connectivity in a state (wake, rest, sleep etc.) and task (active, lying, walking etc.) dependent manner. The origin of these fluctuations are thought to lie in networked neurophysiological mechanisms (Kopell et al., 2014; Medaglia et al., 2015; Medaglia et al., 2017). Under pathological conditions a further layer of time-evolving processes that disrupt the healthy evolution of brain states are thought to be involved (van den Heuvel and Sporns, 2013). The key question from a neurotechnological perspective is how we can develop safe and adaptive stimulation solutions capable of operating efficiently atop the dynamic floor of ongoing activity.

In this chapter, we show that fluctuations and drifts in stimulus-response relations over long time-scales are features of generic neuronal networks *in vitro* that are likely mediated by networked neurophysiological mechanisms. The dependence of evoked response strengths on the latency of the stimulus to prior activity in the network was described in chapter 3. The non-linear model fits a saturating exponential of the form $A(1 - e^{-\lambda t}) + B$. We used this stationary model as a baseline to investigate the presence of long-term trends and fluctuations in the input-output relationship. To test for temporal inhomogeneities we asked if sustained periods of hypo/hyper-excitability w.r.t. the model prevail in the response residual time series.

We make the case that defining a fixed goal – for e.g. clamping response strengths to a pre-defined value – in our model system is strikingly similar in structure to a generic control problem in the clinical context. The challenge then is to adapt to partially observable and poorly characterizable dynamic fluctuations of the underlying network states. Our experimental configuration offers a controlled setting to develop algorithms and better understand the challenges of adaptively achieving targeted interaction with biological neuronal networks.

Slow activity fluctuations, though reported in various studies using cultured neuronal networks *in vitro*, are yet to be comprehensively characterized in literature (Baltz and Voigt, 2015; Haroush and Marom, 2015). Because the mechanisms and relevant time-scales behind these modulations remain unclear, no numerical models describing them are available.

In our data, temporal inhomogeneities were discernible across hours of repetitive interaction with the network during closed-loop optimization experiments described in Chapter 3. Long sequences of stimulation placed at random latencies relative to ongoing activity also exposed underlying slow fluctuations once the expectation value determined from the fitted exponential model was subtracted from individual trials. Further, slow fluctuations were observed in the burst strengths of spontaneous events occurring between stimuli. They could serve as indicators of the background process modulating stimulus-response relationships and help predict successive responses. Simple predictive models, however, were difficult to infer due to the non-stationary nature of the relationship.

4.1 Fluctuation of stimulus-response relations in closed-loop sessions

We analysed the evolution of response strengths during the closed-loop trade-off problem, which involved repeated interaction with the network over a time-scale of hours. Data from both training and testing sessions were pooled. Latencies were explored in steps of 0.5 s during the experiment and grouped according to the chosen latency. A saturating exponential model fit was made to median response strengths at each latency (Fig 4.1A). The data revealed considerable variability around the fitted model across states, particularly in the earlier ones where sampling of the distribution was richer. This was partly due to the fact that during the learning process certain latencies were progressively preferred over the others. The other factor was that trials with stimulation planned for longer latencies were much more likely to be interrupted by ongoing activity, resulting in fewer samples at these latencies. The latency-wise distributions of response strengths are shown in Fig 4.1B.

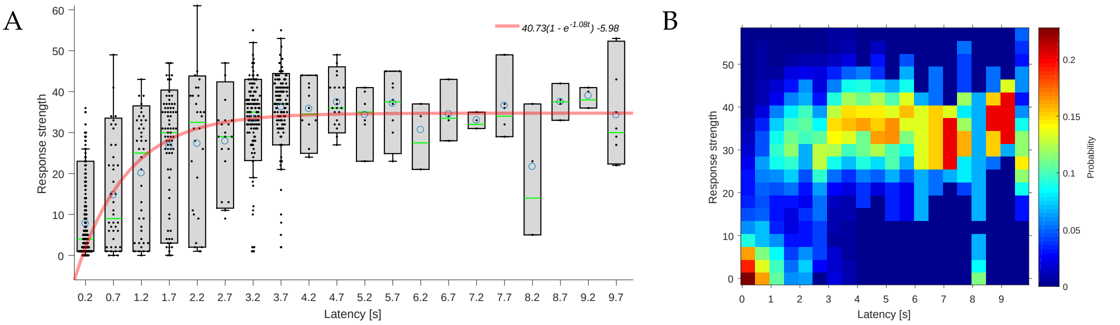

Figure 4.1. **Variability around the recovery function.***(A) Box plot capturing the relationship between response strengths and periods of prior latencies during the closed-loop trade-off experiment for a sample network. Data from both training and testing sessions were pooled. Each box shows the statistics of response strengths recorded at each discrete state. The data were overlaid with a jitter along the x-axis for visualization. The central measures are median (green lines) and mean (blue open circles); each box extends between the 25th and 75th percentiles and whiskers to extreme data points not considered outliers. The least-square fit (red) was made to the medians. (B) The distribution of response strengths at each state. Probabilities are colour coded.*

It is possible that the breadth of the response strength distribution belies just a measure of 'noise' involved in the underlying data generating process, i.e. spiking activity evoked and measured in a recurrent network as a response to focal stimulation. One way to rule this possibility out is to analyse the order plot of the residuals, i.e. the sequence of errors to the model fit. If the model were indeed fitting the central measure of noisy data from a stationary source, the sequence of residuals would be expected to be statistically independent, i.e. without serial interaction. Interestingly, order plots across our networks revealed considerable serial correlations, rhythmic trends and drifts over long time scales. Example order plots for two networks are shown in Fig 4.2. The residual sequence was smoothed over a 3 sample box filter to smooth out fast changes. The network in Fig 4.2A exhibited a general oscillatory trend in residuals while that in Fig 4.2B showed marked drifts over time.

The self-similar nature of the underlying data generating process was captured in the occurrence of closed-orbits of varying circumferences and foci in the session wise second order return map of the residual sequence (Fig 4.3). Residuals appeared more variable during training than in testing sessions.

The recurring pattern in the data was confirmed in the autocorrelogram of the overall data (Fig 4.4A). Correlation coefficients – both positive and negative – alternated with increasing lags. The Ljung-Box Q-test, a quantitative test for residual autocorrelation, rejected the null hypothesis of no autocorrelations at each of the lags tested (5 – 50). The Wald-Wolfowitz runs test also rejected the null hypothesis of randomness in the model residuals ($p < 10^{-6}$).

Amplitudes and time-periods of the fluctuations observed in the residuals appeared widely distributed. To reveal the temporal scales of modulation, we analysed the inter-event interval distribution of data points above and below $\sigma/2$ of positive and negative residuals (Fig 4.4B). Both distributions were bimodal in nature, with the earlier peak capturing intervals within each residual cluster and the later one between clusters. Note the logarithmic scaling of the time axis. Inter-cluster events seemed log-normally distributed over 1 to 10 minutes.

The residuals, during training and testing sessions were shown to be drawn from

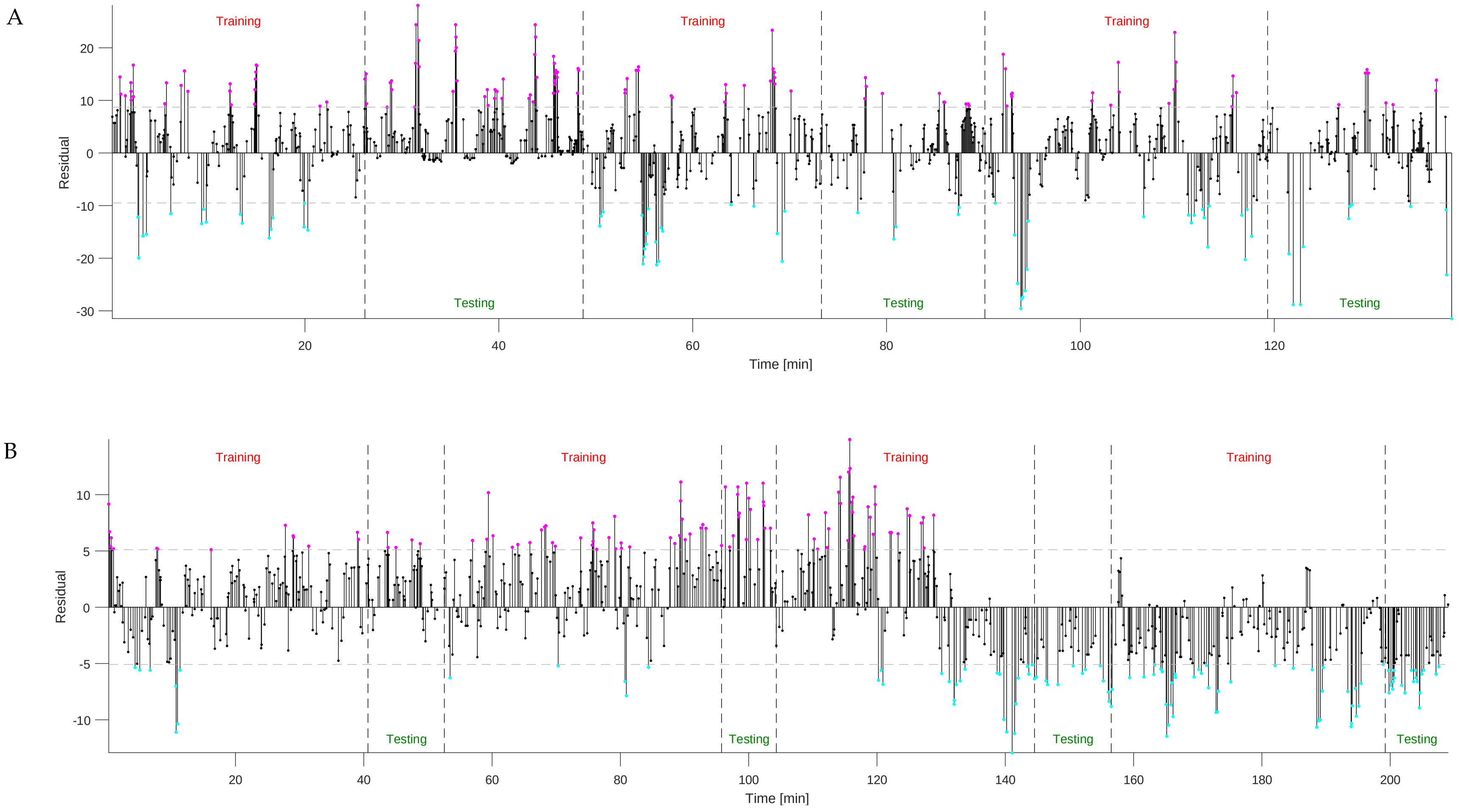

Figure 4.2. *The sequence of residuals relative to the saturating exponential model during the closed-loop trade-off experiment shown for two example networks in (A) and (B). Training and testing sessions are separated by dashed vertical lines. To reveal the modulation of response strengths over long time scales, data points over and below $\sigma/2$ of positive and negative residuals (dashed horizontal lines) are coloured in magenta and cyan respectively. These data points tend to appear in alternating clusters, each of which repeats over a period of about 5-10 minutes. Apart from fluctuations, some networks also displayed drifts in activity. The example network shown in (B) exhibited a pronounced non-stationarity in residuals 2 hours into the closed-loop session.*

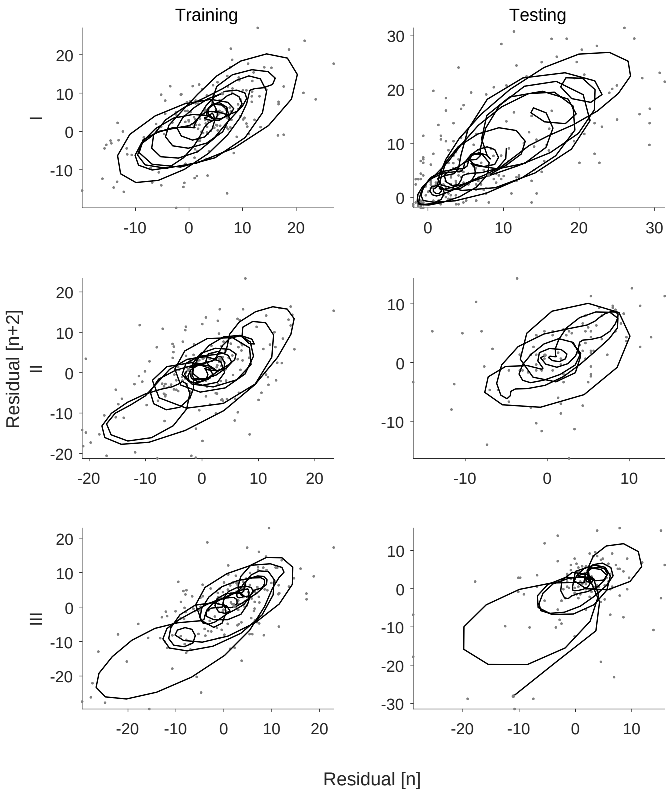

Figure 4.3. *The recurring closed trajectories in the session-wise second order return maps of the residuals sequence during the three training and testing sessions in a closed-loop session (left and right columns respectively) points to an underlying slow modulation of the data generating process. Individual data points are shown in grey. Black lines are a smoothed overlay. Data are from the same network in Fig 4.2A.*

distinct distributions (Kolmogorov-Smirnov test, $p < 0.001$; Fig 4.4C, D). A likely explanation is the asymmetry imposed by the positive-definite nature of response strengths. An alternative hypothesis is heteroscedasticity in the data, i.e. response variability may itself be a function of the stimulus latencies. However, we did not find significant differences when squared residual magnitudes were compared for similarly sampled latencies (e.g. latencies 0.2, 3.2 and 3.7 s in Fig 4.1A, Bartlett's test, $p = 0.33$).

The periodogram of the residual time series revealed dominant frequencies in the range of 2 – 35/hour (red dots in Fig 4.5A). Similar results were observed across the 20 networks we studied for the optimization experiments. The average dominant time period observed in each network is shown in Fig 4.5C. The distribution of dominant periods pooled across all networks is shown in Fig 4.5D (329 frequency peaks from 20 networks).

To quantify the skewed spectral distribution, we used median frequency as a metric (red line in Fig 4.5A, B). This allowed us to test if the lower frequency peaks were

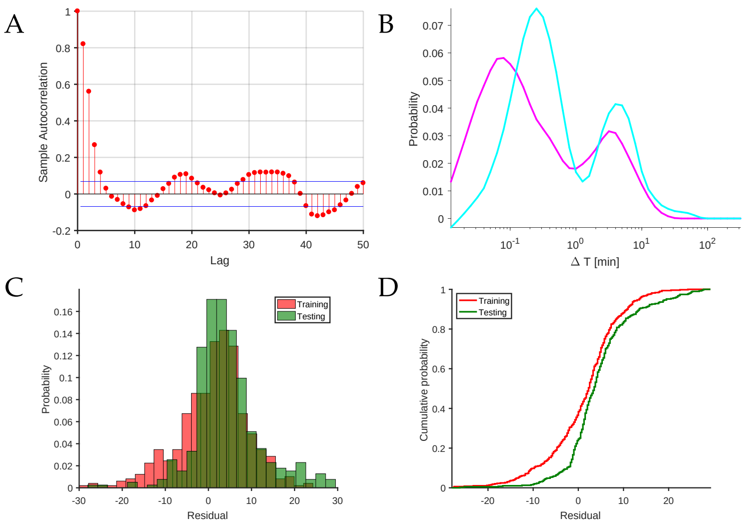

Figure 4.4. *(A) Autocorrelogram reveals the residual autocorrelation structure. Significant autocorrelations–both positive and negative– exist outside the Bartlett two-standard error bands for white noise– given by the blue lines, for various lags. (B) The bimodal distribution of intervals between residuals above and below $\sigma/2$ of positive and negative residuals and were coloured magenta and cyan respectively (see Fig 4.2A). The residuals during training and testing phases were drawn from distinct distributions (Kolmogorov-Smirnov test, $p < 0.001$). (C,D) The distribution of residuals and cumulative distribution functions during the training and testing phases. All panels are based on the network in Fig 4.2A.*

merely artefacts of the time series smoothing. We generated 250 shuffled surrogates for each raw residual series and applied the same smoothing kernel (Hinich et al., 2005). The 95% threshold level $\mu_{0.95}$, for which 95% of the surrogates had a median frequency greater than $\mu_{0.95}$ was calculated. The observed median frequency lay below the $\mu_{0.95}$ in 14 out of the 20 networks studied (Fig 4.6).

4.2 Fluctuation of stimulus-response relations in open-loop sessions

We then analysed the evolution of response strengths during repeated interaction with the network in open-loop sessions. Stimuli were delivered periodically every 10 s. The periods of inactivity prior to stimuli were nevertheless random due to irregularly occurring spontaneous bursts (Fig 4.7). Furthermore, the temporal evolution of response strengths revealed a weak fluctuating trend.

We fitted this data with a saturating exponential model of the form $A(1-e^{-\lambda(t-\tau)})$ (Fig 4.8A). and computed model residuals. The succession of residuals exposed an un-

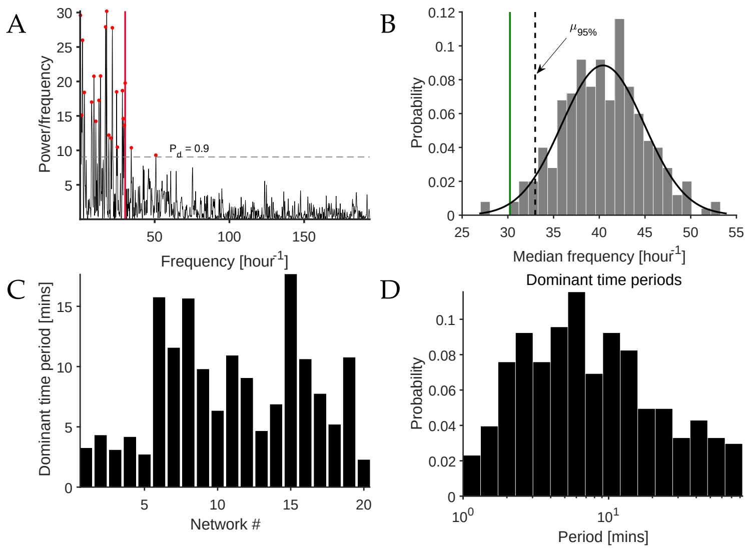

Figure 4.5. *(A) Lomb-Scargle periodogram estimates of the residual time series shows several dominant peaks in the range of 2 – 35/hour. The probability that a peak in the spectrum is not due to random fluctuations of peak detection was set to 0.9 (grey dashed line). Red dots indicate the detected peaks and the red line, the median frequency. (B) The distribution of median frequencies in the 250 shuffled instances of the residual time series. The black line is a Gaussian model fit to the data. The solid line indicates the median frequency observed during the closed-loop session. A threshold, $\mu_{0.95}$ was defined such that 95% of the surrogates had a greater median frequency (dashed line). The observed median frequency was lower than the threshold, suggesting that the slow fluctuations observed in the data were likely not an artefact of smoothing. Both (A) and (B) correspond to the network shown in Fig 4.2A (Nw# 1 in Fig 4.6). (C) The average dominant time periods observed in each of the networks studied varied widely between 2 – 20 minutes. (D) Dominant time periods pooled across networks were broadly distributed over a scale of minutes to tens of minutes.*

derlying temporally inhomogeneous trend despite its distribution being approximately normal (Fig 4.8B, C).

We corrected response strength values with the slowly varying residual trend to illustrate that, as expected, compensating for such modulations would make response strengths more predictable (Fig 4.10). First, response strengths were corrected using the approximate sinusoidal model that we fit to the modulatory trend, resulting in moderate improvement in the predictive power of the model (Adjusted $R^2 = 0.73$, Fig 4.10A). The modulatory trend in the raw residuals was considerably dampened (Fig 4.10B, C) while remaining approximated normally distributed. Even better improvements in the model's predictive power was observed when response strengths were compensated using a history-based lagged estimator ($R^2_{adj} = 0.81$, Fig 4.10D). The repetitive structure of the residual sequence was further weakened. The residual distribution was verified to be approximately normal (Fig 4.10E, F).

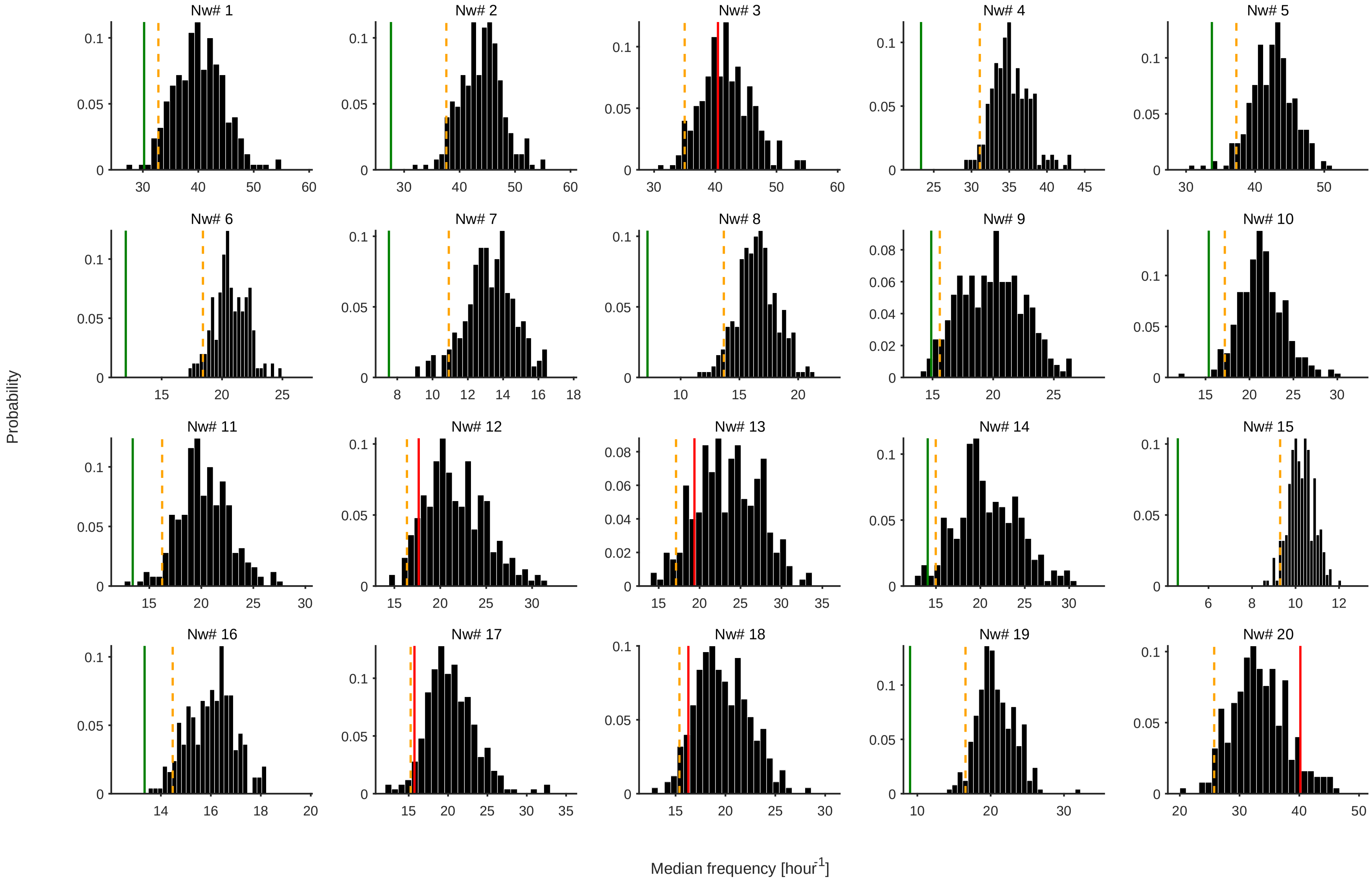

Figure 4.6. *The distribution of median frequencies of shuffled surrogates of the raw residuals for the twenty networks studied. Solid lines indicate median frequencies of the unshuffled residuals. Dashed orange lines are the 95% threshold median frequency threshold computed for each network. The observed median frequency was less than the 95% threshold in 14 out of 20 networks (green). In the remaining 6 networks, the median time periods were in the 1 – 5 minute range, which could be attributed to the smoothing kernel (red).*

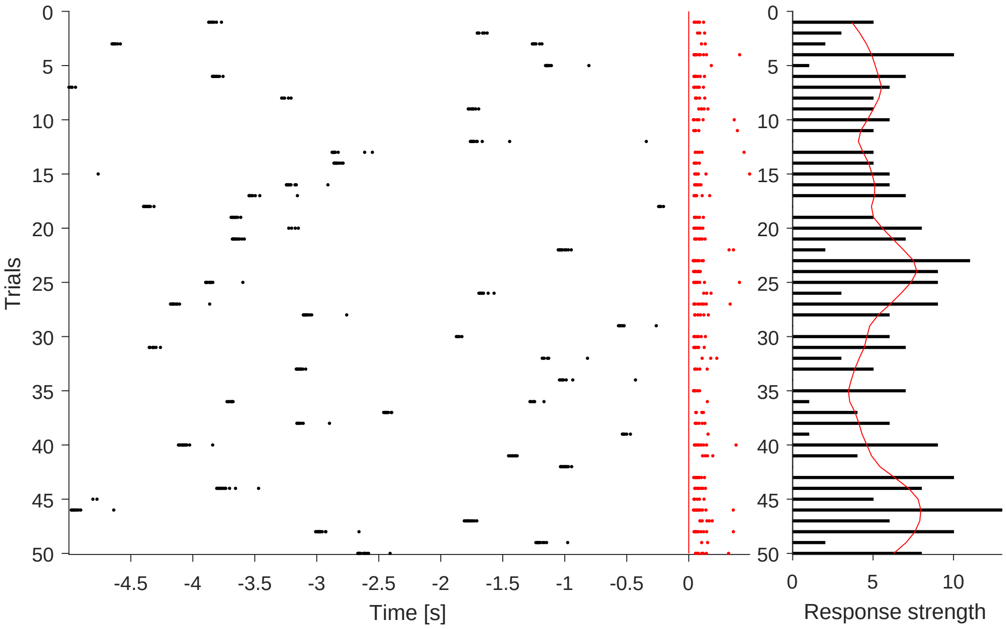

Figure 4.7. *Raster plot shows the result of stimulating the network at a single site periodically every 10 s. The periods of latencies relative to previously occurring spontaneous events was random. Figure shows 50 trials where stimuli were delivered at t=0 (red line). Spontaneous bursts are shown in black and evoked responses in red. The panel in the right shows trial-wise response strengths. The overlaid smoothed trace showing the temporal evolution of response strengths revealed a weak fluctuation with generally stronger responses around trials 25 and 45.*

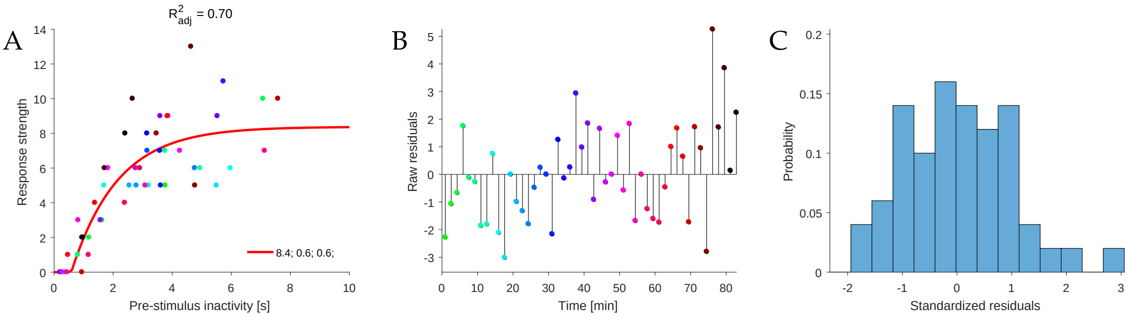

Figure 4.8. *(A) Exponential model fit to response strengths vs. pre-stimulus inactivities from the example in figure 4.7. The model used here was of the form $A(1 - e^{-\lambda(t-\tau)})$; fit parameters A, λ and τ shown in the legend. (B) The sequence of residuals w.r.t to the fitted model point to an underlying autocorrelative structure. The colour code in (A) and (B) corresponds to the temporal order of the trials. (C) The distribution of residuals was approximately normal.*

A sinusoidal model was clearly not a perfect fit to residual dynamics. This was reflected in the limited improvement in the predictability of response strengths not only in this particular example but also in data across networks. In general, residual dynamics exhibited amplitude and frequency modulation in a addition to drifting trends. Fitting a simple generalized modulation model under these circumstances would be cumbersome. A non-parametric history based regressor might be a more feasible approach. Our results imply that an approach seeking to predict or control

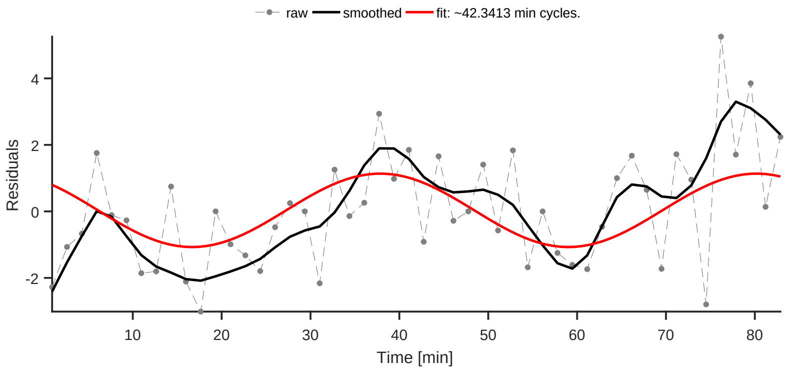

Figure 4.9. Residuals from the same network as in Fig 4.8 shown here overlaid with a smoothed trace and a least square sinusoidal that fits the data with a period of approximately 42 minutes.

properties of the network's response to stimuli may have to factor in not just timing of stimuli relative to ongoing activity but also history dependent processes whose temporal scale of impact remain poorly understood.

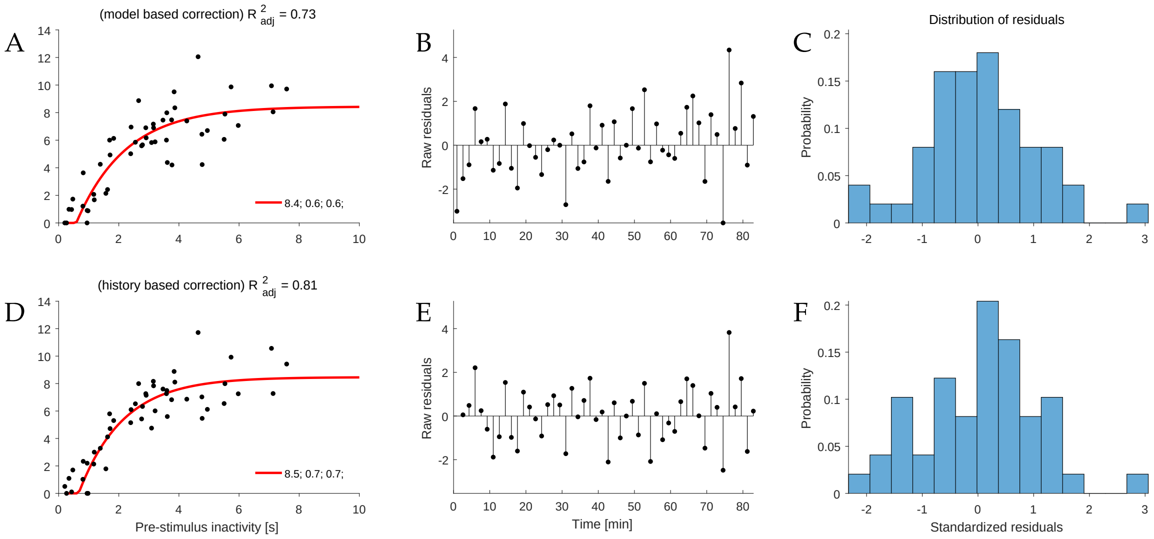

Figure 4.10. Correcting individual response strengths with the estimates of fluctuations based on the sinusoidal fit (A) and history based smoothed trace (D) results in improvements to the goodness of fit to the exponential model while also bringing down serial correlations in the residual time-series (B,E). The remaining residuals were still approximately normally distributed in each case (C,F).

History dependence in the evolution of response residuals on long time scales implied that past interactions could be valuable to make local predictions of future response strengths. Apart from this modality, could other features of network activity help predict future responses? To this end, we investigated if the background modulating process is reflected in features of ongoing activity patterns and if they could help predict modulations observed in stimulus-response relations.

4.2.1 Relationship with spontaneously generated events

In this section we investigated if the processes mediating stimulus response modulations could be reflected in features of ongoing activity and thereby used as predictive indicators of upcoming responses. As discussed in the previous section, the period of inactivity prior to stimulation is a known predictor of response strengths. Therefore, we first removed the influence of pre-stimulus latencies by converting actual response strengths to errors relative to a saturating exponential model fit to the data (residuals). Various features of ongoing spontaneous bursts were tested for correlations to successive residuals. Though long sequences of features like burst strengths (single channel and global), burst widths, peak rates within bursts were uncorrelated to response residuals, slow trends were observed in their temporal dynamics.

Consider the sequence of residuals and spike counts from prior spontaneous events at the same channel, normalized and overlaid in Fig 4.11. Data belong to the same network discussed in the previous section (see Fig 4.8). Rhythmic modulations are evident in both time series suggesting that signatures of the background processes mediating fluctuations of the recovery function may be reflected in ongoing activity as well. A time-resolved cross correlation of the two traces was calculated for 15 minute sliding windows ($dt = 0.25$ minutes) and a maximal lag of ± 5 minutes. Interestingly, over the first 30 minutes of the session residual dynamics were inversely correlated with the strengths of preceding ongoing events. Thereafter the pattern switched to a longer regime where they were positively correlated.

Though such switches in co-modulation were qualitatively observed across networks, a general quantitative framework remained elusive. Our data suggest that indeed the history of ongoing activity might be informative of the magnitude of response residuals, though the mapping remains unclear. A network-specific model fitted to such data would likely be of little worth since the relationship appeared temporally inconsistent. Locally fitted models may help but are cumbersome to realize given the lack of clarity in the nature and time scales of their dynamics.

It is conceivable that the exponential model-based compensation introduces artefacts to the residual time series due to non-uniform sampling of the period of prior inactivity.

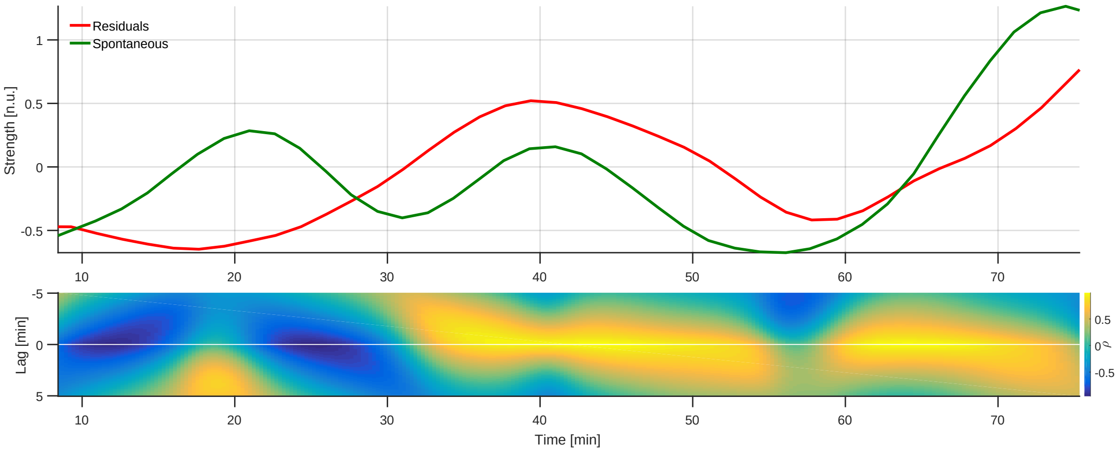

Figure 4.11. *Smoothed traces of normalized residuals (red) and spontaneous (green) origin (top). Only the spontaneous event in the immediate history of the stimulus was considered. Time resolved cross correlation of the standardized snippets of the two traces show periods where the two sequences are negatively and positively correlated. Switches between the two happened over a scale of tens of minutes indicating temporally inhomogeneous system behaviour.*

Further, differences in the quality of fitting and stimulation rates across networks also makes it cumbersome to assess fluctuations observed in the response residuals. One way around the problem is to control for latencies, i.e. stimulate in a time locked fashion at fixed latencies and evaluating residuals with respect to the observed mean response strength. This obviates the need to compensate for the contribution of latency to response strengths.

We devised such a closed-loop experiment where spontaneous activity at a chosen channel was continuously monitored for opportune moments to deliver stimuli. Only a pre-set period of silence had elapsed would the stimulus be delivered. An additional constraint of a 10 s minimal inter-stimulus-interval was placed in order avoid the effects of activity depression due to repetitive stimulation (Fig 4.12).

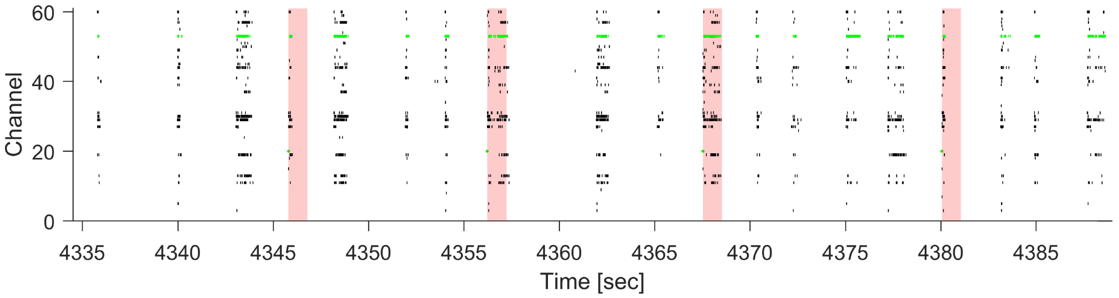

Figure 4.12. *Raster plot showing an example of fixed latency stimulation. Stimuli were delivered in a closed-loop setting after every 2 s in this example; response windows are shaded. Inter-stimulus intervals were not allowed to be less than 10 s. Spikes detected in the recording channel (52) and the stimulating channel at the time of stimulation (20) are highlighted in green.*

The response strength series exhibited a modulatory trend around its mean value (Fig 4.13).

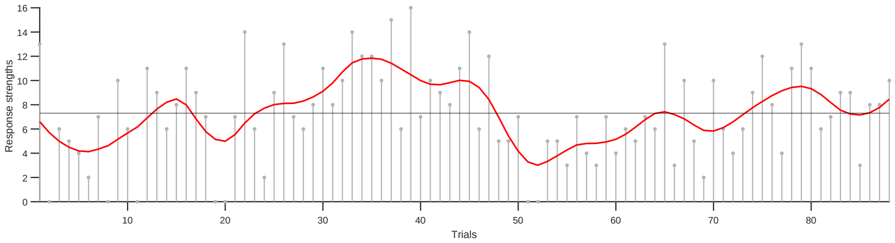

Figure 4.13. *Response strength time-series during the experiment. Black line denotes mean value and red line the time series smoothed using locally weighted linear regression (span=5).*

In order to understand the slow underlying pattern, we isolated strong and weak events and examined their temporal preferences. To this end, we defined the following indicator function:

$$I(n) = \begin{cases} 1 & \text{if } R[n] \geq \mu(R) + \sigma(R) \\ -1 & \text{if } R[n] \leq \mu(R) - \sigma(R) \end{cases} \tag{4.1}$$

The indicator function revealed the tendency of such events to cluster in time (Fig 4.14 top). The rate of occurrence of each event type exposes a slow alternating pattern over the 85 stimulus trials (20 minutes) in the session (Fig 4.14 bottom).

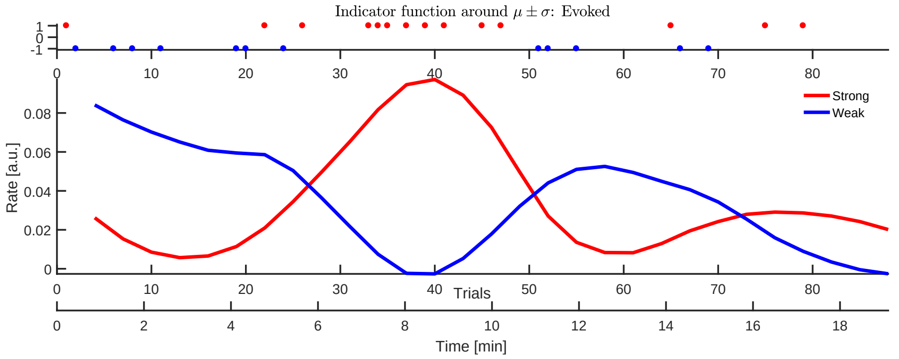

Figure 4.14. *An indicator function was defined to return ±1 if response strengths exceeded $\pm\sigma$ over the mean (top panel, in red/blue respectively). (Bottom) Detected events were binned and smoothed using a locally weighted regressor over sets of 3 trials each. The rate traces of each event type is plotted against time. Periods of higher and lower responsiveness emerge over a scale of tens of minutes.*

Pre-stimulus spontaneously occurring events were also categorized into strong and weak events using the same indicator function (cf. Eq. (4.1)). The rate traces of strong events, both spontaneous and evoked were overlaid in Fig 4.15. Over 20 minutes of the recording, a delayed complementary coupling in the dynamics of these measures persisted.

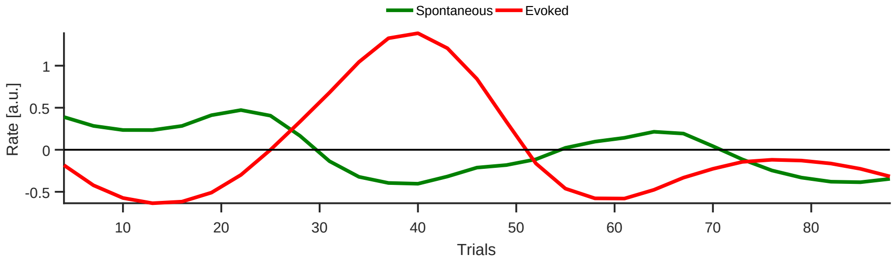

Figure 4.15. *The indicator function described in Fig 4.14 was extended also to spontaneously originating events relative to which latencies were computed. The resulting rates were superimposed on the 'strong event' rate trace from Fig 4.14. The traces are indicative of a delayed complementary coupling in the dynamics of these measures.*

To study the long-term stability of the coupling of response strengths with spontaneously generated events, the network was stimulated in closed-loop with a pre-set latency of 3 s over ≈ 3 hours (Fig 4.16). The left panel shows spontaneous events in the recording channel 5 s preceding the stimulus; the right panel shows the response window (500 ms post-stimulus) in a logarithmic time scale.

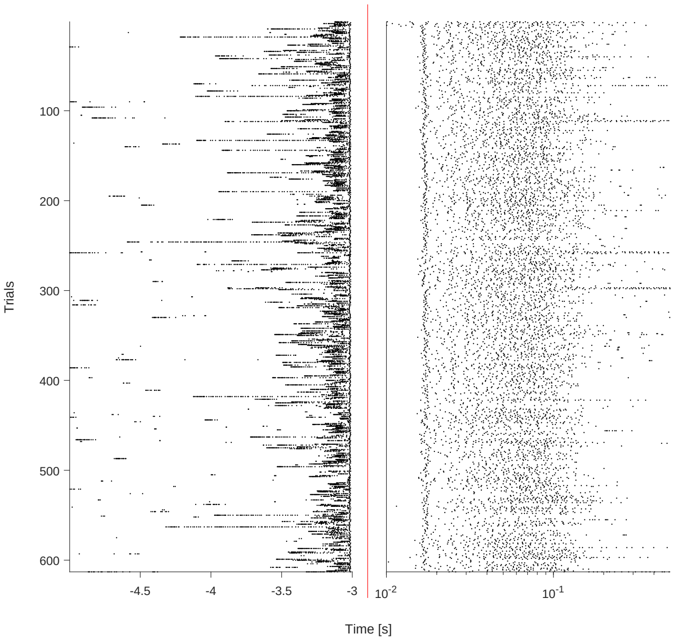

Figure 4.16. *The fixed latency paradigm repeated over longer time scales to study the stability of the coupling of response strengths with spontaneously generated events close to the stimulus. Raster plot shows trials delivered over approx 3 hours with a constant latency of 3 s. The left panel shows spontaneous events in the recording channel 5 s preceding the stimulus which was delivered at t=0 (red line). The right panel shows the response window (500 ms post-stimulus) in a logarithmic time scale.*

Fig 4.17(bottom) shows a time-resolved cross correlation of the two traces were performed over 7 minute sliding windows ($dt = 0.25$ minutes) and a maximal lag of ±2.5 minutes. For each stimulus, only the closest spontaneous event was considered.

Around zero lag, between 20 and 40 minutes, traces were mostly positively correlated. This trend changed thereafter to negative correlations for the next 40 minutes. Another half an hour stretch of positive correlation re-emerged 90 minutes into the recording; negative correlations followed thereafter. Clearly, the coupling of evoked and spontaneous events exhibited dynamics over long time-scales.

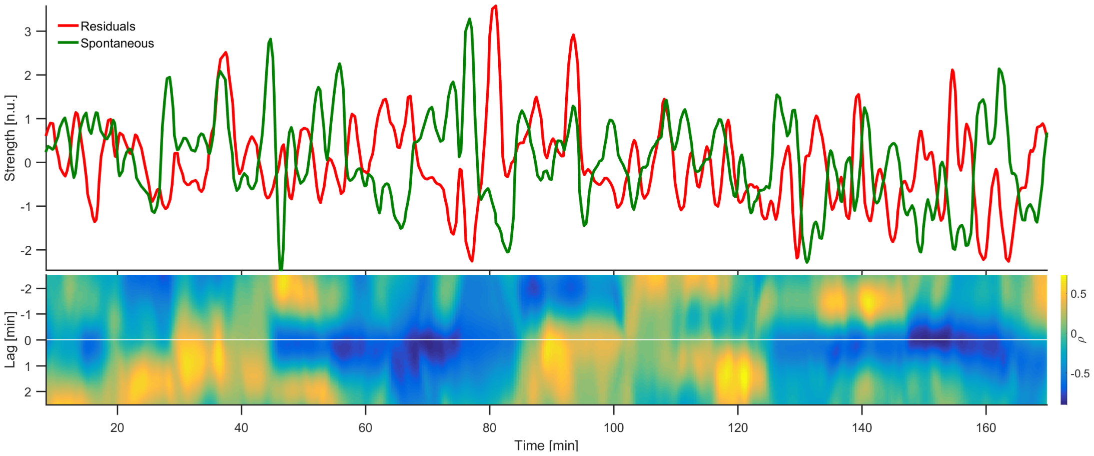

Figure 4.17. *Smoothed traces of response residuals (red) and spike counts in spontaneous events (green) shown for the entire stretch of the recording, normalized (top). For each stimulus, only the closest preceding spontaneous event was considered. Time resolved cross-correlation of standardized snippets of the two traces show alternating periods of broad positive and negative correlations in the time scale of hours, indicating non-stationary system behaviour.*

As described in the beginning of the section, a minimal inter-stimulus interval of 10 s was imposed during the fixed-latency experimental paradigm. This resulted in multiple spontaneously generated events elapsing between successive stimuli. A case could be made that in considering only one spontaneously originating event in the history of each stimulus, numerous other preceding events of varying rates and strengths would have to be ignored. How much of an impact does this selective sampling create in assessing the local excitability of the network?

To understand this, we computed the history relationship of response strengths including all spontaneously generated events and timings (Fig 4.18). The time-resolved cross correlation was performed over 10 minute sliding windows ($dt = 0.25$ minutes) and a maximal lag of ± 2.5 minutes. Interestingly, the dynamic structure of the coupling between the traces was largely comparable with Fig 4.17. Further, periods of positive and negative correlations and switches between them seemed qualitatively more pronounced. This suggests that the immediate history of ongoing activity already captures substantial information about the trend of excitability in the network and may be sufficient to make local predictions on subsequent evoked event strengths.

Such temporally local correlations appeared in most of our data sets involving repetitive stimulation over longer time periods. The distribution of correlation coefficients around the zero lag of such correlograms failed the Hartigan's dip test for uni-modality in 6 out of 7 networks visually verified, indicating the presence of more than one locally stable correlation values in the time series.

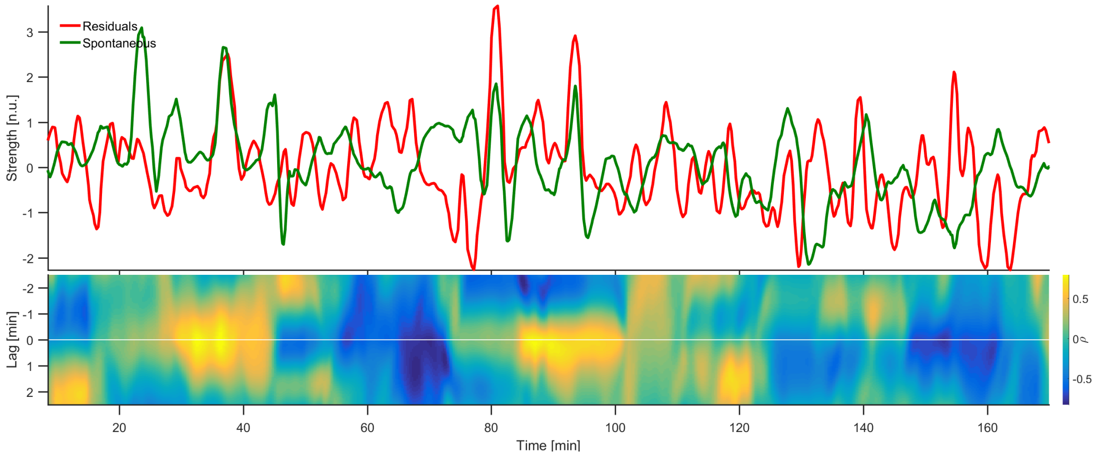

Figure 4.18. *Smoothed traces of response residuals (red) and spontaneous event strengths (green). Unlike in Fig 4.17, all events of spontaneous origin detected at the channel of interest were included in the green trace. Time resolved cross-correlation of the standardized snippets of the two traces show that the periods of broad positive and negative correlations are better distinguishable than in Fig 4.17.*

Our results revealed a slow modulation of stimulus-response relations during repetitive interaction with biological neuronal networks. Their characteristic time scale was widely distributed across networks. Ongoing activity also exhibited features that fluctuated over slow time scales and likely reflect the dynamics of the background process modulating stimulus-response relations. They could thus be useful as predictive indicators of successive responses. The relationship between ongoing activity features and response residuals was itself dynamic over long time scales, suggesting that the network mode may play a pivotal role. Temporary stationary behaviour was nevertheless observed over periods of tens of minutes. The slow nature of such dynamics suggests that it may be possible for RL based algorithms to adaptively learn such relationships.

We exploited these observations to formulate a toy problem: Can response strengths be clamped to a pre-defined value with no prior knowledge of the system dynamics? The problem allows us to capture the structure of a generic adaptive control problem involving biological neuronal networks and develop appropriate algorithms in a controlled setting.

Summary

- Stimulus-response relations in generic neuronal networks *in vitro* exhibit temporal fluctuations in slow time scales.
- Inhomogeneities were autoregressive suggesting history dependence.
- In addition, spontaneously occurring activity close to stimulus events may help predict successive responses.
- However, a stationary model of such dependencies did not exist.
- Our observations hint at locally persistent features that influence both ongoing activity and its relationship to external stimuli.
- We used this intuition to extend the RL based closed-loop framework to clamp response strengths to predefined values in the next chapter.

Chapter 5

Adapting to Temporal Inhomogeneities in Stimulus-Response Relations

A central challenge in neurotechnology is to develop adaptive stimulation solutions capable of operating safely and efficiently, notwithstanding the poorly understood dynamic floor of ongoing activity. We showed in Chapter 4 that, like networks in the brain, generic neuronal networks *in vitro* exhibit elements of stochastic fluctuations and long-term drifts in response features, the mechanistic origins of which remain unclear. In this chapter, we propose to extend our autonomous paradigm for adaptive control problems in biological neuronal networks (BNNs). Specifically, we explore the following questions: (1) How do we design controllers that adapt autonomously to the poorly understood dynamics of stimulus-response relations to remain goal-directed? (2) What are some factors that govern or limit the dynamic stability of the paradigm? (3) What are the implications for the development of safe, clinically relevant neurotechnological solutions?

5.1 Reinforcement learning (RL) framework for adaptive control of neuronal networks

Based on our findings that the history of multiple evoked and spontaneously generated event strengths preceding a stimulus could be information-rich features useful to predict the outcome of an upcoming stimulation (cf. Chapter 4), we designed an RL based autonomous controller to now clamp response strengths to pre-defined quantities by adapting to the ongoing activity dynamics in the network.

A high-dimensional state vector captured the network state in the temporal neighbourhood of the stimulus (see Appendix B for a detailed description). The action set was augmented to include up to three modalities: multiple stimulation sites, latencies relative to ongoing activity at each, and multiple stimulus amplitudes.

Online Q-learning was the learning algorithm used. To cope with the dimensionality of the problem, we used an approximate algorithm based on Least-Squares Policy Iteration (LSPI) to learn the Q-function as a linear combination of state and action features. See Section 2.4.2 in Chapter 2 and Appendix B for a detailed description.

5.2 Dynamic instabilities in the co-adaptive architecture

Feedback control schemes, although resilient to external disturbances and noise, are susceptible to feedback instabilities and runaway dynamics. Additionally, the intrinsic plasticity dynamics of biological neuronal networks and adaptivity of the RL controller introduces further interactions over multiple time scales, i.e. co-adaptivity. Understanding coupled dynamics and assessing feedback stability in such architectures is therefore cumbersome.

Stability alone may not necessarily ensure 'safety' in a neurotechnological context. The notion of 'safety' in this context may involve the imposition of dynamic constraints on the sequences and/or structures of intervention patterns to evade 'unsafe' states. Intervention patterns derived through the proposed autonomous learning approach are unpredictable as is their behaviour under rare conditions and are thus insufficient to guarantee safe operation. Given the lack of tractable models of relevant scale and complexity, it is currently non-trivial, if not impossible to formally estimate stability and safety of such architectures.

Nevertheless, assessments of this kind remain critical for the translation of such methods to the clinical domain. Here, we analyse several cases of failures and instabilities to identify failure classes and potential causes.

Stability in a neurobiological context is cumbersome to define formally. Stability assessments of our learning sessions were therefore qualitative and expressed in terms of the degree to which observed response strengths remained goal-directed, goal-averse or oscillatory.

An additional consideration when investigating the origin on instabilities is the co-adaptive nature of the architecture and the consequent causality dilemma. However in many cases, based on qualitative assessments of individual system dynamics, it was possible to attribute instabilities to specific aspects of the learning algorithm or the biological neuronal network. The origin of unstable dynamics observed in our experiments could in general be attributed to one of the following:

1. Network mode switches
2. Sharp non-linearities in input-output relationships
3. Delays in the learning loop

Such instabilities could potentially arise in any autonomous control scheme involving biological neuronal networks. In the following sections, we report examples illustrating instabilities of each class and use them to develop general strategies that could help address some of these issues.

5.2.1 Instabilities arising from switching network modes

Biological neuronal networks are known to switch between distinct network modes over long time scales. Such poorly predictable behaviour was especially problematic for learning based schemes since a policy learnt from interactions in one mode may not hold in another. The same target may not even be reachable in the new mode. Nonetheless, a policy now potentially sub-optimal and unsafe will continue to be executed until sufficient information has been gathered to adjust the policy to the new mode.

To test if such situations arise in our networks we probed the extent to which the controller′s performance was dependent on the ongoing mode in our networks. We present data from a network where switches in the network's activity mode were ostensible. We distinguished network modes post-hoc, from the analysis of first and second order statistics of ongoing spontaneous burst (SB) event strengths.

If our hypothesis on the ongoing network mode's influence on system controllability were true, we predicted that the time to achieve target would be significantly different over multiple sessions provided learning began with the similar initial conditions. Additionally, we expected the learned policy should break down during switches in activity modes.

Case 1

Spontaneous activity was monitored and found stationary during a short observation window (30 min) prior to the closed-loop session. Based on responses to open-loop randomly placed stimuli, stimulation sites, recording sites and an achievable target response strength was chosen (10 spikes at the evaluation electrode). A 12 episode closed-loop session with 100 trials each was performed (Case 1a).

At around 120 min, evoked response strengths were close to target (black line in Fig 5.1A, B). Over the remainder of the recording, achieved response strengths fluctuated around the target. This was likely due to the discrete latencies available to the controller (steps of 0.5 s) and delays in the learning loop.

The initial 150 trials (40 min) were dominated by action exploration. No particular latency was preferred at this stage (Fig 5.1C). Over the next 150 trials, certain latencies (3, 6, 7, 7.5 and 9 s) were preferred more often (white arrow marks in Fig 5.1C). During this time, response strengths, though close to target in terms of mean values, varied considerably (Fig 5.1B). In the next stretch (trials 300 – 380), a low latency policy was preferred producing a low response regime (green arrow mark in Fig 5.1C). Thereafter, at around 120 min, the controller switched to a policy that evoked response strengths close to target. Between trials 400 and 500, two latencies were mostly chosen. In the last phase (trials 550 onward, white line), the choice of stimulus latency was relatively distributed (Fig 5.1C). Despite this, no departure from desired levels was observed in terms of evoked response strengths. This suggested that the learned controller was able to adapt to ongoing activity patterns over short time scales to clamp response strengths

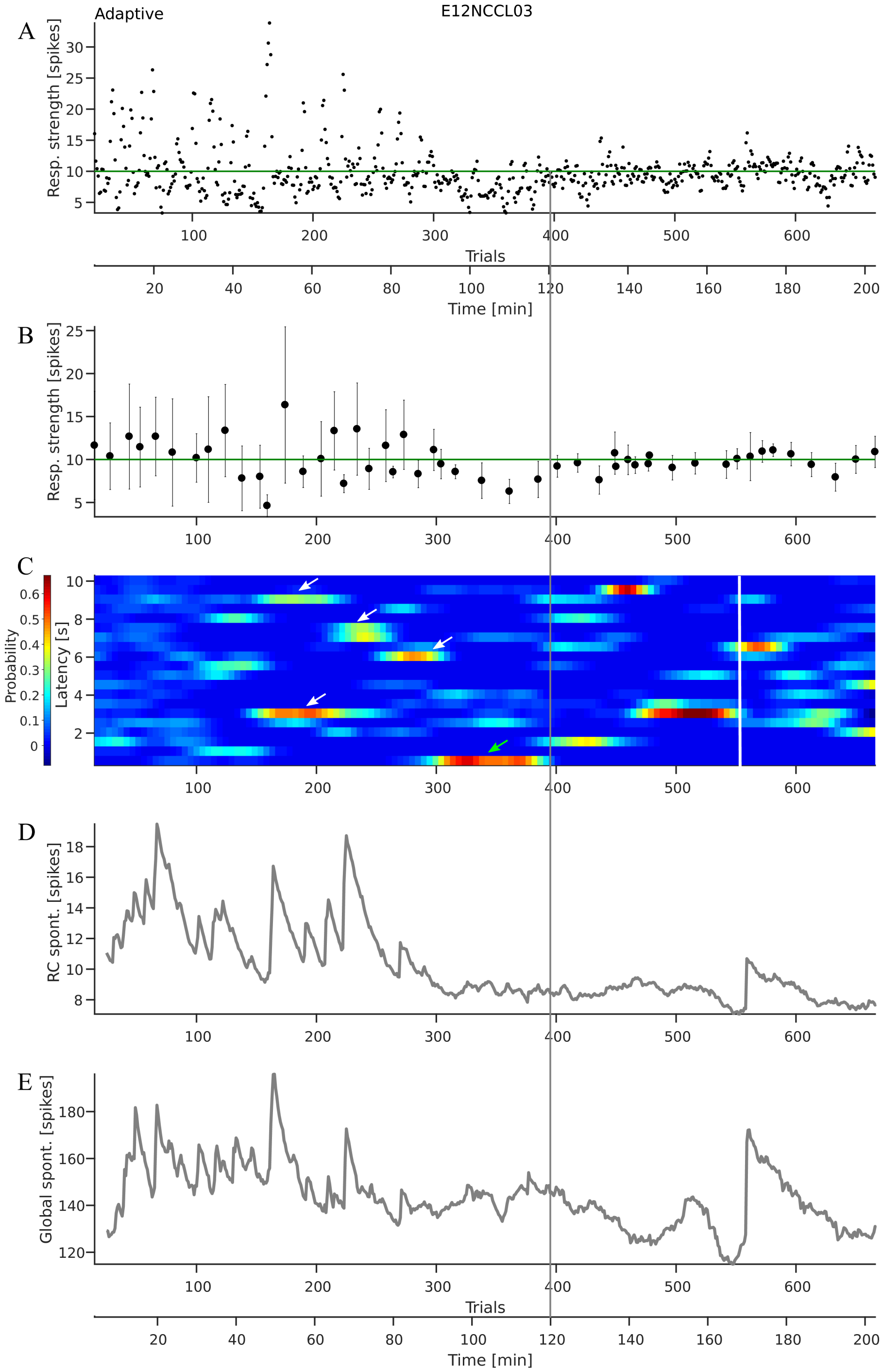

Figure 5.1. **Case 1a:** *(A) Response strengths over trials smoothed with a 5 sample moving gaussian window. (B) Data points in (A) binned ($\mu \pm \sigma$) every 4 min. The target response strength is indicated by the green line. The second x-axis indicates the time course of the session. (C) Evolution of the controller's choice of latencies during the session, computed using a sliding window of 10 trials each. Colour indicates the probability of choosing a latency in each such window. (D) Evolution of spontaneous event strengths in the network preceding each stimulus trial during the session. Both single channel (D) and global (E) spike count time series were smoothed with an exponentially weighted moving average technique ($\alpha = 0.12$).*

to desired levels.

One of the strengths inherent to a closed-loop strategy is its ability to characterize dynamical systems. The key question here is: what are the inputs required to hold the system output at a pre-defined level. The equivalent question in our experimental context would be: what stimulus latencies were chosen when response strengths were clamped to desired levels? In this session, response strengths from 120 min onward were considered clamped to the desired level. The evolution of average latencies chosen during this period is shown in Fig 5.2. Alternating periods with longer, resp. shorter latencies were conspicuous. Departures in and out of these states occurred over a period of tens of minutes and likely expose slow dynamics of excitability in the network, suggesting that this modulated recovery after spontaneous events. Yet the controller was able to follow this background process and clamp responses to a pre-defined level.

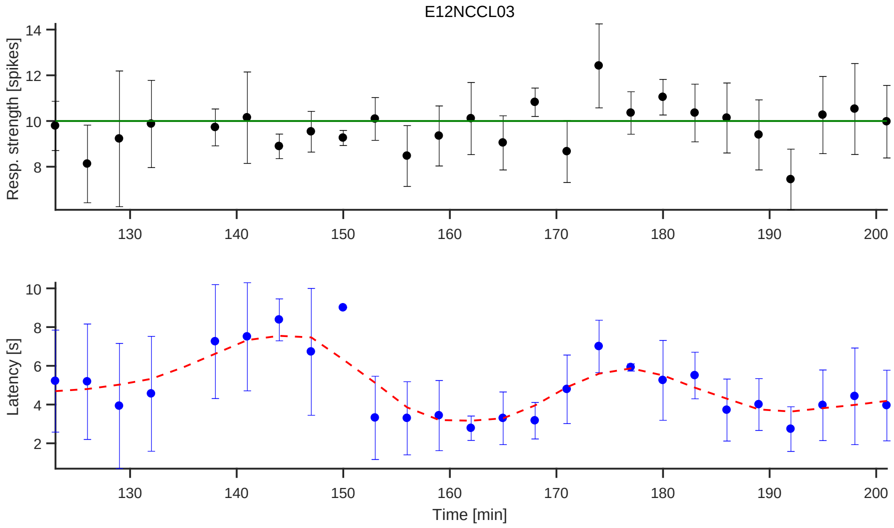

Figure 5.2. **Case 1a:** *Latencies chosen by the controller (bottom) during the period when response strengths were clamped to target (top). Both quantities were binned in 3 min non-overlapping windows. The mean ± std values in each bin are shown. The green line in the top panel indicates the target. Smoothed mean latencies reveal a slow fluctuating trend (red dashed line in bottom panel).*

Fig 5.1D–E shows the evolution of spontaneously occurring events over the session. The initial period was characterized by high variability in event strengths. As time progressed, activity transitioned into a distinctly different mode characterized by relatively stable responses in the recording channel, around the time at which the target was achieved. The network remained in this mode for the remainder of the session.

The controller learned to adaptively clamp response strengths to desired levels during the last hour of the session. However, distinct differences in the strength and variability of spontaneously arising bursts were evident. Moreover, periods with reduced variability of spontaneous events coincided with those where response strengths were close to target.

To test our hypothesis on the network mode's influence on system controllability, we ran another session on the network using the same stimulation electrode (SE), recording electrode (RE) and target with similar initial conditions.

A long-term closed-loop session with a re-initialized Q-function was performed on the network a day after (Case 1b). The session lasted close to 11 hours. The rewards received by the controller fluctuated over long time scales and reached a regime of higher rewards set in only from episode 22 (red arrow in Fig 5.3). Interestingly, the performance deteriorated later in the session.

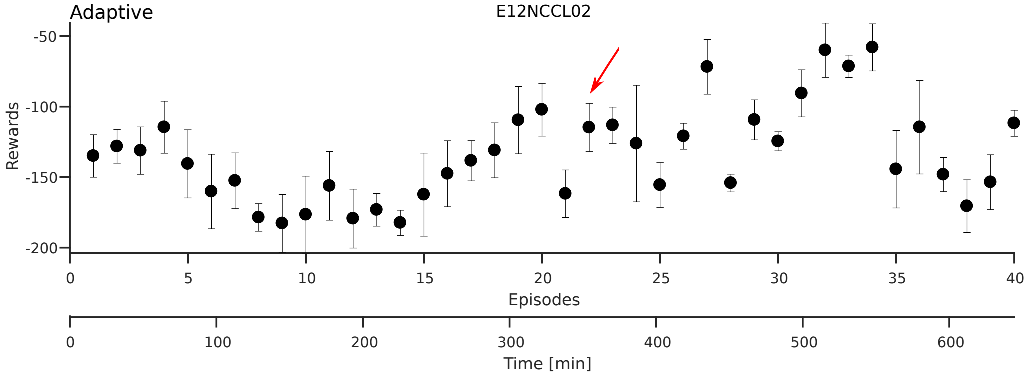

Figure 5.3. **Case 1b:** *Rewards ($\mu \pm \sigma$) received by the controller during each episode made up of 100 trials. The second x-axis shows the session's time-line.The reward was defined as the negative of the absolute error to target and was therefore ideally zero. From episode 22 onward (red arrow), rewards were generally higher.*

At $\approx$ 350 min, close to episode 22 (see Fig 5.3), mean response strengths reached close to target (line in Fig 5.4A–B). However, after a further 5 hour stretch, performance worsened at around 600 min (line in Fig 5.4A–B).

The initial 100 trials were dominated by action exploration (Fig 5.4C). At around 350 min, the controller evoked response strengths close to target. Around trial number 2200, a qualitatively different set of actions were executed. This period also corresponded to the departure from target, pointing to a possible switch in the network mode where

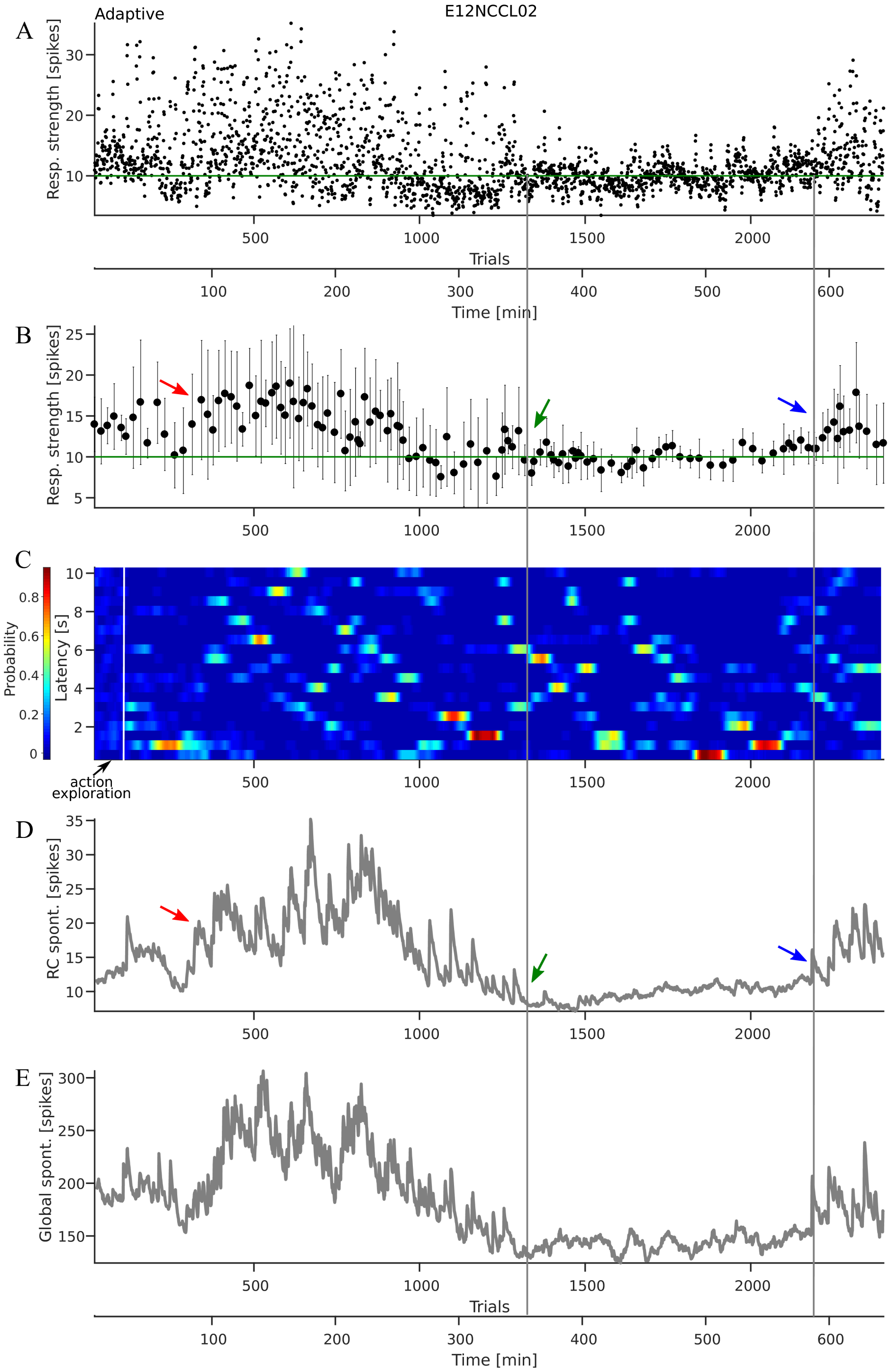

Figure 5.4. **Case 1b:** *(A) Response strengths smoothed with a 5 sample moving Gaussian window. (B) Data points in (A), binned ($\mu \pm \sigma$) every 5 min. The green line indicates the pre-defined target. (C) Evolution of the distribution of the controller's choice of latencies during the session, computed using a sliding window of 10 trials each. (D, E) Evolution of SB strengths in the network preceding each stimulus trial during the session. Both single channel (D) and global (E) spike count time series were smoothed with an exponentially weighted moving average technique ($\alpha = 0.04$). Lines indicate duration for which target response strength levels were achieved. Arrows highlight the correspondence of targeted interaction with features (strength and variability) of ongoing events in the network.*

the learned transfer characteristics of the network-stimulus interaction were no longer valid.

This interpretation is further supported by the qualitative correspondence of targeted interaction with the strength and variability of spontaneously originating events over the same duration (Fig 5.4D–E). Spontaneous event strengths at the evaluation channel (Fig 5.4D) closely followed trends in global events (Fig 5.4E). The initial period was characterized by weaker spontaneous events. This seemed amenable to learning and the naive controller approached target response strengths after 250 trials. Soon after, ongoing activity switched into a mode characterized by strong and fluctuating events during which the controller was unable to remain goal-directed (red arrow marks in Fig 5.5B, D). Hours later, as event strengths diminished and the network transitioned into another mode, the target was once again achieved (green arrow marks in Fig 5.5B, D). The network stayed in this mode for the next 5 hours of the recording. In the last hour of the recording, however, it slipped into the high variability mode and the learned strategy to clamp responses was no longer effective (blue arrow marks in Fig 5.4B, D).

The evolution of latencies the controller chose during the period where it was able to exert control of the network revealed a distinct underlying fluctuation dominated regime (Fig 5.5). Departures in and out of periods of high or low latency choices occurred again over a period of tens of minutes.

Taken together, Cases 1a and 1b show that drastically different durations (2 and 5 hours respectively) were required to achieve the same target in the same network across two closed-loop learning sessions. The learned strategy also ceased to perform well when the network switched into a different mode (Case 1b, Fig 5.4).

These observations suggest that network modes may indeed play a limiting role in the controller's ability to achieve pre-set targets. However, it is not clear what role, if any, such controllers could play in driving the network into a 'controllable' mode and preventing it from switching back to a high variance mode. To study the propensity of the network to switch between modes regardless of the stimulation strategy involved, we attempted a long term recording on the same network after disabling learning

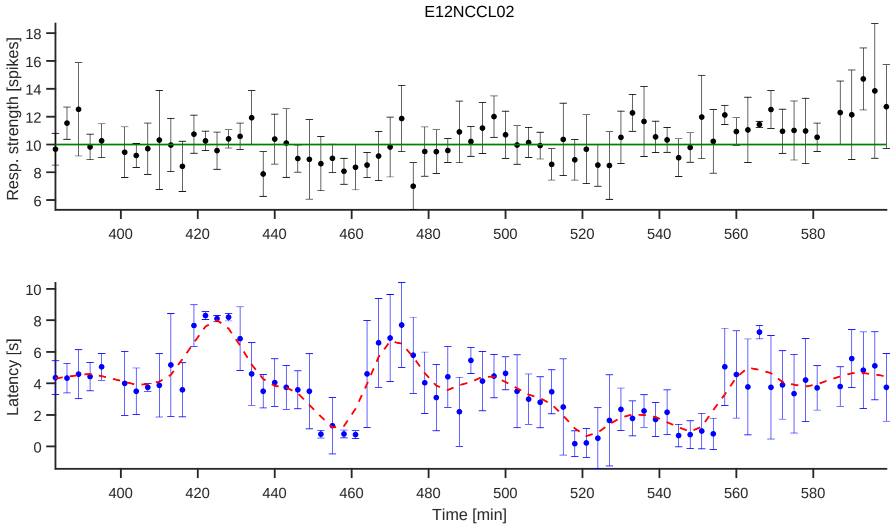

Figure 5.5. **Case 1b:** *Latencies chosen by the controller (bottom) during the period when response strengths were clamped to target (top). Both quantities were binned in 3 min non-overlapping windows. The mean ± std values in each bin are shown. The green line in the top panel indicates the target. Smoothed mean latencies reveal a slow fluctuating trend (red dashed line in bottom panel).*

(Case 1c).

We refer to sessions where learning was disabled as 'non-adaptive'. The choice of actions in these sessions remained random throughout. Fig 5.6C shows that the distribution of latencies chosen was relatively uniform across this 11 hour session. Fig 5.6A–B shows response strengths during the session.

A slow oscillatory rhythm spanning hours affected response strengths as well as their variability. The pattern was similar to the mode switches observed during closed-loop sessions. At around 200 min, response variability across trials decreased drastically. This 'low variability' mode persisted for the next 180 min, following which the network switched back to a high variability mode (lines in Fig 5.6). A subtle trend of progressively increasing excitability preceding the transition back into the high variability mode was also observed (arrow at $300 < t < 400$ min in Fig 5.6B, D).

However, the low variability mode during open-loop interaction was qualitatively distinguishable from that in the previously discussed closed-loop session. First, the low variability mode persisted for a relatively shorter duration compared to the closed-loop sessions. Second, there were no response strength fluctuations (ringing) as in the two closed-loop sessions discussed before (see Figs 5.1B, 5.4B). Further, the variance of response strengths during this mode was higher compared to clamped periods in

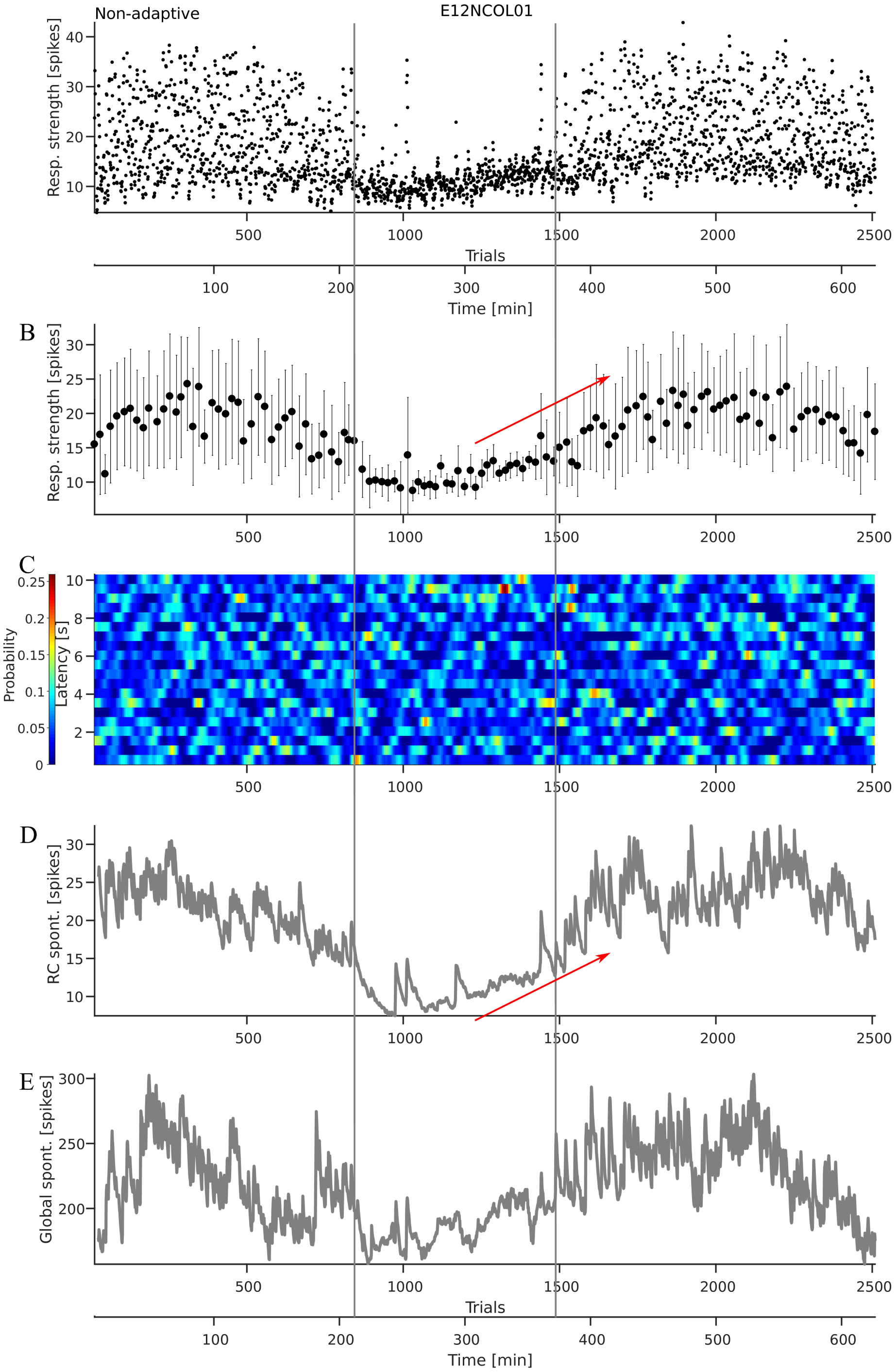

Figure 5.6. **Case 1c:** *(A) Response strengths over trials smoothed with a 5 sample moving Gaussian window. (B) Data points in (A) binned ($\mu \pm \sigma$) every 5 min. The second x-axis indicates the time course of the session. (C) Evolution of the controller's choice of latencies during the session, computed using a sliding window of 10 trials each. Colour indicates the probability of choosing a latency in each such window. (D) Evolution of spontaneous event strengths in the network preceding each stimulus trial during the session. Both single channel (D) and global (E) spike count time series were smoothed with an exponentially weighted moving average technique ($\alpha = 0.04$). Lines denote switches between high and low variability modes. Arrows indicate in the rise in excitability during a mode switch.*

closed-loop sessions (two-sample F-test, $p < 0.001$). These observations indicate that even in a 'favourable' network mode, an appropriate stimulation policy is essential to achieve targeted interaction.

Similar to the closed-loop session, signatures of network mode switching were observed in spontaneous activity traces as well (Fig 5.6D–E). Mirroring the response strength trend, excitability, seen as number of spikes per SB event, increased at the recording channel from trial 1100 onward (arrow in Fig 5.6B, D).

In summary, the presented data sets illustrated long term switches between activity modes and their potential influences on autonomous learning strategies.

5.2.2 Instabilities due to non-linearities in stimulus-response relations

Case 2

The control strategy was found to consistently fail to converge to a stable policy when the stimulus-response relations in the network was characterized by sharp non-linearities. The closed-loop strategy was applied to a network where the exponential recovery function fitting stimulus-response relations had a fast time-constant (a high value of λ) and thereby a step like behaviour during offline random stimulation (Fig 5.10). When stimulated, the network either failed to respond or responded with a high value of response strength. The target was set midway to the high value of the response strength.

The closed-loop session lasted for around 12 hours (Case 2a). Fig 5.7 shows the rewards received by the controller during the session. The reward showed near periodic fluctuations.

Around 100 min into the closed-loop session, response strengths approached target levels for the first time. After a further $\approx$ 30 min, responses returned to high levels. The pattern was found to recur with a period close to three hours.

Fig 5.8C shows the distribution of latencies chosen by the controller as the session progressed. Corresponding to every trough of the response strength sequence, note that lower latencies were preferred (marked in Fig 5.8C). Given the steep nature of the recovery function, such strategies would have resulted in response failures.

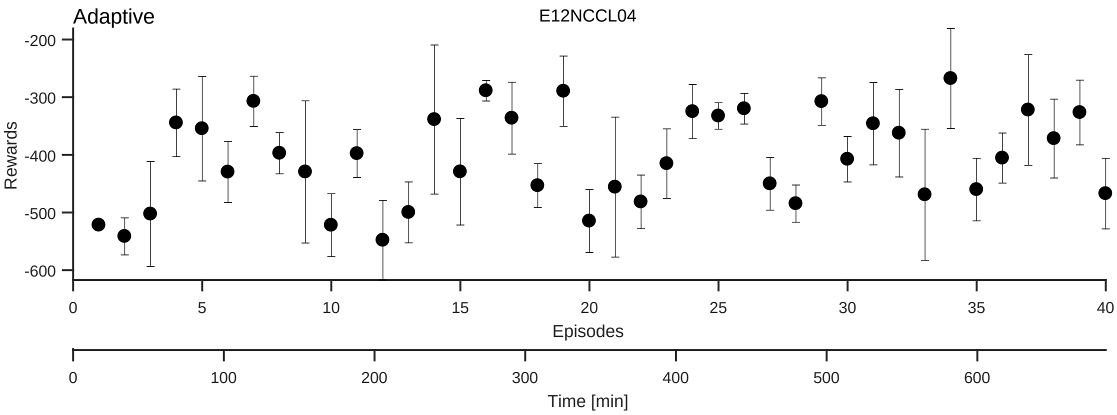

Figure 5.7. **Case 2a:** *Rewards ($\mu \pm \sigma$) received by the controller during each 100 trial learning episode. They fluctuated almost periodically with higher rewards regimes appearing $\approx$ every 10 episodes.*

Repeated punishments forced the policy to choose higher latency values which then yielded responses stronger than the target, thus explaining the oscillatory nature of the sequence.

We did not find qualitative shifts in the network mode as measured by spontaneous event strengths prior to each trial (Fig 5.8D–E). Though a weak fluctuating trend was observed in spontaneous event strengths in the recording channel, we did not find a consistent relationship with response strength fluctuations.

Weak fluctuations in spontaneous event strengths raised the question whether closed-loop interactions with the controller played a role in setting them up and if such fluctuations in turn played a role in the near-periodic appearance of troughs in the response strength sequence (Fig 5.8). One way to approach the problem was to ask if fluctuations in SB strengths and oscillatory response strength sequences were present also during non-adaptive stimulation with randomly chosen latencies.

A 12 hour long ‘non-adaptive’ session (Case 2b) revealed no oscillatory trends in the response strengths (Fig 5.9A).

However, fluctuations were still observed in SB event strengths at the RC (Fig 5.9B). This suggests that fluctuations in SB event strengths may not be related to the closed-loop intervention. It also suggests that the near-periodic ringing of the responses, observed only when learning was active, could be attributed to an unstable controller.

The closed-loop session described above, used LSPI for a linear approximation of the Q-function. This may not have been ideal, particularly when handling sharply non-

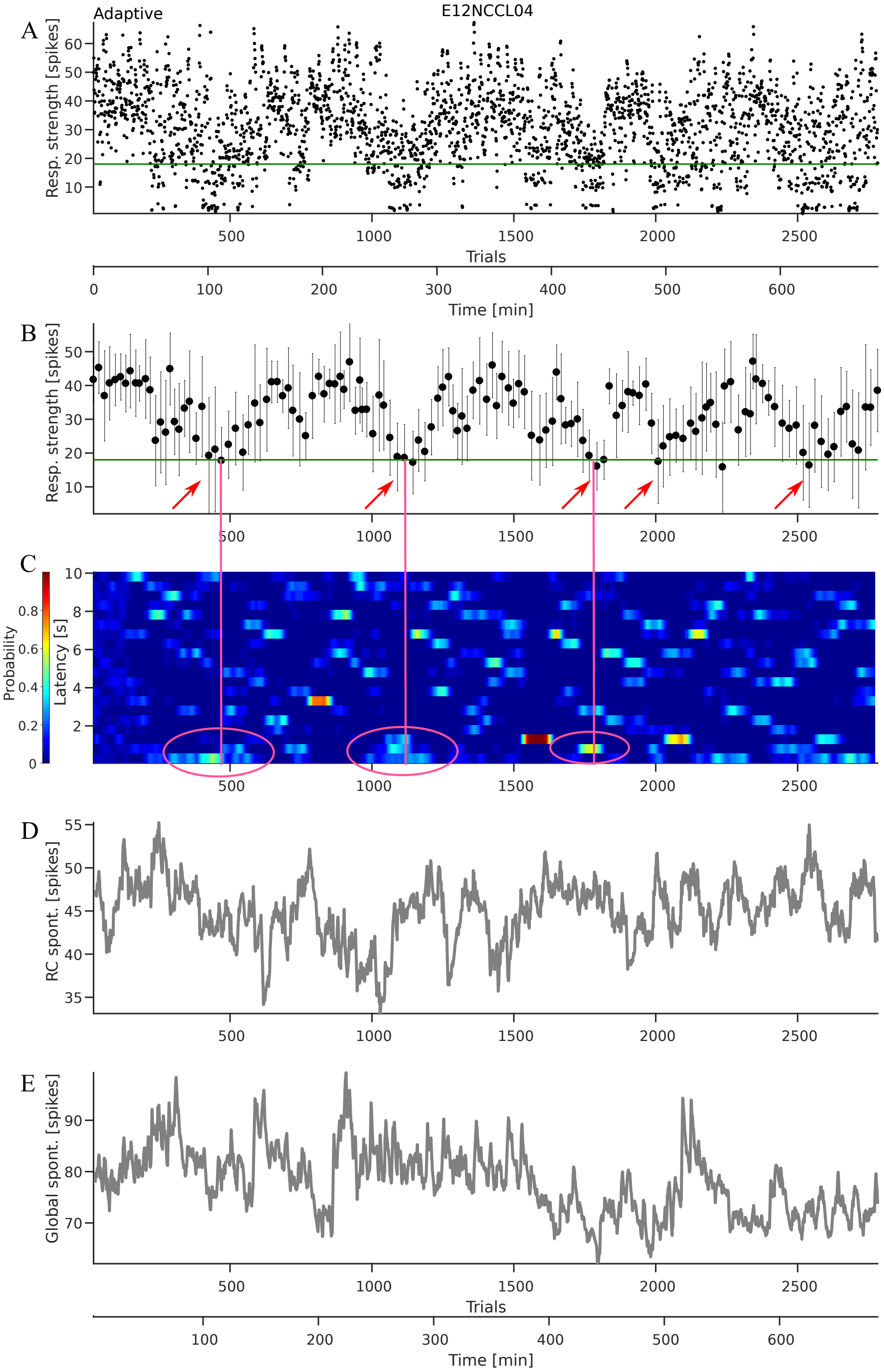

Figure 5.8. **Case 2a:** *(A) Response strengths over trials smoothed with a 5 sample moving gaussian window. (B) Data points in (A), binned ($\mu \pm \sigma$) every 5 min. The green line indicates the goal. Red arrows indicate troughs in the response strength sequence. (C) Evolution of the distribution of the controller's choice of latencies, computed every 10 trials (sliding window of 3 trials). The preference for lower latencies during troughs is marked in magenta. (D) SB strengths (spike counts) in the network preceding each trial at the evaluation channel (D) and array-wide (E). Both time series were smoothed with an exponentially weighted moving average technique ($\alpha = 0.04$).*

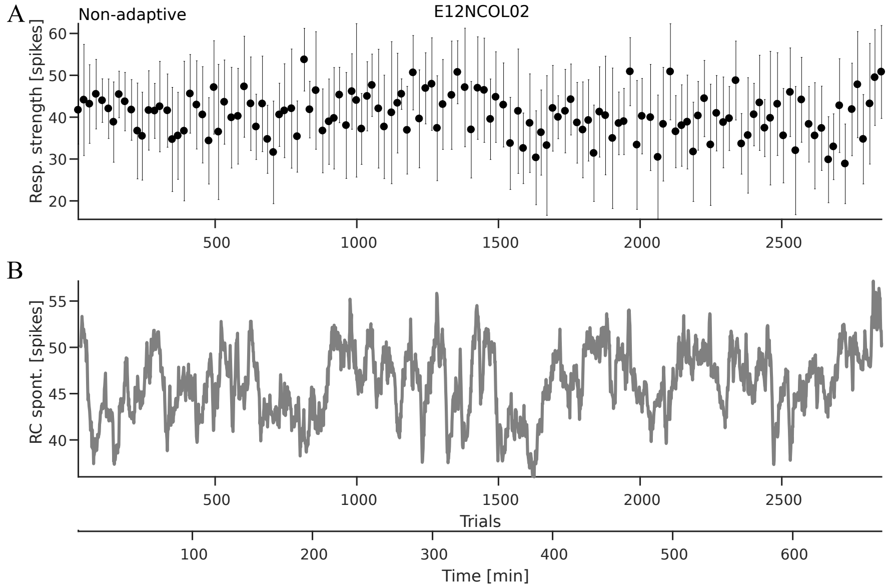

Figure 5.9. **Case 2b:** *(A) Response strengths smoothed with a 5 sample moving Gaussian window and binned ($\mu \pm \sigma$) every 5 min. Response strengths were relatively stable over the entire 12 hours of the session. (B) Evolution of spontaneous event strengths in the network preceding each stimulus trial during the session at the evaluation channel. The spike count time series was smoothed with an exponentially weighted moving average technique ($\alpha = 0.04$).*

linear input-output relationships (Fig 5.10). To test if non-linear function approximators could help in such situations, we asked if Neural Fitted Q-iteration (NFQ) could offer better performance.

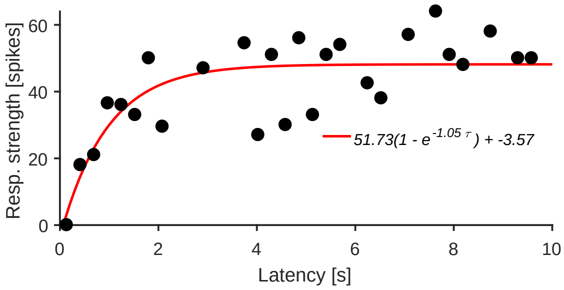

Figure 5.10. *The sharp non-linear stimulus-response relationship in this network. An exponential model of the form $A(1 - e^{-\lambda\tau}) + B$, was fitted (red) to binned response strengths (see Chapter 2). The time constant $\lambda > 1$, points to a particularly fast post-burst recovery in this network.*

Using the same stimulus and recording sites and target response strength, we performed another closed-loop learning session lasting ≈ 11 hours using NFQ as the learning algorithm (Case 2c). Fig 5.11A shows the rewards the controller received during the session. The reward structure though initially fluctuating – as before when the learning proceeded using LSPI – progressively dampened and settled from episode 28 onward to a stable regime of higher rewards. A similar trend was observed in the

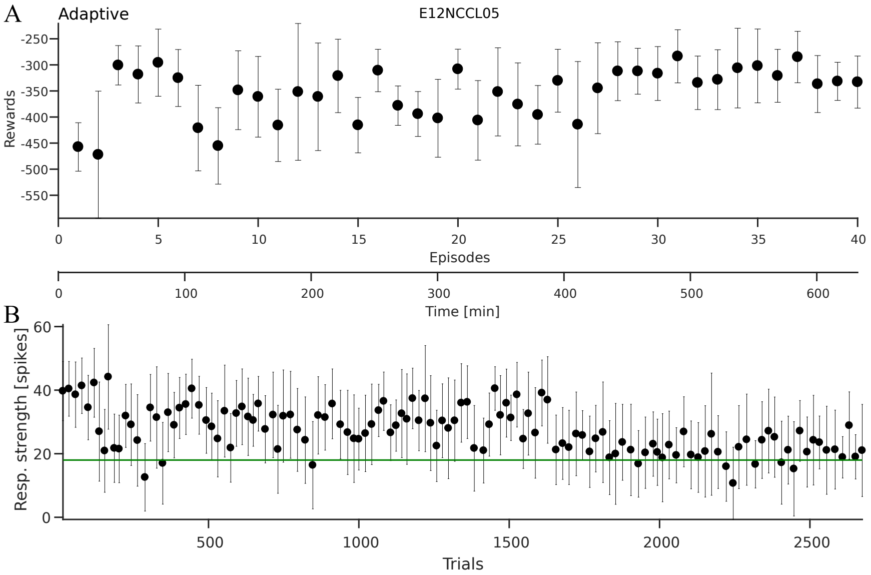

Figure 5.11. **Case 2c:** *(A) Rewards received by the controller during each learning episode consisting of 100 trials. The second x-axis shows the session's time-line. After weak initially fluctuations, the rewards settled down at around episode 28 to a relatively stable regime of higher rewards. (B) Response strengths smoothed with a 5 sample moving Gaussian window and binned ($\mu \pm \sigma$) every 5 min.The green line indicates the pre-defined target response strength. From $\approx$ 400 min onward, response strengths transitioned to values closer to the target.*

binned response strength sequence (Fig 5.11B). At $\approx$ 1 hour into the recording, mean response strengths were closer to target but departed soon after. Response strengths stayed at a higher value for the next 5 hours. However, from 400 min onward, responses transitioned to values closer to target and remained thereabouts for the remainder of the session.

In summary, these sessions demonstrated that the nature of input-output relationships was an important factor to consider in the development of stable autonomous control strategies.

In all networks described heretofore, the available action set was one-dimensional. Only stimulus latencies could be manipulated by the controller. Overall, 11 sessions were run with one-dimensional actions sets involving 8 networks. Three of them were characterized by mode-switches (see Case 1). Three sessions exhibited oscillatory instabilities due to sharp non-linearities (Case 2). In three sessions, the target was achievable by random actions and hence learning did not proceed. In the remaining two sessions, target response strengths were not achieved. The last two categories indicated lack of controllability due to limitations in either observability of the network state or accessibility of the outputs given the action set (see Section 5.3 and Appendix C.1).

5.2.3 Instabilities due to action dimensionality and learning delays

An formidable challenge for neurotechnological devices in a clinical context is the size of the parameter space to be explored. With open-loop devices, stimulator programming involves a highly trained clinician heuristically exploring a large action space (stimulation site, frequency, amplitude pulse-width etc.). The ability to navigate high-dimensional action spaces and converge quickly to optimal policies is therefore extremely desirable in an autonomous paradigm. High dimensional actions spaces pose a challenge to current RL algorithms, many of which do not scale gracefully with dimensionality.

To include this aspect of the challenge in our model, we added more action dimensions to the controller. Based on the hypothesis that response strengths may be dependent not just on temporal relationships to the previous network event, but also spatial considerations, i.e. pathways recruited for propagation of activity in the network, we now included multiple stimulus locations and latencies at each to the set of actions available to the controller.

Additionally, to compare random against learned strategies, experiments started with an initial 'non-adaptive' period where random actions were delivered. Response-strengths during this phase was monitored and used to learn the Q-function. When the open-loop constraint was removed, the controller transitioned into a closed-loop learning based policy where it would execute optimal actions based on observations made thus far. Further along the session, learning proceeded with a forgetting scheme, i.e. older samples were discarded as and when newer experiences were acquired.

Given the large action set that was available for exploration, to make efficient use of the collected data and stabilize the learning process we performed batch RL (Lange et al., 2012). The Q-function was updated in batches of 100 trials each. However the procedure also introduced a delay of $\approx$ 30 min between observing the system and learning from interactions. The choice of permissible delays may play an important role in the dynamic stability of the paradigm since input-output relations were found to fluctuate over similar time scales. LSPI was used as the learning algorithm.

The high dimensionality of the action set and the attendant learning delays resulted in an oscillatory instability that was reproducible across networks. We report here, two such cases.

Case 3

A 7 hour long session with a target response strength at the evaluation site set to 5 spikes was performed. The experiment proceeded in batch RL mode with each episode comprising 100 trials.

During the random phase, mean response strengths were already close to the predefined target (Figs 5.12A, 5.13A). At around 140 min, the experiment switched to the learning based adaptive phase. In this phase, though response variability was found to decrease slightly, mean responses fluctuated around the goal, pointing to a system instability (Fig 5.13A).

As the controller transitioned into the adaptive phase, certain latencies and stimulation sites were preferred (Fig 5.12B–C). The latencies chosen by the controller over time revealed a pattern where the smallest latency (0.5 s) was repeatedly preferred approximately every 50 min. These periods also corresponded to troughs in the response strength time series (see solid magenta lines in Fig 5.12A–C). Crests in the response strength sequence were correlated with the choice of stimulation site 83 (see dashed magenta lines in Fig 5.12A–C). These observations suggest that it was the controller's choices that mediated fluctuations in response strengths. The repeated choice of poor actions points to an unstable closed-loop configuration.

The evolution of SB strengths did not resemble the oscillatory trends of the response strength time series (Fig 5.12D–E). This supports our contention that in this session, it was the adaptive controller and the policy it learned, and not changes in ongoing activity modes that mediated excursions in the response strength sequence. The ineffective policy could be attributed to the long delays in the learning loop and the use of linear approximations for potentially non-linear mappings. The example illustrates the susceptibility of our paradigm to feedback instabilities.

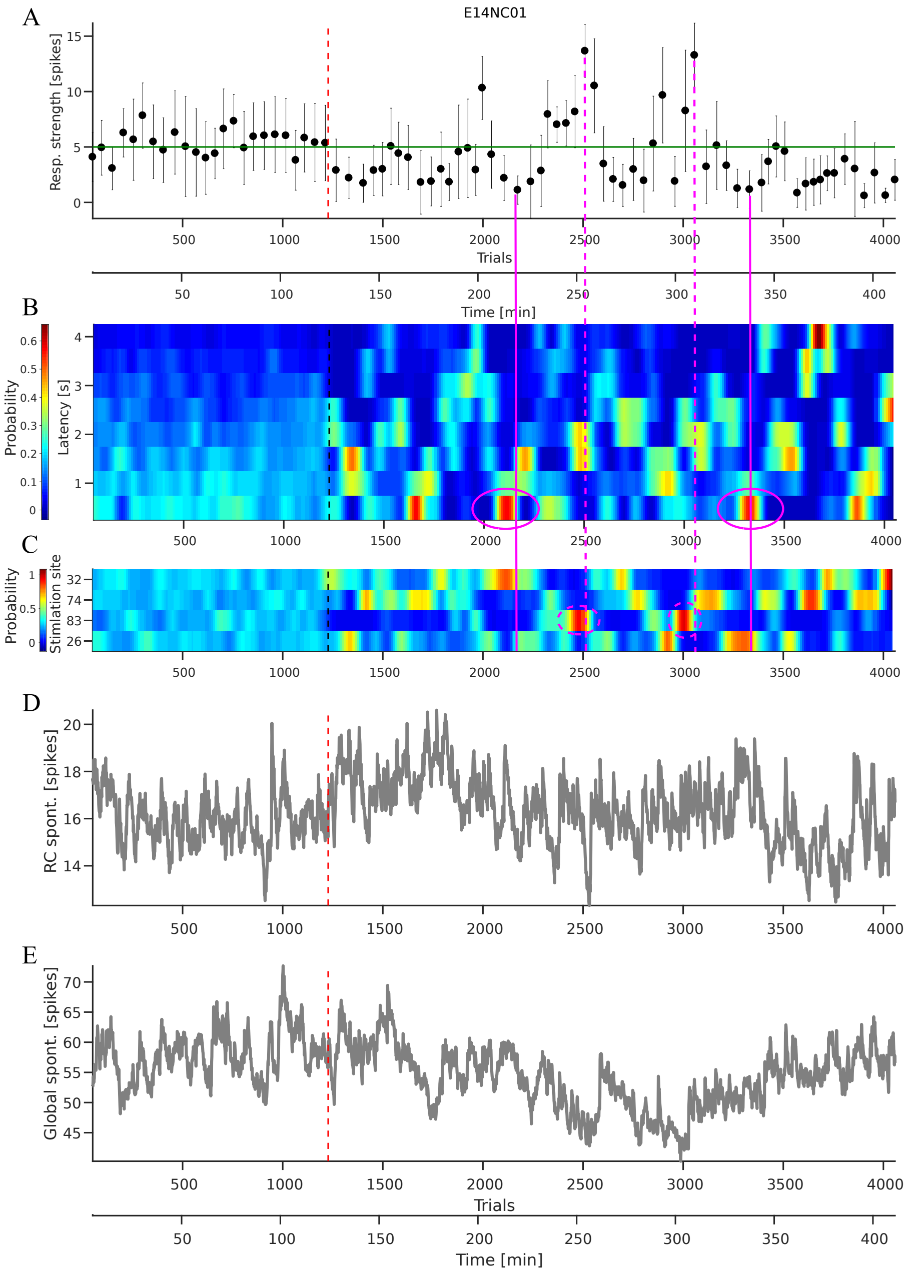

Figure 5.12. **Case 3:** *(A) Response strengths smoothed with a 5 sample moving Gaussian window and binned ($\mu \pm \sigma$) every 5 min. The green line indicates the goal. (B, C) Evolution of the distribution of the controller's choice of latencies (B) and stimulation sites (C) during the session, computed using a sliding window of 10 trials each. The preference of lower latencies during troughs (solid lines) and site 83 during crests in the response strength sequence (dashed lines) are marked in magenta. (D, E) SB strengths preceding each trial during the session. Both single channel (D) and global (E) spike count time series were smoothed with an exponentially weighted moving average technique ($\alpha = 0.04$). Dashed line in panels A–E indicate the end of the non-adaptive phase.*

Case 4

We report another example of an unstable controller under similar settings. Mean response strengths were already around target levels during the random phase (Fig 5.14A).

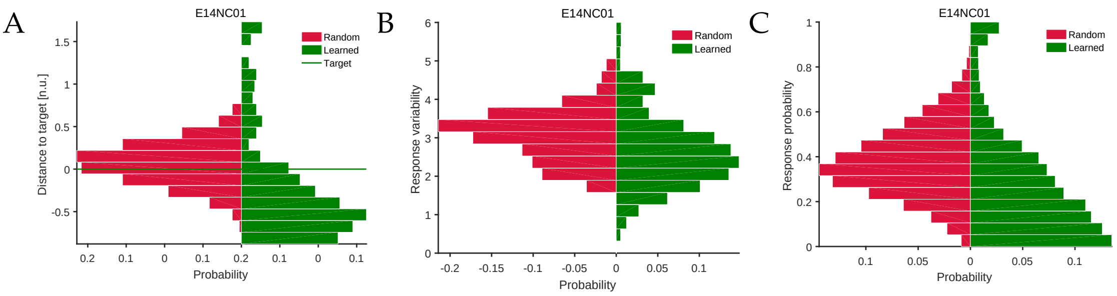

Figure 5.13. **Case 3:** *Distributions of the (A) means and (B) standard deviations of response strengths during random and learned strategies. Response strengths were binned every 3 min during the recording session (see Fig 5.12A). Normalized distance to target (A) was obtained by subtracting and dividing with the target response strength. While variability was significantly lower, the mean response strengths were off target after learning (green line in A). (C) The response probability distribution before and after learning showed that response failures, possibly due to poor choices of stimulation policy, were more common after learning.*

At $\approx$ 210 min, the controller switched to the learned adaptive phase (dashed line in Fig 5.14A). As in Case 3, response variability decreased, though mostly in periods of low responsiveness (Figs 5.14A, 5.15B). Mean responses weakened and fluctuated over a period of hours with recurring windows of response failures, likely due to a lousy stimulation policy (Fig 5.14A).

Figs 5.14B-C shows the evolution of action selection, i.e., latencies and stimulation sites chosen by the controller. In the adaptive phase, certain actions were preferred by the controller. Troughs in the response strength sequence corresponded to the choice of stimulation site 46 (marked in magenta in Fig 5.14A–B), suggesting that the controller's choices likely mediated fluctuations in response strengths. However, unlike in Case 3, such an interpretation is confounded by the fluctuations observed in SB event strengths at the RC in the adaptive phase of the session (Fig 5.12D). Troughs in the response strength sequence corresponded to crests in the SB strength sequence, indicating that the ongoing activity mode may also have a role to play (magenta lines, Fig 5.14A–D). The example illustrates the causality dilemma inherent to analyses of coupled co-adaptive schemes.

In summary, experimental sessions that progressed with high dimensional action sets and delayed learning were dominated by a trend of dynamic instabilities during closed-loop learning. Overall, 21 networks were studied using two-dimensional action sets and large learning delays. Distinct oscillatory excursions in response strengths were found in all 14 networks where sustained closed-loop interaction was viable. In the rest

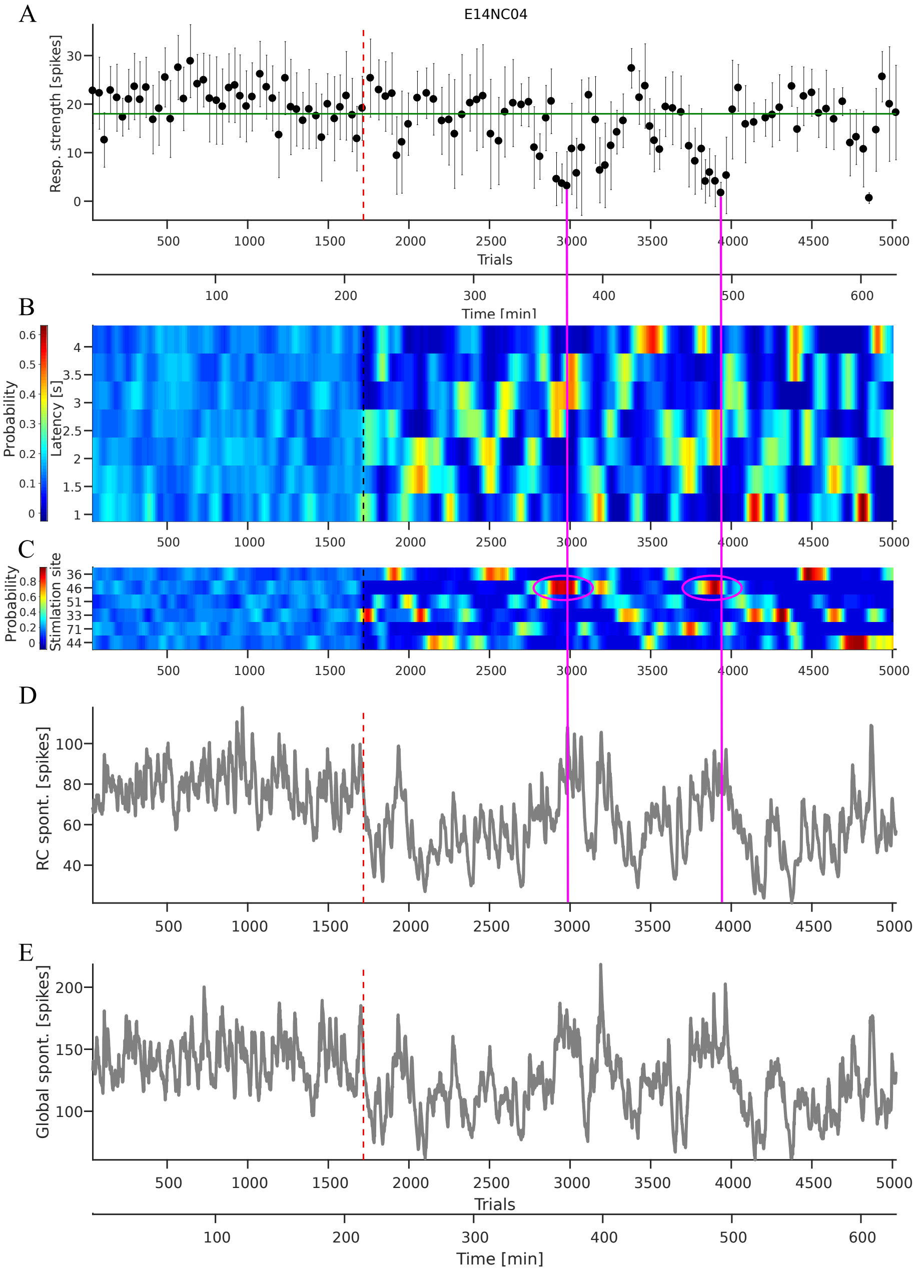

***Figure 5.14.* Case 4:** *(A) Response strengths smoothed with a 5 sample moving Gaussian window and binned ($\mu \pm \sigma$) every 5 min. The green line indicates the goal. (B, C) Evolution of the distribution of the controller's choice of latencies (B) and stimulation sites (C) during the session, computed using a sliding window of 10 trials each. Site 46 was preferred during troughs in the response strength sequence (marked in magenta). (D, E) SB strengths preceding each trial during the session. Both single channel (D) and global (E) spike count time series were smoothed with an exponentially weighted moving average technique ($\alpha = 0.04$). Dashed line in panels A–E indicate the end of the non-adaptive phase.*

activity switched to a regime characterized by multiphasic synchronized bursting (n=3) or responsiveness of the RE was lost (n=4; see Fig 5.26 and Appendix C.2).

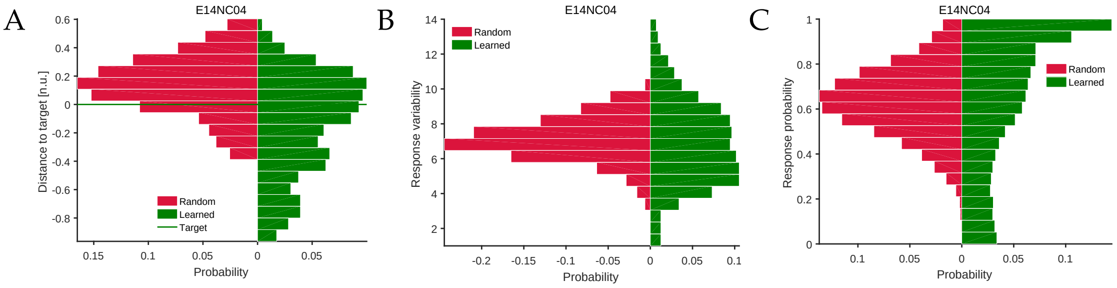

Figure 5.15. **Case 4:** *Distributions of the (A) means and (B) standard deviations of response strengths during random and learned strategies. Response strengths were binned every 5 min during the recording session (see Fig 5.14A). Normalized distance to target (A) was obtained by subtracting and dividing with the target response strength. While the variability distribution flattened, the mean response strengths turned multi-modal due to the recurring periods of low or no responses (see Fig 5.14A). (C) The flattening of smaller probabilities in the response probability distribution indicates that response failures were frequent after learning.*

5.3 Observability and controllability issues

The stability issues discussed heretofore were not the only factors that mediated the performance of our controllers. In some of the networks studied, the target response strength was unachievable.

A formal determination of system observability or controllability are generally not feasible for control problems involving neurobiological networks. The model-system was thus assumed observable and controllable. In networks for which the controller was unable to achieve the goal, it is likely that these assumptions were violated. Learning impaired closed-loop sessions could be broadly classified into the following categories.

When pre-defined target response strengths were already achievable with a random strategy, learning was impaired due to insufficient discriminability among actions. Fig 5.16 shows two such examples.

Temporal drifts and loss of activity in the sole recording channel also hampered network observability in some of our networks. In such cases, the controller continued to execute random policies presumably due to the lack of meaningful rewards.

Another limiting factor was the lack of sustained accessibility of the target state given the limited repertoire of actions available to the controller. In a few cases, the target was found consistently unattainable using the action space available to the controller (Fig 5.17).

In other cases, after hours of interaction, certain stimulation sites were no longer

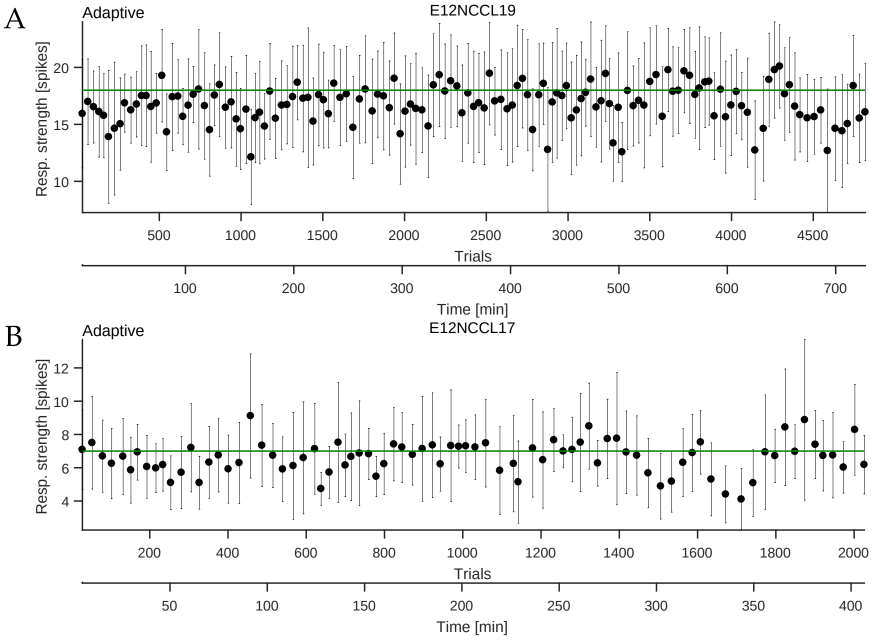

Figure 5.16. *When the target was achievable with a random strategy, the controller was limited in its ability to improve the stimulation policy. In both panels, response strengths ($\mu \pm \sigma$) recorded over 5 min bins are shown. The green line indicates the target.*

able to evoke responses, essentially limiting the controller's accessibility to the network.

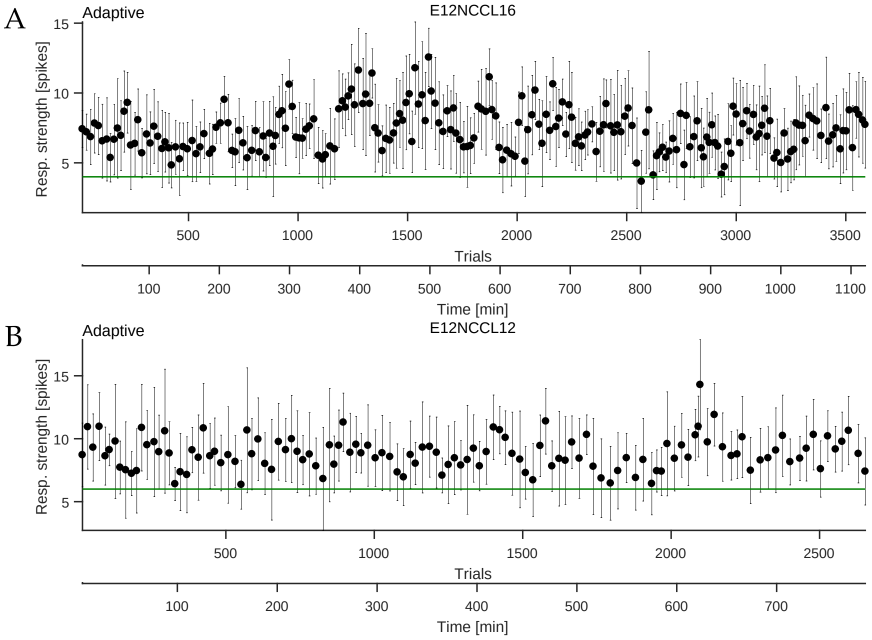

Figure 5.17. *When the target was not achievable with the available actions, the controller was limited in its ability to improve the stimulation policy. Data as in Fig 5.16.*

5.4 Stable high-dimensional adaptive control of response strengths

Though autonomous approaches could help optimize interactions with biological neuronal networks (see Chapter 3), ensuring dynamic stability of such interactions over long periods of time remains a challenge. The failure classes discussed thus far illustrate various aspects of this challenge and raise the question of whether maintaining dynamic stability in such a co-adaptive architecture is feasible at all.

Stability, in a formal sense is impossible guarantee for such a paradigm, given the absence of a computational description of the system's dynamics. Nonetheless, since the instabilities discussed in our case studies could be attributed to specific aspects of the co-adaptive system, we asked if it was possible to address them individually by suitably adapting our learning algorithms. We followed a failure driven algorithm development strategy assuming that the biological conditions were favourable (i.e., 'control-friendly network mode', sustained observability and accessibility to the thorough the chosen sites and so on). Data from networks where these conditions were met were not analysed.

To dissociate the controller from possible slow mode-switches in ongoing activity as discussed in Case 1, we introduced a forgetting scheme. Specifically, the agent retained the memory of transitions only close to the current time instead of the entire history. In addition, we used non-linear methods to approximate the Q-function (see Section 2.4.2 and Appendix B for a detailed description). NFQ had already shown promise compared to linear methods in dealing with sharp non-linear interactions in a previously discussed case (cf. Case 2). Further, to mitigate the effects of delayed learning on closed-loop performance (cf. Cases 3 and 4), we shortened our update cycles.

Along with the higher dimensional action sets we included one or more dud-sites – channels that were manifestly ineffective to evoke responses in the network. These served as a sanity check for the learning algorithms while also contributing to clearly distinguishing between naive and learned policies by making the set of effective actions a sparser subset of the available ones. Here we report two cases in this category (Cases 5 and 6).

Case 5

A 16 hour long experimental session with a target response strength at a chosen site set to 12 spikes was conducted. The experiment proceeded with shorter learning episodes of 50 trials each. Improvement in goal-directed behaviour was observed as the learned controller was deployed after twenty episodes of random action exploration. Thereafter, from $\approx$ 400 min onward, response strengths remained around the target albeit with small oscillations.

Figs 5.18A shows the individual response strengths evoked during the session. During the random phase, mean response strengths were low (5 – 7 spikes) with many response failures (Fig 5.19C). From $\approx$ 250 min onward, the learned adaptive controller was deployed and performance improved until response strengths settled around the target (Fig 5.19A). Mild overshoot and subsequent oscillatory behaviour persisted for the remainder of the session. Notably, response variability did not change appreciably from one phase to the next (Fig 5.19B).

During the initial 800 trials (250 min), latencies and stimulation sites were randomly selected from the available choices (Fig 5.18C–D). As the controller switched to the learned adaptive phase, certain actions were found to be preferentially selected. A range of latencies between 2 – 4 s were chosen in the learned phase. When executing the learned policy, the controller relied mostly on two stimulus sites (42 and 53, Fig 5.18C). As the experiment proceeded, the choice shifted exclusively to channel 53. Channels 67 and 55 were dud-sites planted for a quick validation of the learned policy. During the learnt phase, these sites were avoided indicating the efficacy of the learned policy. Fig 5.18D–E shows the evolution of spontaneous events in the network prior to each action. The closed-loop phase had a depressing effect on global spike counts, although events at the recording site were slightly stronger.

Fig 5.20 shows the evolution of chosen stimulus latencies during the period over which response strengths were reasonably on target. Considerable fluctuations of the chosen latencies were observed. Particularly, during the last 100 min, a switch to a different policy was evident although there was no impact on the evoked response strengths, pointing to adaptive capabilities of the learning algorithm. Cross correlation

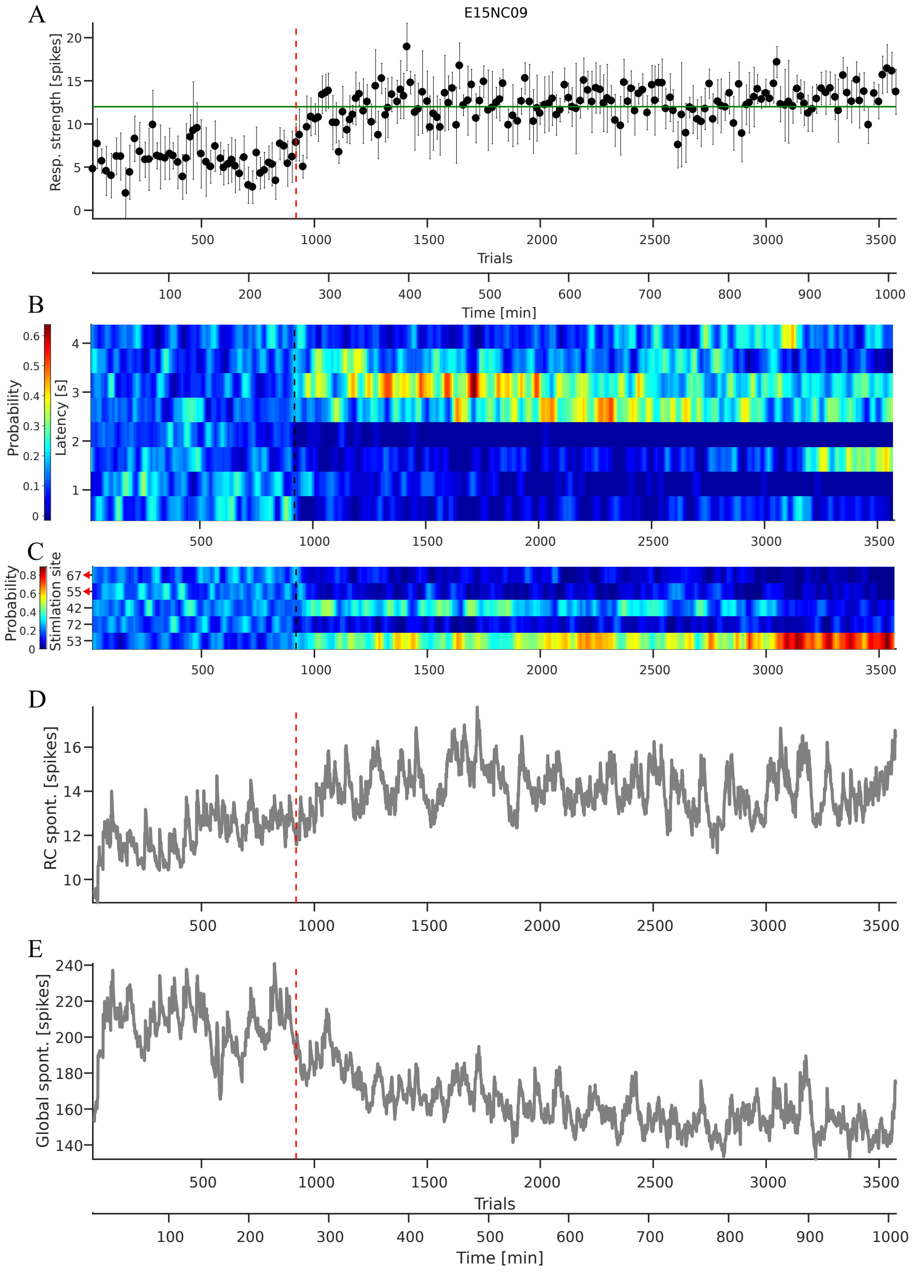

Figure 5.18. **Case 5:** *(A) Response strengths smoothed with a 5 sample moving Gaussian window and binned ($\mu \pm \sigma$) every 5 min. (B, C) Evolution of the controller's choice of latencies (B) and stimulation sites (C) during the session, computed using a sliding window of 10 trials each. Colour indicates the probability of choosing a particular latency or site in each such window. Red triangles indicate the designated dud-sites. (D, E) Evolution of spontaneous event strengths in the network preceding each stimulus trial during the session. Both single channel (D) and global (E) spike count time series were smoothed with an exponentially weighted moving average technique ($\alpha = 0.04$).*

of the mean latencies with the mean response strengths revealed a negative peak at a lag of ≈ 18 min. The negative value points to the ability of the controller to correct overshoots and modulations in response strengths and stay on target by choosing

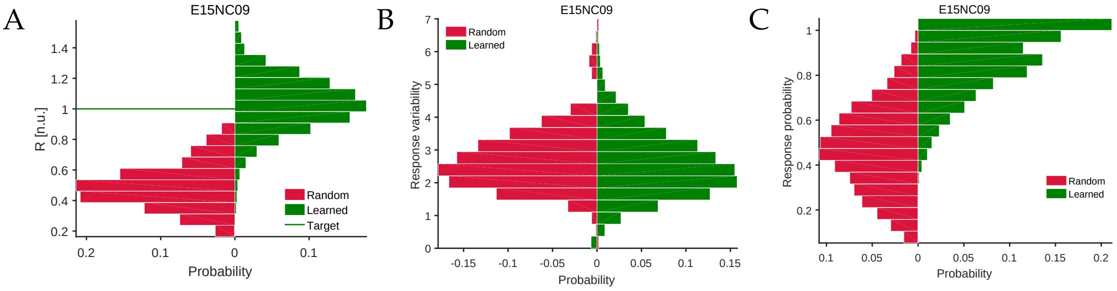

Figure 5.19. **Case 5:** *(A) The distribution of binned mean response strengths relative to the target (green line) and normalized by the standard deviation of response strengths during random and learned strategies showed that after learning, mean response strengths shifted stayed around the pre-defined target. (B) The distribution of standard deviations of binned response strengths during random and learned strategies suggested that short time-scale trial-to-trial response variability was not much impacted in this network by the control strategy. (C) The distribution of Response probabilities before and after learning showed that far fewer response failures occurred after learning suggesting that the controller learned to avoid actions that were likely to fail (see also Fig 5.18B–C).*

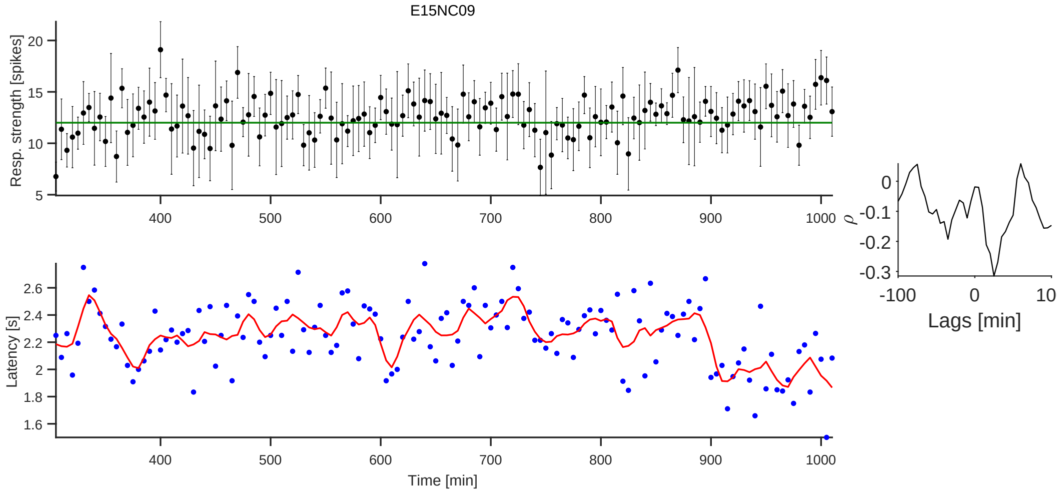

Figure 5.20. **Case 5:** *Snippet of the clamped phase, response strengths computed over 7 min bins (top) and the corresponding mean stimulus latencies chosen during the clamped period (bottom). The cross correlation of the two measures indicate a negative peak at around 18 min (inset).*

latencies that tend to oppose response strength fluctuations. The lag corresponds to the approximate episode duration or the delay after which the controller updates itself.

Case 6

We report another example of a stable controller under similar settings. Performance was found to improve as soon as the learned adaptive controller was active (Fig 5.21A).

During the random phase, mean response strengths were low (≈ 3 – 5 spikes). At ≈ 175 min, the learned controller was active. Thereafter, performance improved until it settles around the target near the 250 min mark. Mild overshoot and subsequent

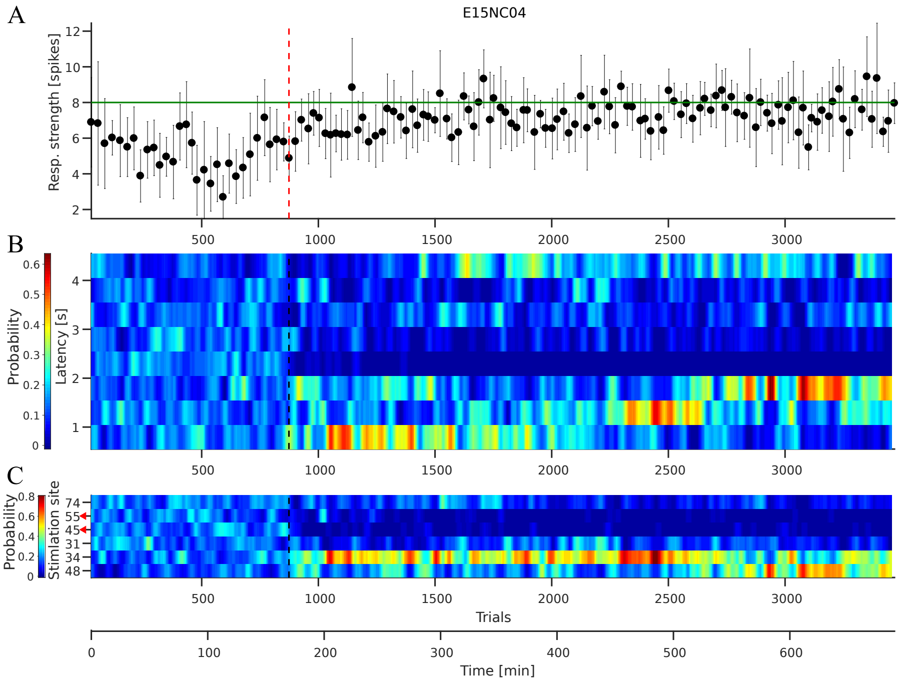

Figure 5.21. **Case 6:** *(A) Response strengths smoothed with a 5 sample moving Gaussian window and binned ($\mu \pm \sigma$) every 5 min. (B, C) Evolution of the distribution of the controller's choice of latencies (B) and stimulation sites (C) during the session, computed using a sliding window of 10 trials each. Red triangles indicate the designated dud-sites. Dotted lines in all panels indicate the end of the non-adaptive phase where random actions were chosen by the controller.*

oscillatory behaviour persisted for the remainder of the session. As response strengths moved toward the target, slight reduction in variability was also observed in this session (Fig 5.22B).

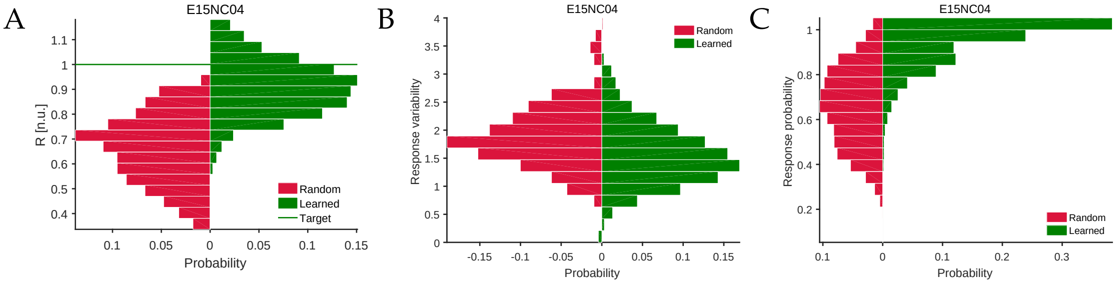

Figure 5.22. **Case 6:** *(A) The distribution of binned mean response strengths relative to the target (green line) and normalized by the standard deviation of response strengths during random and learned strategies showed that after learning, mean response strengths shifted stayed around the pre-defined target. (B) Distributions of the standard deviations of binned response strengths during random and learned strategies suggested that trial-to-trial response variability was considerably reduced by the control strategy. (C) The distribution of response probabilities before and after learning showed that far fewer response failures occurred after learning suggesting that the controller learned to avoid actions that were likely to fail (see also Fig 5.21C).*

During the initial 800 trials (175 min), latencies and stimulation sites were randomly selected from the available action set (Fig 5.21B–C). In contrast, certain actions were found to be preferred in the learned phase. When executing the learned policy, the controller relied mostly on two stimulus sites (34 and 48, Fig 5.21C). Among these, 34 was preferred in the first 5 hours of the closed-loop session and 48, thereafter. Channels 55 and 45 were dud-sites planted for a quick validation of the learned policy. During the learnt phase, these sites were avoided, validating the efficacy of the learned policy.

Considerable fluctuations in the chosen latencies were observed during the period when response strengths were clamped to the pre-set target (Fig 5.23). Moreover, their cross correlation revealed a negative peak at a lag of ≈ 18 min. The negative value points to the controller's tendency to correct overshoots and modulations in response strengths and stay on target by choosing latencies that tend to oppose response strength fluctuations. The lag corresponds to the approximate learning delay after which the controller updates itself.

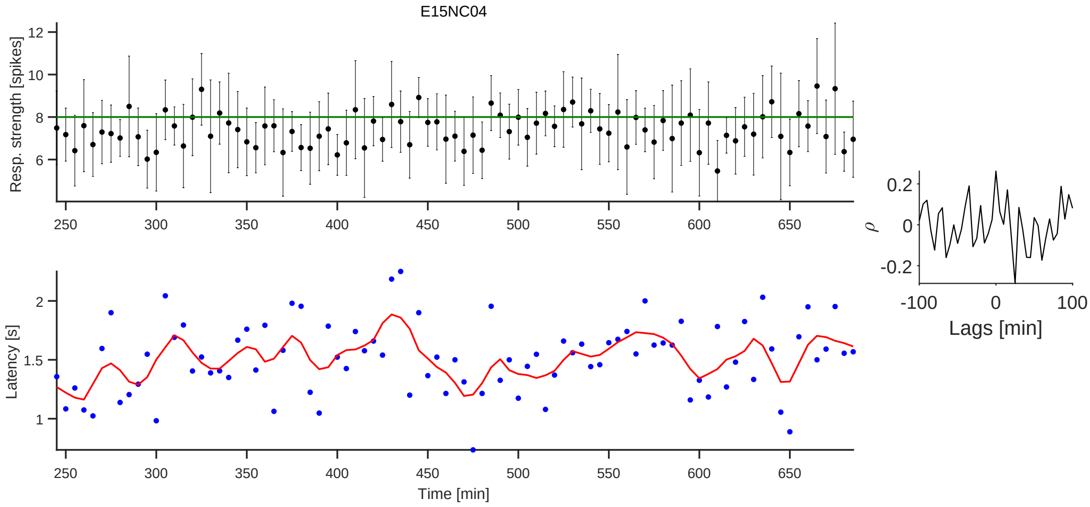

***Figure 5.23.* Case 6:** *(Top) Snippet of the clamped phase, (top) response strengths computed over 5 min bins and the corresponding mean stimulus latencies chosen during the clamped period (bottom) . The cross correlation of the two measures indicate a negative peak at around 18 min (inset).*

The relationship between binned previous SB event strengths and the latencies chosen by the RL controller, when responses were clamped to the pre-set target, was non-linear and non-monotonic (Fig 5.24).

In summary, following a failure driven strategy, we tuned potentially destabilizing factors in the learning algorithm. This led to substantial improvements in the controller's performance across networks (Fig 5.26). Stability and improved target

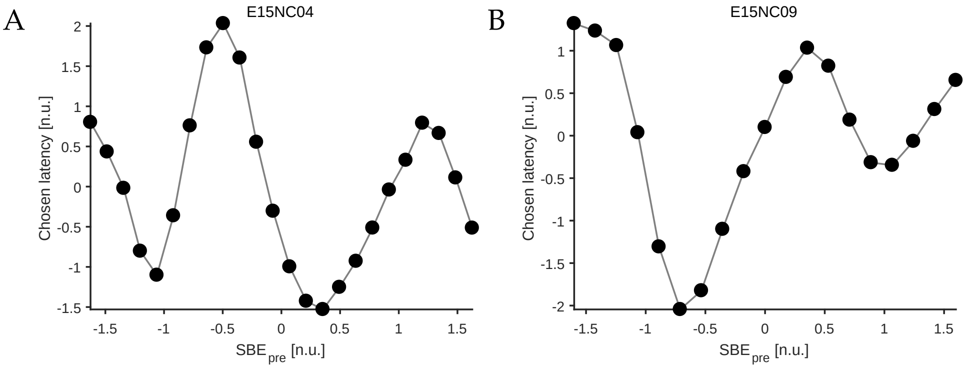

Figure 5.24. *Relationship between the previous SB event strengths and chosen stimulus latency during periods when responses were clamped to the pre-set target. Panels (A) and (B) show data corresponding to the networks discussed in Cases 5 and 6 respectively. Both time series were z-scored. The non-monotonic relationship shared qualitative similarities.*

reachability relative to a random policy was achieved in 27 out of the 29 networks where the RE was active throughout the session (Fig 5.25). Feedback instabilities were observed in 2 networks. We demonstrate that achieving stable RL based adaptive control of biological neuronal networks is indeed feasible though non-trivial. A number of caveats govern the stable operation of such co-adaptive systems, as illustrated in our networks. More data are reported as supplementary material in Appendix C.3.

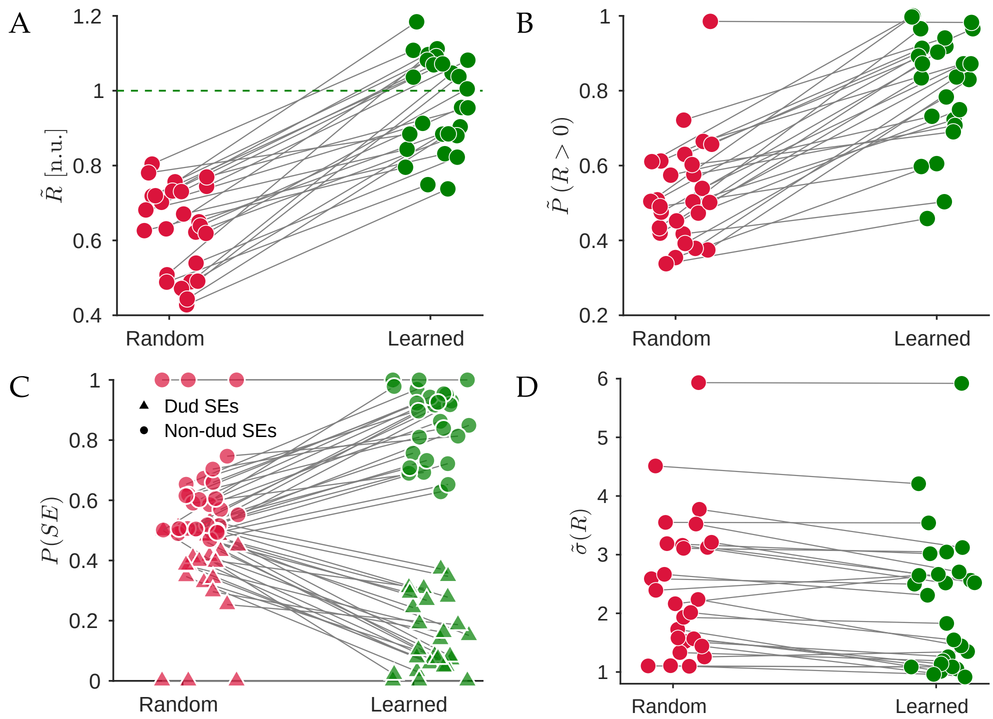

Figure 5.25. *(A) Response strengths after learning moved closer to the pre-set target in each network studied (n=27). Circles correspond to the median binned response strength $\tilde{R}$, normalized by the predefined target for each network (see Figs 5.19A, 5.22A). (B) Failed trials were less likely after learning. The circles correspond to median response probabilities (see Figs 5.19C and 5.22C) in each network. (C) Dud SEs were the least preferred SEs after learning, in each network. Circles and triangles correspond to the assigned non-dud and dud SEs respectively for each network. The probability of choosing an SE of each type before and after learning is shown for each network. No dud SEs were assigned in the 3 networks with zero probabilities. (D) Learned stimulation policies were, however, unable to reduce response variability. The median σ (see Figs 5.19B, 5.22B for examples) for each network is shown during the random and learned phases. Solid lines in each panel connect data points corresponding to each network. The jitter along the x-axis was added to aid visibility.*

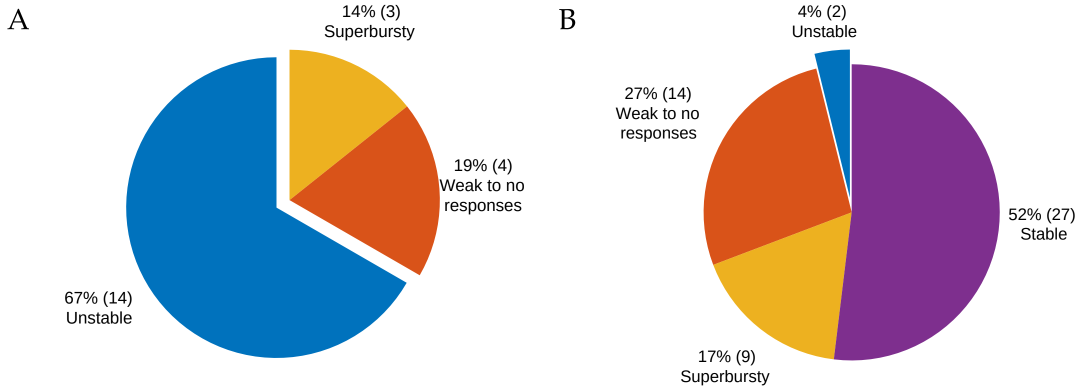

Figure 5.26. *(A) In the 21 networks where learning delays were large (updated every 100 trials) and LSPI was the learning algorithm, sustained interactions were possible in 67% of the networks. They all exhibited stochastic instabilities and response fluctuations (see Section 5.2.3 and Appendix C.2). (B) 52 networks were studied with shorter learning delays (updated every 50 trials) and NFQ as the learning algorithm. Stable performance was achieved in 93% (i.e. 27 out of 29) of the viable networks.*

Summary

- We defined a toy problem to develop algorithms for the autonomous control of BNNs. The aim was to clamp response strengths to pre-defined levels.
- Closed-loop performance was found to vary widely across networks.
- Cases exhibiting dynamic instabilities were analysed to understand the factors mediating failures.
- Mode switches in ongoing activity and sharp non-linearities in the interaction model were notable biological factors that contributed to dynamic instabilities.
- High dimensional controller specifications (state and action spaces) and learning delays in the loop were technical factors destabilizing the system.
- A failure-driven algorithm development strategy allowed us to mitigate the likelihood of failures and improve performance.
- Our results show that though autonomous methods may indeed be feasible for stable adaptive control of BNNs, guaranteeing robust performance remains an open problem.
- Formal methods for safety and stability assessments are necessary; our approach offers tentative directions to aid their development.

Chapter 6

Discussion

[*] Closed-loop stimulation has been proposed as a promising strategy for targeted interaction with the activity dynamics of pathological networks in the brain. Realizing such a framework in the clinical context is, however, riddled with challenges. It remains unclear how to identify appropriate signal features to measure the 'state' of a neuronal network. How stimulation settings can be optimized relative to a given target is also not understood. Further, the plastic nature of biological neuronal networks makes it cumbersome to understand stability properties and guarantee safe operation of such a co-adaptive scheme.

In this thesis, we proposed a novel RL based closed-loop paradigm capable of addressing some of the aforementioned challenges. RL based methods are distinguished from other approaches in that they can learn from direct interaction without relying on exemplary supervision or complete models of the system. However, the feasibility and specific challenges of coupling such controllers in a co-adaptive architecture with high-dimensional and plastic ensembles of biological neuronal networks has remained unexplored.

To explore the viability of such a paradigm and develop learning algorithms to achieve goal-directed interaction, we used generic neuronal networks *in vitro* as a model system. Cultured neuronal networks, in contrast to *in vivo* brain networks, are devoid

[*] Parts of Section 6.1 were included in the "Discussion" section of our published article: Kumar SS, Wülfing J, Okujeni S, Boedecker J, Riedmiller M, Egert U (2016) Autonomous optimization of targeted stimulation of neuronal networks. *PLoS Comput Biol* 12(8): e1005054.

of anatomical or functional specialization and offers a generic network that preserves low-level networked neurophysiological mechanisms thought to mediate challenges relevant to closed-loop interactions with neurobiological systems. Moreover, they are easier to maintain in a controlled bi-directionally accessible environment, independent of cognitive or behavioural states.

We carefully formulated toy problems that translated the conceptual structure of the closed-loop neurotechnological problem and approached goal-directed RL algorithm development with little *a priori* information. In the first part of this thesis (Chapter 3), we asked if such a controller could address a multi-objective optimization problem defined on the model system. Our results offer the first proof-of-principle of an autonomous controller solving an optimization problem in biological neuronal networks by interacting with them (Kumar et al., 2016). The details are discussed in Section 6.1.

In the second part (Chapter 5), we investigated the extend to which such controllers could adapt to the poorly understood temporal non-stationarities corresponding to changes of the interaction model, typically found in neuronal networks *in vivo* and *in vitro*. Using the history of ongoing and evoked activity to track temporal trends and continuously revise stimulation policies, we devised a closed-loop strategy responsive to the dynamics of input-output relations. Analysis of long term performance across networks suggested that the co-adaptive paradigm was prone to feedback instabilities. Using post-hoc analysis, we describe few classes of instabilities that biological neuronal networks coupled to adaptive controllers could fall into. An informal failure-driven algorithm development strategy enabled us to reduce failure rates and improve performance in many networks, suggesting that stable solutions may indeed be feasible. However, formal strategies to guarantee robust and stable operation remain an open problem. The challenges posed by system non-stationarity, tentative strategies to address them and the challenges involved in ensuring safe and stable operation of the paradigm are discussed in Section 6.2.

6.1 Optimal interactions with biological neuronal networks

It is often the case in a neurotherapeutic context that one may not be able to specify *a priori* an explicit desired neuronal output pattern, but can only state overall system objectives such as maximizing activity correlations, minimizing synchrony, etc. Examples include maintaining asynchronous activity in PD or prolonging residence in states that minimize susceptibility to epileptic seizures. The main features of such a goal is that a quantitative measure of the target response cannot be clearly identified *a priori*. It is instead intrinsically determined, variable, or emerges as a result of multiple interacting processes. It might also be desirable to qualitatively constrain interactions as is heuristically done in open-loop stimulator programming to balance the trade-off between clinical improvements and stimulus-induced side-effects. Additionally, the ability to include technical constraints in the cost functional (e.g. energy expenditure), could significantly add value to such therapeutic technologies.

The extremum-seeking problem structure underlying such situations was captured in a trade-off scenario identified in our model system. We discuss the nature of the problem, our approach to using autonomous RL algorithms and interpret the results in this section.

6.1.1 Translating the optimization problem to generic networks *in vitro*

The extremum seeking problem with no numerical set-point available *a priori* was translated in our model system to a trade-off problem involving the interplay of ongoing and evoked activity patterns in neuronal networks. The problem involved identifying the optimal stimulus latency relative to ongoing spontaneous bursts (SB) that maximized the response strength evoked per trial. Maximal achievable response strengths was network dependent and unknown *a priori*. We found that each network had a unique optimal latency depending on the ongoing and evoked activity patterns it supported.

The temporal relationship of SB events in these networks was approximated by a log-normal function. Moreover, the strengths of responses to electrical stimuli is known to fit a saturating exponential model, dependent on the period of inactivity following an SB event (Weihberger et al., 2013). Response strength was defined as the number of spikes detected in a predefined temporal window (typically 500 ms) from stimulus onset. Note that Weihberger et al. (2013) defined response lengths in temporal terms (time to the last spike in the detected response) and used a two parameter model of the form $A(1 - e^{\lambda t})$ to capture the interaction. Our data showed that spike counts in a given post-stimulus window was proportional to response lengths measured in time. Additionally, when used with a three parameter model $A(1 - e^{-\lambda t}) + B$, improved fits were obtained compared to the scheme described in Weihberger et al. (2013). The spike count implementation was also simpler and more robust for closed-loop processing and allowed a straightforward definition of the reward function. Therefore, we used the improved interaction model based on the spike count scheme for all our experiments.

Interaction of these underlying processes gave rise to an abstract objective function predicting a network specific and unique stimulus latency that maximized the number of response spikes evoked over repeated stimulation. The goal set for the RL controller was to autonomously identify this optimal latency.

Our toy problem captures crucial elements of the challenges that closed-loop paradigms face in a biomedical application, i.e. in a very complex, adaptive environment. Balancing the trade-off between response strengths and interruptions involves figuring out their interdependence and simultaneously factoring in the likelihood of ongoing activity to choose an appropriate stimulus latency. With every network being distinct in the properties of its spontaneous dynamics and response to stimuli, the paradigm was tested for robust operation over a range of parameters (Kumar et al., 2016). Furthermore, ongoing activity is highly variable and subject to unpredictable modulation over a wide range of time scales. Stimulation may induce plasticity of synaptic coupling and thus could lead to further challenges.

6.1.2 Does a unique optimal latency exist for each network?

One of the advantages of the chosen problem in the context of our model system was the availability of approximate phenomenological models describing stimulus-response relations. This allowed us to investigate the well-posedness of the control problem using numerical methods. We studied the nature of the objective function and how each model parameter contributed to it.

Simulations revealed the multi-modal nature of the control problem – in that two separate measurable modalities were simultaneously involved – the exponential recovery function and the statistical model of ongoing event occurrence. Combining them enabled us to visualize the non-linear convex cap input-output dependence $f(t)$. The optimal stimulus efficacy, $f(t^*)$, was unique, distinct for each network and necessarily attainable, because the corresponding t^*, the optimal latency, always belonged to the domain of interest (0 – 10 s) throughout the span of parameter values observed experimentally. Each network, being a parameter combination determined from fits to open-loop data, therefore mapped to a single non-linear input-output curve that belonged to the set of objective functions described earlier. In other words, a unique and optimal solution existed for all parameter sets within the observed range (i.e. all networks).

6.1.3 Assessing the quality of autonomously learned solutions

In the generic trade-off problem considered, the maximal value of the objective function for each network was not known *a priori*. Consequently, evaluating the quality of a solution autonomously learned was non-trivial. This is in general an issue in any intervention where only overall, i.e. non-parametric system objectives and not their quantitative specifics are known upfront.

To address this hurdle, we relied on prior studies on such networks (Weihberger et al., 2013). This information was used to capture the dynamic interplay of ongoing and evoked activities using parametric models under the assumption of system stationarity. By fitting open-loop interaction data to these phenomenological models, we were able to predict network-specific optimal stimulus latencies to evaluate the quality of the learned strategy for each network.

However, predictions were based on objective functions estimated from models fits made to spontaneous activity and noisy samples of response strengths evoked during open-loop interactions. Predictions made from these models were therefore only as good as the chosen model and how well it could be fitted to the data-set. Based on the coefficient of determination of model fits to data from each network, we estimated the 99% confidence interval (CI) for the predicted peak stimulus efficacies.

Experimentally achieved efficacies after learning were found to lie within this interval in 8 of the 11 networks studied and could be above or below predicted efficacy levels (Fig 3.9D). A further reason for departures from predicted levels could be temporal inhomogeneities that built up in network activity between the time of the original estimate of the objective function and the learning sessions eventually available for evaluation. The gain in efficacies from a naive to a learned controller also varied across networks. As Figs 3.5B and 3.6A illustrate, the time constant of the exponential stimulus-response relationship is a significant factor affecting the achieved gain in stimulus efficacies.

In short, stimulus efficacies improved after learning in all networks studied and lay within the 99% CI of predicted optima in most networks. This positively validated not only the quality of the learned solutions but also the robustness of the autonomous strategy. To our knowledge, this is the first demonstration of the ability of artificial agents to learn to interact optimally with biological neuronal networks (Kumar et al., 2016).

6.1.4 Is the assumption of stationarity valid?

In this study, stationarity of the input-output relationship was a necessary assumption to compare optimal stimulus latencies predicted from open-loop data at one point in time, to those learned later in closed-loop sessions. However, this is not necessarily correct. Given that stimulus-response interactions in *in vitro* and *in vivo* neuronal networks are known to undergo activity dependent changes, how valid is our assumption (Minerbi et al., 2009; Arieli et al., 1996; Hasenstaub et al., 2007; Azouz and Gray, 1999)?

Differences in the magnitudes of correlations between parameters A (strongly correlated), B (less positively correlated) and λ (no correlation) in open vs. closed-loop data fits are possible indicators that some parameters were perhaps more strongly modulated over time than others.

In spite of such sources of variability, the correlation of the learned latencies of the controller with the preceding, temporally distant, open-loop predictions was surprisingly strong. The likely explanation is that the temporal resolution of the controller – the chosen state-space discretization – was relatively coarse at 0.5 s. The parametric model of the trade-off problem showed that the impact of parameter fluctuations on optimal latencies was small relative to this resolution of the state-space (Figs 3.3, 3.4 and 3.5). The actual optimum, thus, could fall in the neighbourhood of the learned latencies. Such a tendency is indeed visible in the error between the learned and optimal times (Fig 3.8F), which are centred around the optimum.

Therefore, in the context of our experiment, we argue that since the objective function and hence the optimal solution remained reasonably invariant under the range of temporal fluctuations observed in our model parameters, the specific optimization problem could be assumed stationary.

6.1.5 The perils of learning in sessions

For experiments discussed heretofore, our closed-loop framework was designed to learn session-wise. Blocks of trials were dedicated to exploring the action space and learning an optimal stimulus policy (training session). Thereafter, the controller continually executed a learned optimal policy, during which period no further learning occurred. Such an experimental design aided the separation of performance measures before and after learning.

However, the strategy has to be generalized with caution. It assumes that the problem at hand is stationary and that repetitive interactions with the network are serially independent. Though the specific trade-off problem we addressed was arguably within the scope of these assumptions, it has to be noted that they may not be generally true of interaction problems involving biological neuronal networks. Such strategies are unprepared to respond to qualitative changes in stimulus response interactions occurring either spontaneously or as a consequence of interacting with the network.

Unconstrained and repetitive execution of a particular learned policy may introduce serial dependencies and induce activity features beyond the repertoire experienced during training sessions. This could lead to errors between the 'true' and estimated or learned optimal policies that relegate the controller to a sub-optimal regime. Even where the objective function per se is invariant under induced modifications in features of network activity, such interactions should be deemed unsafe because of the poor predictability of the activity patterns so induced.

Effects indicating this situation were observed during some of our sessions where the nature of interactions was distinctly different across adjacent training and testing sessions. During initial training sessions, we found that the naive controller was likely to learn a low and incorrect value of optimal latency. When such policies were repetitively delivered to the network during testing, increased response delays were observed (see Fig 3.15). A possible explanation is the resource depletion and the attendant network refractoriness resulting from the relatively higher stimulation rates during this period (Weihberger et al., 2013). An alternative explanation is the selective adaptation of the network to the frequently presented stimulus – a phenomenon described previously in such networks (Eytan et al., 2003). Changes in synaptic transmission (excitatory synaptic depression and increased background inhibition) were shown to underlie such selective gain control (Eytan et al., 2003).

Our controller ignored such unintended consequences, not just because it was in the testing phase, but also since response delays were not a feature of the state of the environment it was designed to measure. This illustrates in principle two pitfalls of the paradigm: separating learning into sessions, and more generally, working with potentially deficient states.

In our experiments, the case could be made that response delays are orthogonal to response strengths in feature space and that their independence meant the objective function remained invariant. But the principle, when extended to a yet unconsidered dependent feature, could be consequential. In general, the point remains that constraints on action selection and their sequential execution may have to be imposed in the interest of safe operation. Features to inform the constraints will need to be identified online rather than post-hoc. Further, they may need to operate dynamically for safe and efficient learning based control. The design of such an autonomous supervisory framework remains an open problem.

6.2 Autonomous adaptive control

A pervasive element of neuronal activity in the brain is fluctuation (spatial and temporal), the scale and origin of which remains poorly understood (Cabral et al., 2014, and references therein). Under pathological conditions a further layer of time-evolving processes that disrupt the healthy evolution of brain states is thought to be involved (van den Heuvel and Sporns, 2013). An ideal neurotechnological intervention would have to operate efficiently notwithstanding the ongoing dynamics of activity. Apart from finding optimal stimulus policies, it has to adaptively reconcile the temporal evolution of the context within which the policy was optimal.

Neurotechnological devices capable of dynamically adapting to ongoing activity are however challenging to develop, mainly because of a poor understanding of the network state, its spatio-temporal extent and evolving influence on network-stimulus interactions. Further, closing the loop on a partially understood dynamic system runs the risk of feedback instabilities, oscillations and runaway behaviour.

In the second part of the thesis, we approached the design of an autonomous adaptive controller coupled to biological neuronal networks. We assessed viability of the approach, identified some of the factors governing dynamic stability and highlighted specific challenges that need formal treatment for the development of a safe and stable closed-loop neurotechnological framework.

6.2.1 Translating the adaptive control problem to the model system

Generic neuronal networks *in vitro* are known to undergo activity dependent changes (Minerbi et al., 2009). Long term activity fluctuations, though reported in various studies using cultured neuronal networks *in vitro*, are yet to be comprehensively characterized in literature (Baltz and Voigt, 2015; Haroush and Marom, 2015). Our data revealed slow fluctuations around the quantitative model describing stimulus-response relations when stimuli were delivered serially over long time scales. We hypothesized that to repeatedly evoke the same response strength, a stimulation policy would need to be adjusted to accommodate the slow dynamics of underlying network states. Our data also suggest that the background process modulating stimulus-response relations may also reflect in features of spontaneously occurring events proximal to the stimulus. Using these insights, we devised an RL based autonomous adaptive controller to clamp response strengths to a pre-set levels.

The controller had to learn an evolving stimulation policy to clamp response strengths to target levels. Over long time scales, it had to maintain goal-directed behaviour by appropriately adapting its policies using information from the history of evoked and ongoing activity in the network. Neither the mapping between state and action spaces nor the time scale of fluctuations of stimulus-response relations was available *a priori* to the controller. In any case, they were network dependent and often difficult to characterize precisely. This framework recreates the essential structure of the control problem in a clinical context. Importantly, it offers an opportunity to develop algorithms and better understand the challenges of achieving adaptive control with biological neuronal networks in a controlled setting.

6.2.2 Stability and safety concerns of the autonomous paradigm

Dynamical systems coupled in a feedback loop could lead to instabilities. Since our paradigm is based on model-free interactions with biological neuronal networks, it is impossible using current techniques to formally guarantee dynamic stability of the feedback loop. In our experiments, we did indeed observe dynamic instabilities arising in the autonomous loop. Here we discuss qualitatively the nature and properties associated with such instabilities and attempt to attribute their origin to either the biological network or the controller though the closed architecture made it cumbersome to identify the origin of instabilities.

Instabilities driven by the network

A characterizing feature of biological neuronal networks is the poorly predictable migration between 'network modes' and the consequent modulation of network-stimulus interactions over long time scales. Such non-stationary development of MDP dynamics is problematic for RL algorithms since we need to store and re-use experiences or samples from the past to train them. When learning is hampered due to experiences that are inconsistent with observed activity dynamics, situations could arise where stimulation is delivered inappropriately, despite being in a mode where intervention was perhaps not necessary or even 'unsafe'.

Our experiments exposed limitations of this kind in some our networks. The predefined goal was seemingly unattainable during certain periods in the closed-loop session (see Case 1). Drastically different durations were needed to achieve the same target in two closed-loop sessions on the same network (Figs 5.1, 5.4). Further, when the network switched to a different mode, the hitherto successful strategy failed to achieve target (Fig 5.4). They suggest that while learning impacts stimulation outcomes, the ability to learn such policies may be constrained to 'favourable' network modes.

However, such conspicuous mode fluctuations were a feature of only a small subset of the networks we studied (see Fig C.1 in Appendix C). Not all of them were amenable for closed-loop learning. During characterization studies prior to the learning sessions, the network had to remain in a stable mode (the low-variability mode in the example discussed here) so that recovery function fits could be made and a reachable target assigned. Further, the time scales of mode switches had to be contained within the experimental duration. We did not find other networks where such conditions were met. Nevertheless, the described case study illustrates one of the major pitfalls involved in interacting with the temporally non-stationary activity dynamics of biological neuronal networks.

As an emergent property of the network, such mode-switches pose a direct challenge to autonomous learning algorithms and thus map to the structure of the challenge in a neurobiological context of encountering non-stationary network modes that switch in a state (sleep, wake etc.) or task (lying, walking) dependent manner. In the context of a clinical application, such situations could lead to undesirable or unsafe interactions. Crucially, such periods of unsuccessful exploration could also eclipse the learnt (once) successful stimulation policy, which may have to be re-learned in a future mode.

In our experiments, we tried to partly address the issue by resorting to a sliding window approach on the controller's memory. To this end, we kept an arbitrarily chosen number of samples in memory and 'forgot' older ones. Needless to say, such operational rules of thumb may not be optimal. Current RL algorithms do not offer elegant solutions to such challenges. Formalizing problems of this nature will drive the development of elegant algorithms capable of identifying and robustly handling conflicting experiences online.

Another cause of network-driven instabilities was sharp non-linearities in network-stimulus interactions. Biological neuronal networks, being composed of non-linear elements, exhibit sharp threshold phenomena across various scales of observation. Quantitative relationships binding stimulus-response interactions exhibit non-linear recovery like behaviour explained by a saturating exponential model (Weihberger et al., 2013; Kumar et al., 2016). However, time-constants of the model varied widely across networks. Of particular interest were cases when recovery was fast and the non-linearity, step-like.

When the recovery function was shallow with optimal strategies falling roughly on its rising phase, the Q-function was found to be successfully approximated using linear methods. This strategy, however, failed when the time-constant of the exponential was small. Much better performance was obtained in such cases by using non-linear methods to approximate the Q-function (Neural Fitted Q-iteration (NFQ)). Experimental data described in Case 2a and 2b demonstrate that the nature of input-output relationships was an important factor to consider while developing autonomous control strategies. The observation is notable especially since neuronal networks in the brain are often characterized by non-linear threshold like phenomena.

Our results suggest that for the paradigm to be stable, the choice of algorithm has to be made after considering the nature of the system-controller interaction and the desired activity patterns or features. What are the stability bounds on each class of algorithms? What techniques could help formally assess system stability? Further investigations to address such questions are necessary before autonomous paradigms could be of broader clinical appeal.

Instabilities due to the controller

The enormity of the parameter space to be explored poses a challenge for neurotechnological devices in a clinical context. The ability to navigate high-dimensional action spaces and converge quickly to optimal strategies is extremely desirable for an autonomous paradigm that seeks to optimize stimulation policies. We captured this aspect of the control problem by providing the controller an augmented action set. Ineffective actions were also included in many cases to serve as worst-case validators of the algorithm.

Since the set of actions were large, we set the Q-function to be updated only every 100 trials to gather enough samples involving each action (see Cases 3 and 4). Interestingly, such a configuration led to a high likelihood of oscillatory instabilities. A prominent signature of instabilities of this kind was the systematic evolution of the chosen actions. The sequence of actions chosen by the controller closely mirrored the structure of these excursions but not that of ongoing activity patterns in the networks, suggesting that the observed instabilities were likely mediated by an unstable policy.

Likely causes were the higher dimensional action set (resulting in potentially nonlinear Q-functions), and long learning delays. Instabilities were particularly accentuated, the higher the cardinality of the action set. A likely explanation is that with more actions to explore with the same number of trials, approximations of the Q-function were unreliable given that fewer samples were available to capture the underlying mapping.

Instabilities of this kind demonstrate the sensitivity of the paradigm even to technical factors involved in its controller design.

Pursuing an incremental and failure-driven algorithm development strategy allowed us to mitigate the likelihood of failures across networks. Considerable improvements were observed in the target centricity and long-term stability of the interactions. Our results demonstrate that given suitable conditions, stable and adaptive control of biological neuronal networks may indeed be feasible with autonomous paradigms. These conditions are broadly determined by characteristics inherent to biological neuronal networks and the nature of the control problem. Only when control parameters were tuned to accommodate properties unique to such systems, did the paradigm operate satisfactorily. This opens up the following questions: How can the limits of a 'suitable' parameter space be determined relative to the given control problem? How can a formal strategies to test and validate the existence of such parameter spaces be designed?

Further theoretical studies are necessary to formulate benchmark problems and design algorithms to handle the identified problem framework elegantly. Alongside, formal methods to guarantee stable performance will be key to adding translational capabilities to our findings.

6.3 Are networks always controllable?

Controllability describes the ability of an external input to move the output (in our case) from any initial to any final condition in a finite interval of time. Formal determination of system controllability is typically not possible in neurobiological control problems. An empirical approach is often the feasible alternative.

In our experiments, we relied on a brief stretch of offline random stimulation of each network to qualitatively assess the range of response strengths (output) corresponding to input at each channel chosen for stimulation. A target response strength typically lay within the second to third quartile of the response strengths and was assumed to be 'reachable'. The strategy was, however, not always successful and on many occasions failed to produce improvements after closed-loop learning.

In few of our experiments, the target was achievable by a random strategy. Insufficient discriminability among actions likely impaired the learning process and led to cases where we were unable to distinguish between naive and learned response properties (Fig 5.16). Such cases could be indicative of deficiencies in defining, and therefore observing the network state. Temporal drifts and loss of activity in the sole recording channel was also found to hamper network observability in some of our recordings. In such cases, the controller learned or degenerated eventually into (when drifts and non-stationarities were involved) sub-optimal policies. In a general application context, such situations represent an inherent limitation of our approach, leaving the paradigm prone to unsafe operation when network states were 'insufficiently' observable.

Another limiting factor was the assumption of sustained accessibility of the target state given the limited repertoire of actions available to the controller, even after the offline characterization. This assumption was not borne out in a few cases (Fig 5.17). The target was found consistently unattainable within the action space available to the controller. In other cases, after hours of interaction, certain stimulation sites were no longer able to evoke responses, thus cutting out accessibility of the controller to the network. In a more general context, the consequent unpredictable operation of the controller, represents another limitation of the paradigm. When random actions are continually delivered in an attempt to access potentially inaccessible states, harmful consequences may result. One strategy to handle situations of this nature is by increasing degeneracy in the action set. Increasing the number of stimulus sites could improve the likelihood of sustained target accessibility, although at the cost of higher dimensionality.

In summary, a valid definition of network states (observability), availability of appropriate actions (accessibility) and existence of a well-posed problem are crucial antecedents to the success of autonomous learning based control schemes. Methods to ensure these given a control problem are beyond the scope of this thesis. Here we stress the importance of these modalities to the success of the paradigm. Furthermore, deficiencies in these assumptions could very well lead to unsafe interactions. In the absence on formal solutions, an online supervisory framework administering the controller using running estimates of performance metrics may be a promising strategy from a translational perspective.

6.4 Describing the state of a network

The ‘state’ of a biological neuronal network can be thought of as the set of variables that completely characterize its dynamic behaviour and response to a given set of inputs. The veridical ‘state’ remains hidden in the myriad internal variables possibly pervading multiple scales, both spatial (from molecules to the network) and temporal (from milliseconds to hours). From observable signals, the challenge is to abstract a functional measure minimally rich to predict features of interest in upcoming activity, e.g. response to subsequent stimuli.

In generic networks *in vitro*, the notion of 'network excitability' has been proposed as a continuously evolving variable emerging from dynamic interactions of its internal state variables (Tabak et al., 2001; Weihberger et al., 2013). Though hidden, it can be sampled by perturbing i.e. stimulating the network, though the perturbation itself may affect the excitability measure. Models describing the modulation of excitability, relating bursting and periods of inactivity have been proposed to explain the emergence of ongoing activity as well (Tabak et al., 2001). *A posteriori* modelling in perturbed networks suggests that excitability may be modulated by ongoing spontaneous bursts (SBs), such that it remains low directly after a burst and recovers gradually thereafter, over a time scale in the order of seconds (Weihberger et al., 2013).

For our optimization problem (Chapter 3), we exploited this understanding of stimulus-response relationship to define the low-dimensional state-space for the controller. The discretized latencies after an SB were exploited as 'states'. Rewards and punishments were assigned based on the consequences of interacting with the network at each state. The strategy focused the state-space to a relevant low-dimensional sub-space, thus circumventing the need to explore other less informative options and limiting the iterations necessary for the controller to converge. The approach could serve as the basis of a generalisable strategy where consistent phenomenological models describing activity dynamics are available.

The downside of the approach is that the algorithm fails to respond to processes not described or captured by the model. An example is the serial structure in response strength sequences due to possibly higher rates of stimulation, activity dependent plasticity or damage to neurons. In a general context, such situations may result in potentially unsafe operating regimes, if not externally constrained by additional means.

We noticed the presence of slow and systematic trends in the fluctuation of stimulus-response relations, suggesting a deficient state specification. To cope with this situation, we modified our state definition. Since recorded response strengths were a reflection of instantaneous excitability, we added it directly to the state vector. Since they were serially correlated, we assumed that relevant information was present also in the history of such measurements. Thus, we included responses of prior perturbations as higher dimensions of our state vector. Alongside, intervening SBs could also be informative on the network's local excitability. The history of global SB strengths were therefore added as further dimensions of the state vector.

Control of activity patterns in an RL framework hinges on the quality of the available measure of 'network state'. Our strategy of incrementally expanding the state was heuristic and based on trends observed in data as the experiments progressed. Improvements in the state definition require a better mechanistic understanding of information processing in networks. Machine learning methods to automatically extract informative features may be promising in this regard. But it remains unclear if they will be able to cope with the spatio-temporal complexity and dimensionality of activity in neuronal networks.

Alternative methods to expose hidden states have involved the so called 'clamp rationale' or reverse functional characterization using closed-loop frameworks (Wallach et al., 2011; Wallach, 2013; Keren and Marom, 2014) which we discuss in further detail in the next section.

6.5 What can we 'learn' from the machine?

The idea of combining techniques of machine learning and reverse engineering in a neuroscientific context has received considerable interest in recent times (Wallach, 2013; Barak, 2017). Such approaches are expected to help characterize functional relationships, provide alternatives for modelling neuronal networks and serve as hypothesis generation tools (Barak, 2017).

One such technique advances the 'clamp rationale' to invert the experimenter's perspective and aims to exploit automatic control as a tool to characterize a complex dynamical system (see Wallach (2013) for review). It assumes controllability of a given response feature by manipulating some input parameter. In this context, our paradigm could be effective, especially when input-output relationships are complex and time-variant. What opportunities exist for learning from the machine? What are their limitations?

When input-output relationships are simple and time-invariant, as we showed in Chapter 3 (see Figs 3.12–3.14), model parameters could be inferred post-hoc from Q-functions. They were strongly correlated to those from model fits made to response strength series from the closed-loop session, further supporting the reverse characterization principle.

During adaptive control with higher dimensional state and action sets, action sequences demonstrated non-trivial trends over slow time-scales when responses were clamped (Chapter 5). The principle persisted across networks, though the nature and time scales of fluctuations varied widely across networks and a general principle remained elusive. The complexity and dimensionality of the relationship restricted a further nuanced interpretation using simple techniques. Such limitations have to be kept in mind when assessing the impact and scope of reverse learning. Further investigations are necessary to test if dimensionality reduction techniques could yield meaningful insights from such data sets.

Finally, it has to be stressed that data-based characterizations essentially represent a pragmatical process. Caution has to be exercised in interpreting such functional relationships. Degeneracy – i.e. multiple equivalent solutions to the same problem – being a property inherent to biological systems that impedes a classical black-box reconstruction approach (see Marom (2009), Braun and Marom (2015) for discussion).

To conclude, while general principles underlying information processing capabilities may be cumbersome to infer, such relationships will certainly be of value in a neurotechnological context and may help frame testable hypotheses for scientific studies on such networks.

6.6 Translating to real world clinical problems

How do our findings contribute to the development of smart neurotechnological solutions? Our study was not focused on addressing challenges specific to any particular neurological disorder or its treatment. Instead, we focused on developing stimulation paradigms for verifiable RL control strategies and on the bottlenecks impeding advances in general closed-loop interaction strategies with neuronal networks in the brain. The problem structure, along with the biological complexity of the underlying high-dimensional substrate was replicated in a controlled setting. Our experiments demonstrated how such challenges could be addressed within an autonomous framework. Formalizing the problem for an RL controller, developing learning algorithms with minimal *a priori* information and assessing the quality of the resultant controllers, involved strategies that could be generalized across applications. Further, the stability caveats we describe in our experiments are likely relevant, regardless of scale or application. A formal understanding of the dynamic characteristics of such co-adaptive architectures will be key to the safe translation of the paradigm to a neurotechnological context.

Driven by recent technological advancements, there has been a push to develop feedback driven stimulation strategies. So far, event based solutions are the furthest that have been explored for specific neurobiological control problems (Rosin et al., 2011; Little et al., 2013; Cagnan et al., 2017; Raspopovic et al., 2014). With parsimonious stimulus regimes that detect the need for intervention, they are expected to improve therapeutic efficacy, diminish associated side-effects and open avenues to optimize energy expenditure.

To illustrate the shared context and relevance of our approach vis-a-vis these event-driven paradigms (see Table 6.1), we break the framework up into the following functional questions:

Feedback: What should be measured? The term biomarker is often used in medical literature to refer to signals that serve as measurable indicators of normal biological processes, pathogenic processes or pharmacological responses to therapeutic interventions (Miller and O'Callaghan, 2015; Colburn et al., 2001). In feedback control terms, these approximate the 'state' of the system. Biological and technical limitations hinder an unambiguous measurement of the network state. Often, a working solution based on prior knowledge of the pathophysiology under consideration is used to choose anatomical structures and signal features to measure.

Of the various biomarkers drawn from literature on the specific disorder, a low-dimensional choice is heuristically made to constrain the state-space (see column

Table 6.1. *Elements of the feedback based interaction strategy is illustrated along with selected recent studies on Parkinson's Disease and pathological tremor. These strategies were all event-driven. The reference anatomical structure and the signal of interest represents the state space of the pathological network. Target structures and stimulus parameters denote the action set. Abbreviations: AP - action potential, GPi - internal globus pallidus, M1 - primary motor cortex, LFP - local field potential, STN - sub-thalamic nucleus, VLT - ventro-lateral thalamus*

Disorder	Study	Model system	Reference	Target	Parameters	Stability
Parkinson's disease	Rosin et al. (2011)	Non-human primates	APs detected at GPi or M1	GPi	- Site(s) and amplitude determined manually - Single pulses or package - Package - 7 pulses at 130 Hz - Delay of 80 ms heuristically set	Unknown
	Little et al. (2013)	Human patients	Power in the beta band (13 – 30 Hz) of LFP at STN	STN	- Site(s) and amplitude determined manually - Threshold chosen heuristically - Stimulus frequency set to 130 Hz - 250 ms voltage ramping at stimulus on/offset to avoid parasthesias	Unknown
Essential/dystonic tremor	Cagnan et al. (2017)	Human patients	Tremors detected peripherally	VLT	- Site(s) and amplitude determined manually - Frequency, pulse-width set patient-wise - Ideal phase chosen manually	Unknown

'Reference' in Table 6.1). Our approach to defining the state of the network followed similar lines. We relied on previously discovered phenomenological models of input-output relations in generic neuronal networks *in vitro*, to help quantify the network state. However, unlike in the *in vivo* studies mentioned above, our paradigm – being based on a systematic algorithmic framework – will be able to handle higher dimensional state representations (see also Section 6.4).

Input: How to find effective actions? The set of actions is usually large and may include stimulation targets, stimulus parameters (e.g. amplitude, frequency, pulse-width, latency), thresholds etc. Often multiple stimulus contacts are available at a given anatomical target. Crucially, event-driven paradigms do not offer any improvement in action exploration or selection strategies relative to an open-loop setting. The contribution of event-driven paradigms is limited to a potential reduction in the amount of stimulation delivered. In *in vivo* studies, while target anatomical structures were chosen based on the prior knowledge, stimulus settings were typically set using trial and error methods independently for each subject (see columns 'Target' and 'Parameters' in Table 6.1).

Our paradigm is particularly poised to address the problem of choosing optimal stimulation policies. It offers the potential to systematically explore a relatively high-dimensional action set and autonomously converge to network-wise optimal solutions (Kumar et al., 2016). Moreover, as demonstrated in our proof-of-principle study, our paradigm could also help autonomously balance trade-offs involving multiple interacting processes (Kumar et al., 2016). Features or empirical models capturing the adverse effects of interactions, energy expenditure, etc., where available, may help formulate more robust cost-functions for such controllers.

Is the system adaptive, safe and stable? Network activity in the brain varies incessantly under the influence of multiple underlying dynamic processes including plasticity, cognitive and motor load and progression of the pathology. Long-term inhomogeneities do not feature in event-driven paradigms. Their ability to sustain therapeutic improvements over long time-scales remains untested.

Our paradigm, being a continuously adapting one, offers co-adaptive capabilities. Co-adaptivity, however, adds to the risk of oscillations and runaway behaviour in the system. Assessment of the stability and safety of such paradigms are therefore necessary but remains a non-trivial challenge. In this thesis, we qualitatively explored various aspects of long-term stability and uncovered a few determining factors – both biological and technical. Such exploratory studies, we hope, will lay the ground for a more formal framework to approach these concerns.

6.7 Perspectives

Clinical techniques do not typically arise *de novo*, but are the result of gradually evolving ideas and technologies (Gildenberg, 2005). Hence it is important to reflect on the proposed framework in the evolving context of neurostimulation therapy. What specific elements of a state-of-the-art therapeutic intervention do the principles studied in this thesis map onto? Would you be willing to consider an autonomous therapeutic technology for yourself? If not, what are the gaps that need to be addressed?

Much work remains in the journey from feasibility to fruition for autonomous neurostimulation paradigms. On the philosophical front, we are only beginning to grapple with the ramifications of such recursive (agent within an agent) interventions on the notion of identity and freedom of the will. On the more empirical front, the mark of a mature technology – in my opinion – is the ability to unambiguously demarcate boundaries of its operational capabilities.

As a next step in this direction, we propose to use insights from our experiments to augment current RL algorithms and formulate novel benchmark problems for challenges unique to biological neuronal networks. However, guaranteeing the safety of autonomous interventions remains cumbersome to approach with the methods of verification and model checking as they have been developed until now. Our experiments revealed distinct classes of instabilities such systems are susceptible to. Perhaps a statistical inductive approach to the safety problem could be a fruitful starting point.

Questions on the link between controllability and complexity (and therefore information processing capabilities) of biological networks also remain open. What features in the collective phenomena observed in these networks are germane to target reachability and persistence? Can they be expressed in terms of the functional and structural complexity of the network? These questions may be addressed by pharmacologically manipulating the functional and structural properties of the network. Answers to these questions may help formally articulate limitations of control strategies and estimate the safety and stability of such interactions.

Our paradigm could also be a potent research tool to characterize functional relationships in biological neuronal networks and understand the basis of their computational capabilities.

To maximize the potential of electrical stimulation as a therapeutic, augmentative or research tool, there is a clear necessity to re-imagine it in a closed-loop framework. This thesis demonstrated how some of the challenges involved could be mitigated using an RL based autonomous paradigm. Extending the framework may be a promising step forward for clinical applications involving neurostimulation.

Bibliography

Al-Tamimi A, Lewis FL, Abu-Khalaf M (2007) Model-free Q-learning designs for linear discrete-time zero-sum games with application to H-infinity control. *Automatica* 43:473–481.

Arieli A, Sterkin A, Grinvald A, Aertsen A (1996) Dynamics of ongoing activity: Explanation of the large variability in evoked cortical responses. *Science* 273:1868–1871.

Åström KJ, Kumar PR (2014) Control: A perspective. *Automatica* 50:3–43.

Åström KJ, Murray RM (2010) *Feedback systems: An introduction for scientists and engineers* Princeton University Press.

Åström K, Wittenmark B (2008) *Adaptive Control* Dover Books on Electrical Engineering. Dover Publications.

Athans M, Falb P (2013) *Optimal Control: An Introduction to the Theory and Its Applications* Dover Books on Engineering. Dover Publications.

Azouz R, Gray CM (1999) Cellular mechanisms contributing to response variability of cortical neurons *in vivo*. *J. Neurosci.* 19:2209–2223.

Baizabal-Carvallo JF, Kagnoff MN, Jimenez-Shahed J, Fekete R, Jankovic J (2014) The safety and efficacy of thalamic deep brain stimulation in essential tremor: 10 years and beyond. *J. Neurol. Neurosurg. Psychiatry* 85:567–572.

Baltz T, Voigt T (2015) Interaction of electrically evoked activity with intrinsic dynamics of cultured cortical networks with and without functional fast GABAergic synaptic transmission. *Front. Cell. Neurosci.* 9.

Barak O (2017) Recurrent neural networks as versatile tools of neuroscience research. *Curr. Opin. Neurobiol.* 46:1–6.

Bassett DS, Wymbs NF, Porter MA, Mucha PJ, Carlson JM, Grafton ST (2011) Dynamic reconfiguration of human brain networks during learning. *Proc. Natl. Acad. Sci. USA* 108:7641–7646.

Bassett DS, Wymbs NF, Rombach MP, Porter MA, Mucha PJ, Grafton ST (2013) Task-based core-periphery organization of human brain dynamics. *PLOS Comput. Biol.* 9:e1003171.

Benabid AL, Pollak P, Seigneuret E, Hoffmann D, Gay E, Perret J (1993) Chronic VIM thalamic stimulation in Parkinson's Disease, essential tremor and extra-pyramidal dyskinesias In Meyerson BA, Broggi G, Martin-Rodriguez J, Ostertag C, Sindou M, editors, *Advances in Stereotactic and Functional Neurosurgery 10: Proceedings of the 10th Meeting of the European Society for Stereotactic and Functional Neurosurgery Stockholm 1992*, pp. 39–44. Springer Vienna, Vienna.

Benabid A, Pollak P, Hoffmann D, Gervason C, Hommel M, Perret J, de Rougemont J, Gao D (1991) Long-term suppression of tremor by chronic stimulation of the ventral intermediate thalamic nucleus. *Lancet* 337:403 – 406.

Benabid AL, Chabardes S, Mitrofanis J, Pollak P (2009) Deep brain stimulation of the subthalamic nucleus for the treatment of Parkinson's Disease. *Lancet Neurol.* 8:67–81.

Benabid AL, Pollak P, Louveau A, Henry S, De Rougemont J (1988) Combined (thalamotomy and stimulation) stereotactic surgery of the VIM thalamic nucleus for bilateral Parkinson's Disease. *Stereotact. Funct. Neurosurg.* 50:344–346.

Bittar RG, Burn SC, Bain PG, Owen SL, Joint C, Shlugman D, Aziz TZ (2005) Deep brain stimulation for movement disorders and pain. *J. Clin. Neurosci.* 12:457 – 463.

Braun E, Marom S (2015) Universality, complexity and the praxis of biology: Two case studies. *Stud. Hist. Philos. Sci. C* 53:68–72.

Breakspear M (2017) Dynamic models of large-scale brain activity. *Nat. Neurosci.* 20:340–352.

Bressler SL (1995) Large-scale cortical networks and cognition. *Brain Res. Rev.* 20:288–304.

Buonomano DV, Maass W (2009) State-dependent computations: Spatiotemporal processing in cortical networks. *Nat. Rev. Neurosci.* 10:113–125.

Butovas S, Schwarz C (2003) Spatiotemporal effects of microstimulation in rat neocortex: A parametric study using multielectrode recordings. *J. Neurophysiol.* 90:3024–39.

Buzsaki G (2006) *Rhythms of the Brain* Oxford University Press.

Cabral J, Kringelbach ML, Deco G (2014) Exploring the network dynamics underlying brain activity during rest. *Prog. Neurobiol.* 114:102–131.

Cagnan H, Brittain JS, Little S, Foltynie T, Limousin P, Zrinzo L, Hariz M, Joint C, Fitzgerald J, Green AL, Aziz T, Brown P (2013) Phase dependent modulation of tremor amplitude in essential tremor through thalamic stimulation. *Brain* 136:3062–3075.

Cagnan H, Pedrosa D, Little S, Pogosyan A, Cheeran B, Aziz T, Green A, Fitzgerald J, Foltynie T, Limousin P, Zrinzo L, Hariz M, Friston KJ, Denison T, Brown P (2017) Stimulating at the right time: Phase-specific deep brain stimulation. *Brain* 140:132–145.

Carron R, Chaillet A, Filipchuk A, Pasillas-Lepine W, Hammond C (2013) Closing the loop of deep brain stimulation. *Front. Syst. Neurosci.* 7:112.

Chen X, Xiong Y, Xu GL, Liu XF (2012) Deep brain stimulation. *Intervent. Neurol.* 1:200–212.

Colburn W, DeGruttola VG, DeMets DL, Downing GJ, Hoth DF, Oates JA, Peck CC, Schooley RT, Spilker BA, Woodcock J et al. (2001) Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. biomarkers definitions working group. *Clin. Pharmacol. Ther.* 69:89–95.

Coubes P, Roubertie A, Vayssiere N, Hemm S, Echenne B (2000) Treatment of DYT1-generalised dystonia by stimulation of the internal globus pallidus. *Lancet* 355:2220–2221.

Destexhe A, Rudolph M, Paré D (2003) The high-conductance state of neocortical neurons in vivo. *Nat. Rev. Neurosci.* 4:739–751.

Deuschl G, Herzog J, Kleiner-Fisman G, Kubu C, Lozano AM, Lyons KE, Rodriguez-Oroz MC, Tamma F, Tröster AI, Vitek JL, Volkmann J, Voon V (2006) Deep brain stimulation: Postoperative issues. *Mov. Disord.* 21:S219–S237.

Doebelin EO (1985) *Control System Principles and Design* John Wiley & Sons, Inc., New York, NY, USA.

Egert U, Knott T, Schwarz C, Nawrot M, Brandt A, Rotter S, Diesmann M (2002) MEA-Tools: An open source toolbox for the analysis of multi-electrode data with MATLAB. *J. Neurosci. Methods* 117:33 – 42.

Eytan D, Brenner N, Marom S (2003) Selective adaptation in networks of cortical neurons. *J. Neurosci.* 23:9349–9356.

Eytan D, Marom S (2006) Dynamics and effective topology underlying synchronization in networks of cortical neurons. *J. Neurosci.* 26:8465–8476.

Fisher R, Salanova V, Witt T, Worth R, Henry T, Gross R, Oommen K, Osorio I, Nazzaro J, Labar D et al. (2010) Electrical stimulation of the anterior nucleus of thalamus for treatment of refractory epilepsy. *Epilepsia* 51:899–908.

Fraint A, Pal G (2015) Deep brain stimulation in Tourette's syndrome. *Front. Neurol.* 6:170.

Franzini A, Ferroli P, Leone M, Broggi G (2003) Stimulation of the posterior hypothalamus for treatment of chronic intractable cluster headaches: First reported series. *Neurosurgery* 52:1095–1101.

Fridley J, Thomas JG, Navarro JC, Yoshor D (2012) Brain stimulation for the treatment of epilepsy. *Neurosurg. Focus* 32:E13.

Fritsch G, Hitzig E (1870) Über die elektrische Erregbarkeit des Grosshirns. *Arch. Anat. Physiol. Wissen.* 37:300–32.

Gal A, Eytan D, Wallach A, Sandler M, Schiller J, Marom S (2010) Dynamics of excitability over extended timescales in cultured cortical neurons. *J. Neurosci.* 30:16332–16342.

Gildenberg PL (2005) Evolution of neuromodulation. *Stereotact. Funct. Neurosurg.* 83:71–79.

Grosenick L, Marshel JH, Deisseroth K (2015) Closed-loop and activity-guided optogenetic control. *Neuron* 86:106–139.

Ham M, Bettencourt L, McDaniel F, Gross G (2008) Spontaneous coordinated activity in cultured networks: Analysis of multiple ignition sites, primary circuits, and burst phase delay distributions. *J. Comput. Neurosci.* 24:346–357.

Hariz MI, Shamsgovara P, Johansson F, Hariz GM, Fodstad H (1999) Tolerance and tremor rebound following long-term chronic thalamic stimulation for parkinsonian and essential tremor. *Stereotact. Funct. Neurosurg.* 72:208–218.

Haroush N, Marom S (2015) Slow dynamics in features of synchronized neural network responses. *Front. Comput. Neurosci.* 9:1–9.

Harris KD, Thiele A (2011) Cortical state and attention. *Nat. Rev. Neurosci.* 12:509–523.

Hasenstaub A, Sachdev RNS, McCormick DA (2007) State changes rapidly modulate cortical neuronal responsiveness. *J. Neurosci.* 27:9607–9622.

He BJ (2013) Spontaneous and task-evoked brain activity negatively interact. *J. Neurosci.* 33:4672–4682.

Hickey P, Stacy M (2016) Deep Brain Stimulation: A paradigm shifting approach to treat Parkinson's Disease. *Front. Neurosci.* 10.

Hinich MJ, Mendes EM, Stone L (2005) Detecting nonlinearity in time series: Surrogate and bootstrap approaches. *Studies in Nonlinear Dynamics & Econometrics* 9.

Holtzheimer PE, Mayberg HS (2011) Deep brain stimulation for psychiatric disorders. *Annu. Rev. Neurosci.* 34:289.

Isaias IU, Alterman RL, Tagliati M (2009) Deep brain stimulation for primary generalized dystonia: Long-term outcomes. *Arch. Neurol.* 66:465–470.

Joshua M, Elias S, Levine O, Bergman H (2007) Quantifying the isolation quality of extracellularly recorded action potentials. *J. Neurosci. Methods* 163:267–282.

Kandler S (2011) Spatiotemporal embedding of individual neurons into generic neuronal networks Ph.D. diss., Faculty of Biology, University of Freiburg.

Keren H, Marom S (2014) Controlling neural network responsiveness: Tradeoffs and constraints. *Front. Neuroeng.* 7:11.

Kermany E, Gal A, Lyakhov V, Meir R, Marom S, Eytan D (2010) Tradeoffs and constraints on neural representation in networks of cortical neurons. *J. Neurosci.* 30:9588–9596.

Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. *CoRR* abs/1412.6980.

Kisley MA, Gerstein GL (1999) Trial-to-trial variability and state-dependent modulation of auditory-evoked responses in cortex. *J. Neurosci.* 19:10451–10460.

Koller WC, Lyons KE, Wilkinson SB, Pahwa R (1999) Efficacy of unilateral deep brain stimulation of the vim nucleus of the thalamus for essential head tremor. *Mov. Disord.* 14:847–850.

Kopell NJ, Gritton HJ, Whittington MA, Kramer MA (2014) Beyond the connectome: The dynome. *Neuron* 83:1319–1328.

Krack P, Batir A, Van Blercom N, Chabardes S, Fraix V, Ardouin C, Koudsie A, Limousin PD, Benazzouz A, LeBas JF, Benabid AL, Pollak P (2003) Five-year follow-up of bilateral stimulation of the subthalamic nucleus in advanced parkinson's disease. *N. Engl. J. Med.* 349:1925–1934 PMID: 14614167.

Kringelbach ML, Jenkinson N, Owen SLF, Aziz TZ (2007) Translational principles of deep brain stimulation. *Nat. Rev. Neurosci.* 8:623–635.

Kumar SS, Wülfing J, Okujeni S, Boedecker J, Riedmiller M, Egert U (2016) Autonomous optimization of targeted stimulation of neuronal networks. *PLOS Comput. Biol.* 12:1–22.

Lagoudakis MG, Parr R (2003) Least-squares policy iteration. *J. Mach. Learn. Res.* 4:1107–1149.

Landau ID, Lozano R, M'Saad M, Karimi A (2011) *Adaptive control: algorithms, analysis and applications* Springer Science & Business Media.

Lange S, Gabel T, Riedmiller M (2012) Batch reinforcement learning In Wiering M, van Otterlo M, editors, *Reinforcement learning*, pp. 45–73. Springer.

Le-Yi W, Wen-Xiao Z (2013) System identification: new paradigms, challenges, and opportunities. *Acta Automatica Sinica* 39:933–942.

Leone M, Franzini A, Broggi G, Bussone G (2006) Hypothalamic stimulation for intractable cluster headache: long-term experience. *Neurology* 67:150–152.

Lewis FL, Vrabie D, Syrmos VL (2012) *Optimal Control* John Wiley & Sons.

Limousin P, Krack P, Pollak P, Benazzouz A, Ardouin C, Hoffmann D, Benabid AL (1998) Electrical stimulation of the subthalamic nucleus in advanced parkinson's disease. *N. Engl. J. Med.* 339:1105–1111.

Little S, Pogosyan A, Neal S, Zavala B, Zrinzo L, Hariz M, Foltynie T, Limousin P, Ashkan K, James F, Green AL, Aziz TZ, Brown P (2013) Adaptive deep brain stimulation in advanced Parkinson's disease. *Ann. Neurol.* 74:449–457.

Mallet L, Polosan M, Jaafari N, Baup N, Welter ML, Fontaine D, Montcel STd, Yelnik J, Chéreau I, Arbus C et al. (2008) Subthalamic nucleus stimulation in severe obsessive–compulsive disorder. *N. Engl. J. Med.* 359:2121–2134.

Marom S (2009) On the precarious path of reverse neuro-engineering. *Front. Comput. Neurosci.* 3:3–6.

Matharu MS, Zrinzo L (2010) Deep brain stimulation in cluster headache: Hypothalamus or midbrain tegmentum? *Curr. Pain Headache Rep.* 14:151–159.

Mayberg HS, Lozano AM, Voon V, McNeely HE, Seminowicz D, Hamani C, Schwalb JM, Kennedy SH (2005) Deep brain stimulation for treatment-resistant depression. *Neuron* 45:651 – 660.

McIntosh AR (2000) Towards a network theory of cognition. *Neural Netw.* 13:861–870.

Medaglia JD, Lynall ME, Bassett DS (2015) Cognitive network neuroscience. *J. Cogn. Neurosci.* 27:1471–1491.

Medaglia JD, Pasqualetti F, Hamilton RH, Thompson-Schill SL, Bassett DS (2017) Brain and cognitive reserve: Translation via network control theory. *Neurosci. Biobehav. Rev.* .

Mesulam MM (1998) From sensation to cognition. *Brain* 121:1013–1052.

Miller DB, O'Callaghan JP (2015) Biomarkers of Parkinson's disease: Present and future. *Metabolism* 64:S40–S46.

Minerbi A, Kahana R, Goldfeld L, Kaufman M, Marom S, Ziv NE (2009) Long-term relationships between synaptic tenacity, synaptic remodeling and network activity. *PLOS Biol.* 7:e1000136.

Moro E, Poon YYW, Lozano AM, Saint-Cyr JA, Lang AE (2006) Subthalamic nucleus stimulation: Improvements in outcome with reprogramming. *Arch. Neurol.* 63:1266–1272.

Morrell MJ (2011) Responsive cortical stimulation for the treatment of medically intractable partial epilepsy. *Neurology* 77:1295–1304.

MEA Manual (2017) *Microelectrode Array (MEA) Manual* Multi Channel Systems MCS GmbH, Aspenhaustraße 21, 72770 Reutlingen, Germany Available at `https://www.multichannelsystems.com/sites/multichannelsystems.com/files/documents/manuals/MEA_Manual.pdf`.

Murray RM, Åström KJ, Boyd SP, Brockett RW, Stein G (2003) Future directions in control in an information-rich world. *IEEE Control Systems* 23:20–33.

Nuttin B, Cosyns P, Demeulemeester H, Gybels J, Meyerson B (1999) Electrical stimulation in anterior limbs of internal capsules in patients with obsessive-compulsive disorder. *Lancet* 354:1526.

Okujeni S, Kandler S, Egert U (2017) Mesoscale architecture shapes initiation and richness of spontaneous network activity. *J. Neurosci.* 37:2552–16.

Pedrosa DJ, Auth M, Pauls KAM, Runge M, Maarouf M, Fink GR, Timmermann L (2014) Verbal fluency in essential tremor patients: The effects of deep brain stimulation. *Brain Stimul.* 7:359–364.

Petersen CCH, Hahn TTG, Mehta M, Grinvald A, Sakmann B (2003) Interaction of sensory responses with spontaneous depolarization in layer 2/3 barrel cortex. *Proc. Natl. Acad. Sci. USA* 100:13638–13643.

Raspopovic S, Capogrosso M, Petrini FM, Bonizzato M, Rigosa J, Di Pino G, Carpaneto J, Controzzi M, Boretius T, Fernandez E, Granata G, Oddo CM, Citi L, Ciancio AL, Cipriani C, Carrozza MC, Jensen W, Guglielmelli E, Stieglitz T, Rossini PM, Micera S (2014) Restoring natural sensory feedback in real-time bidirectional hand prostheses. *Sci. Transl. Med.* 6:222ra19–222ra19.

Rehncrona S, Johnels B, Widner H, Törnqvist AL, Hariz M, Sydow O (2003) Long-term efficacy of thalamic deep brain stimulation for tremor: Double-blind assessments. *Mov. Disord.* 18:163–170.

Riedmiller M (2005) Neural fitted Q iteration - First experiences with a data efficient neural reinforcement learning method In *Lecture Notes in Computer Science: Proc. of the European Conference on Machine Learning, ECML 2005*, pp. 317–328, Porto, Portugal.

Riedmiller M (1999) Concepts and facilities of a neural reinforcement learning control architecture for technical process control. *Neural. Comput. Appl.* 8:323–338.

Rosin B, Slovik M, Mitelman R, Michal R, Haber SN, Israel Z, Vaadia E, Bergman H (2011) Closed-loop deep brain stimulation is superior in ameliorating parkinsonism. *Neuron* 72:370–384.

Rowland NC, Jaeger D (2008) Responses to tactile stimulation in deep cerebellar nucleus neurons result from recurrent activation in multiple pathways. *J. Neurophysiol.* 99:704–17.

Sarem-Aslani A, Mullett K (2011) Industrial perspective on deep brain stimulation: history, current state, and future developments. *Front. Integr. Neurosci.* .

Schiff SJ (2010) Towards model-based control of Parkinson's disease. *Phil. Trans. R. Soc. A* 368:2269–2308.

Schuurman PR, Bosch DA, Bossuyt PM, Bonsel GJ, van Someren EJ, de Bie RM, Merkus MP, Speelman JD (2000) A comparison of continuous thalamic stimulation and thalamotomy for suppression of severe tremor. *N. Engl. J. Med.* 342:461–468 PMID: 10675426.

Seber G, Wild C (2003) *Nonlinear Regression* Wiley Series in Probability and Statistics. John Wiley & Sons.

Shahaf G, Eytan D, Gal A, Kermany E, Lyakhov V, Zrenner C, Marom S (2008) Order-based representation in random networks of cortical neurons. *PLOS Comput. Biol.* 4.

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15:1929–1958.

Sutton RS, Barto AG (1998) *Reinforcement learning: An introduction*, Vol. 1 MIT press Cambridge.

Sutton RS, Barto AG, Williams RJ (1992) Reinforcement learning is direct adaptive optimal control. *IEEE Control Systems* 12:19–22.

Tabak J, Rinzel J, O'Donovan MJ (2001) The role of activity-dependent network depression in the expression and self-regulation of spontaneous activity in the developing spinal cord. *J. Neurosci.* 21:8966–8978.

Teich MC, Heneghan C, Lowen SB, Ozaki T, Kaplan E (1997) Fractal character of the neural spike train in the visual system of the cat. *J. Opt. Soc. Am. A* 14:529–546.

van den Heuvel MP, Sporns O (2013) Network hubs in the human brain. *Trends Cogn. Sci.* 17:683–696.

Vidailhet M, Vercueil L, Houeto JL, Krystkowiak P, Benabid AL, Cornu P, Lagrange C, Tézenas du Montcel S, Dormont D, Grand S, Blond S, Detante O, Pillon B, Ardouin C, Agid Y, Desté A, Pollak P (2005) Bilateral deep-brain stimulation of the globus pallidus in primary generalized dystonia. *N. Engl. J. Med.* 352:459–467 PMID: 15689584.

Wagenaar D, DeMarse TB, Potter SM (2005) Meabench: A toolset for multi-electrode data acquisition and on-line analysis In *Conference Proceedings. 2nd International IEEE EMBS Conference on Neural Engineering, 2005.*, pp. 518–521.

Wagenaar DA, Pine J, Potter SM (2004) Effective parameters for stimulation of dissociated cultures using multi-electrode arrays. *J. Neurosci. Methods* 138:27–37.

Wagenaar DA, Pine J, Potter SM (2006) An extremely rich repertoire of bursting patterns during the development of cortical cultures. *BMC Neurosci.* 7:11.

Wallach A (2013) The response clamp: Functional characterization of neural systems using closed-loop control. *Front. Neural Circuits* 7.

Wallach A, Eytan D, Gal A, Zrenner C, Marom S (2011) Neuronal response clamp. *Front. Neuroeng.* 4:3.

Wang D, Liu D, Wei Q, Zhao D, Jin N (2012) Optimal control of unknown non-affine nonlinear discrete-time systems based on adaptive dynamic programming. *Automatica* 48:1825–1832.

Watkins CJ, Dayan P (1992) Technical note: Q-learning. *Mach. Learn.* 8:279–292.

Weihberger O, Okujeni S, Mikkonen JE, Egert U (2013) Quantitative examination of stimulus-response relations in cortical networks *in vitro*. *J. Neurophysiol.* 109:1764–1774.

Werner G (2011) Letting the brain speak for itself. *Front. Physio.* 2:60.

Wiering M, van Otterlo M (2012) *Reinforcement Learning: State-of-the-Art* Adaptation, Learning, and Optimization. Springer Berlin Heidelberg.

Xu T, Barak O (2017) Dynamical timescale explains marginal stability in excitability dynamics. *J. Neurosci.* 37:4508–4524.

Zhang K, Bhatia S, Oh MY, Cohen D, Angle C, Whiting D (2010) Long-term results of thalamic deep brain stimulation for essential tremor: Clinical article. *J. Neurosurg.* 112:1271–1276.

Zhao D, Xia Z, Wang D (2015) Model-free optimal control for affine nonlinear systems with convergence analysis. *IEEE Trans. Autom. Sci. Eng.* 12:1461–1468.

Appendices

Appendix A

Protocol for culturing neuronal networks on MEAs

[7] Here we present the protocol to isolate, dissociate and culture cortical neurons from fresh born rats on microelectrode arrays (MEAs). The preparation of MEAs, buffers and solutions took about 2 days before tissue extraction from the animal. The dissection and plating took 3 – 4 h. Cultures were typically maintained for around 6 weeks.

A.1 Materials and Methods

A.1.1 Chemicals

5% CO_2–air mixture	Air Liquid, Freiburg, Germany
Deionized water	
DNAse	Sigma-Aldrich, Munich, Germany
Ethanol	neoLab, Heidelberg, Germany
Gentamicine	Invitrogen, Paisley, UK
Glucose	Invitrogen, Paisley, UK
Horse serum (heat inactivated)	Invitrogen, Paisley, UK
L-Glutamine	Invitrogen, Paisley UK
MEM-Eagle w/o L-Glutamine	Invitrogen, Paisley, UK
Phosphate buffered saline (PBS)	Invitrogen, Paisley, UK
Trypsin	Invitrogen, Paisley UK
Polyethylenimine (PEI)	Sigma-Aldrich, Munich, Germany

[7] Thanks to Samora Okujeni and Ute Riede for providing details of the protocol.

A.1.2 Tools

Eppendorf tubes	Roth, Karlsruhe, Germany
Pipette tips	Roth, Karlsruhe, Germany
Scalpels	Bayha GmbH, Tuttlingen, Germany
Scissors	Dumont & Fils, Switzerland
Serological pipettes (5, 10 and 25 mL)	Becton-Dickinson, NJ, USA and Falcon, Munich, Germany
Spatulas	neoLab, Heidelberg, Germany
Syringe (1 mL)	Roth, Karlsruhe, Germany
Syringe filters (0.22 µm)	Roth, Karlsruhe, Germany
Tubes	Falcon, Munich, Germany
Tweezers	Dumont & Fils, Switzerland

A.1.3 Devices

Automated cell counter	Casy - Schärfe Systems GmbH, Germany
Camera	Microcular PCE-ME100, PCE Deutschland GmbH, Germany
Centrifuge	Rotofix 32A, Hettich, Tuttlingen, Germany
Incubator	CB210 – Binder, Tuttlingen, Germany
Incubator	Heracell 240 – Thermo Fisher Scientific, Germany
Laminar flow bench	H-190 Ehret, Emmendingen, Germany
Phase contrast microscope	Axiovert 40C, Zeiss, Germany
Plasma cleaner	model: Femto A, Diener Electronic, Germany)
Ultrasonic bath	Elmasonic, ELMA, Schmidbauer GmbH, Germany
Vortex	7-2020; neoLab, Heidelberg, Germany
Water bath	Medingen GmbH, Dresden, Germany

A.1.4 Rats

1-24 h old rat pups (Wistar strain)	Breeding facilities, University of Freiburg, Freiburg, Germany

A.1.5 Microelectrode arrays

60MEA500/30iR-Ti and 60MEA500/30iR-ITO	Multichannel Systems (MCS), Reutlingen, Germany

A.2 Protocol

A.2.1 Cleaning

MEA surfaces were rinsed with deionized (DI) water and hydrophilized in humidified air plasma (10–30 min, 40 kHz, 100 W; Femto plasma cleaner, model: Femto A, Diener electronic, Germany). Freshly cleaned MEAs were used for cultures only after (at least) 2 days.

MEAs were placed in a beaker filled with distilled water and boiled for 1 h. They were then transferred into sterile Petri-dishes and the water in the chamber pipetted out. MEAs were left under the sterile hood to dry.

A.2.2 Coating and sterilization

Clean surfaces were subsequently covered with a drop of 150 µL 0.1–0.2% PEI solution and incubated at room temperature for 2 h. Non-adherent PEI on MEAs was removed in 1–3 rinsing steps with 1 mL DI water, each. Coated substrates were left to dry for at least 24 h. All substrates were then sterilized by ultraviolet radiation (2 min; UVP XX-15 s bench lamp, Upland, CA, USA) prior to their usage for culturing.

A.2.3 Preparation

Cortical tissue was extracted from brains of neonatal wistar rat pups, minced with a scalpel and transferred into ice-cold phosphate buffered saline (PBS; Invitrogen, Germany). Tissue pieces were subsequently incubated with trypsin (0.05%, 15 min at 37 °C; Invitrogen) to digest the extracellular matrix. Proteolysis was stopped with horse serum (20%; Invitrogen) and the cells were further dissociated by trituration with a serological pipette (10 mL). DNase (50 µg mL^{-1}; Sigma) was added to eliminate cell aggregation through DNA strings, if needed.

The suspension was centrifuged and the cell pellet resuspended in growth medium (see next paragraph) in a second trituration step. Cell densities were determined with an automated cell counter (CASY, Schärfe Systems GmbH, Germany).

Cells were seeded in drops of 100-200 µL suspension and left to settle and adhere for 1–2 h. Defined seeding densities were achieved by appropriately diluting the cell suspensions. Once the cells attached to the substrate, growth medium was added to a final volume of 1 mL.

MEA culture chambers were sealed with cast polydimethylsiloxane lids or with teflon membranes (ALA scientific, USA) to avoid water evaporation and contamination. Cultures developed in 1 mL growth medium comprised of minimal essential medium (MEM; Invitrogen) supplemented with heat-inactivated horse serum (5%, Invitrogen), L-glutamine (0.5–1 mM; Invitrogen), glucose (20 mM; Invitrogen) and gentamycin (20 µg mL^{-1}; Invitrogen).

Three quarters of the medium was exchanged after the first day to remove non-adherent cells and debris. During incubation, a third of the medium was exchanged twice per week. After the first week, the L-glutamine concentration was reduced to 0.5 mM. Cultures were maintained in a humidified incubator (Thermo Fisher Scientific, Germany) at 5% CO_2 and 37 °C.

Culture development was continuously inspected with a phase contrast microscope (Axiovert 40C; Zeiss, Germany). For documentation, images were taken with a digital camera (Microcular PCE-ME100, PCE Deutschland GmbH, Germany). Animal handling and tissue extraction were done in accordance with the University of Freiburg and German guidelines on the use of animals in research.

Appendix B

Reinforcement learning for autonomous control

[*] [8] Reinforcement learning (RL) involves an active agent interacting with its environment and learning to map its perceived state to actions so as to maximize a numerical reward signal. Learning a controller for a given task with reinforcement learning requires formalizing it as a Markov Decision Process (MDP). An MDP is defined as a four-tuple ($\mathcal{S}$, $\mathcal{A}$, R, P), where $\mathcal{S}$ is a set of states and $\mathcal{A}$, a set of actions. The reward function $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ defines the reward the RL controller receives when it applies action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$ and transitions into $s' \in \mathcal{S}$. The probabilistic transition model $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ defines the probability of transitioning from state s to state s' under the action a. The goal of RL is to find a control law (policy) $\pi : \mathcal{S} \to \mathcal{A}$ that maximizes the expected accumulated discounted reward, $V^{\pi}(s)$, i.e.,

$$V^{\pi}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^{t} R(s_t, \pi(s_t), s') \mid s_0 = s\right] \tag{B.1}$$

$$\pi(s) = \arg\max_{a} \sum_{s'} \mathcal{P}(s, a, s') V^{\pi}(s') \tag{B.2}$$

where $\gamma \in [0, 1)$ is a discounting factor on future rewards.

[*] Portions of Sections B.1, B.2, B.3 and B.4 were included in the subsection 'Reinforcement Learning' of our published article: Kumar SS, Wülfing J, Okujeni S, Boedecker J, Riedmiller M, Egert U (2016) Autonomous optimization of targeted stimulation of neuronal networks. *PLoS Comput Biol* 12(8): e1005054.

[8] Author thanks Jan Wülfing for reviewing this Appendix.

Value Iteration is commonly used to find V if the transition model $\mathcal{P}$ is available. In general, for neurobiological systems, such models are typically not available *a priori*. Hence, we considered a model-free setting and used Q-learning (Watkins and Dayan, 1992) to learn an action-value function $Q(s, a)$ ($Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$) which represents the value of choosing action a in state s. The greedy policy π was then derived as $\pi(s) = \arg\max_a Q(s, a)$.

To apply Q-learning, we first had to define the state and action space as well as a suitable reward function. These sets were defined slightly differently for the two control problems taken up in this thesis.

B.1 State space

In the first part (see Chapter 3), where we tackled an optimization problem in biological neuronal networks, our definition of $\mathcal{S}$, the set of states, was motivated by the following considerations. Solving the trade-off problem involved reconciling the dynamic interplay of the initiation of synchronous spontaneous bursts (SBs) in the network and the recovery of network excitability after SB termination. A simple statistical model of the initiation of synchronous SBs was a lognormal function of the period of inactivity between SBs. The cumulative of this distribution indicated the probability of SB initiation as a function of time after the preceding SB (Eq (3.2)). At the same time, recovery could equally be modelled by an exponential function of the time after the end of an SB (Eq (3.3)). Stimulation at a certain latency thus effectively probed the level of recovery at that time. This latency was defined as the quantitative state variable accessible to the learned controller, providing information on the dynamics of both processes. Therefore the time after SB termination was a simple and intuitive choice of a low dimensional state feature. We discretized this latency in 0.5 s steps, corresponding to states $1, \ldots, N$. These made up the set of states, $\mathcal{S}$, together with terminal states that reflect the outcome of the stimulation T_i (i indicating the response strength) or an "interruption" state F.

For the second part of the thesis (see Chapter 5), we pursed a different approach to achieve adaptive control of response strengths. Our data suggested that the history of response strengths and spontaneously generated events proximal to the stimulus were likely information rich features to predict the outcome of an upcoming stimulation. We used this finding to define the new $\mathcal{S}$. Event strengths detected at the recording channel was used to construct two time series, one with the spike counts in spontaneously originating bursts n^i_{sp}, and the other, with evoked response strengths, n^j_{st} (i, j indicate positions in the respective time series). At each trial t, the current state $\mathbf{s_t} \in \mathbb{R}^{2h}$, was computed by concatenating h previous response and spontaneous event strengths relative to the current trial. h was typically set to 2. In many sessions, time to the previous stimulus was added as an additional dimension to the state vector, i.e. $\mathbf{s_t} \in \mathbb{R}^{2h+1}$.

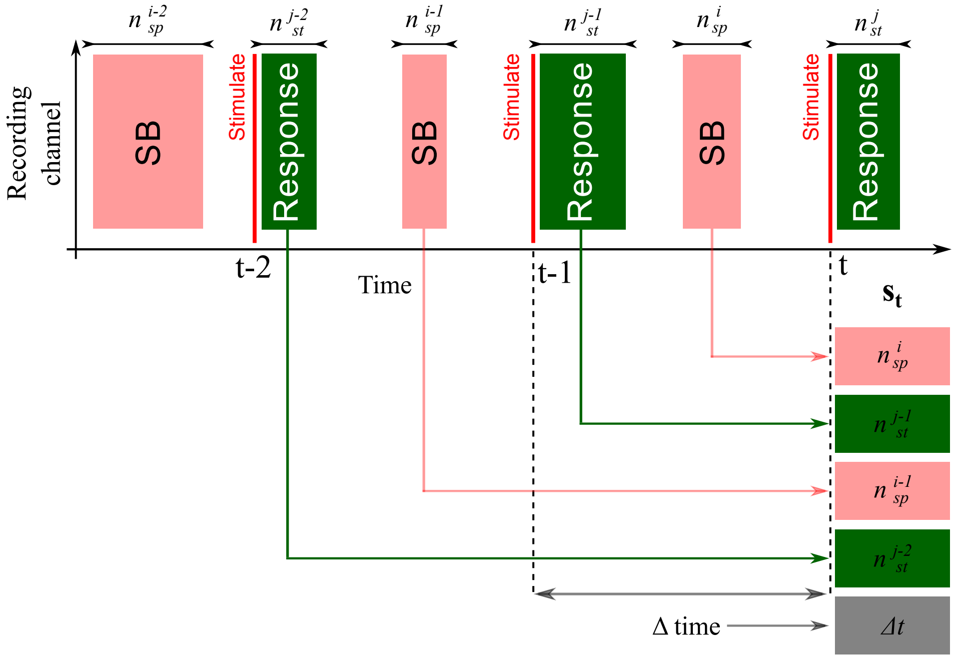

Figure B.1. The composition of the high-dimensional state used for adaptive control of response strengths is illustrated. Spontaneous bursts (SBs) at the recording electrode (RE) were detected and stimuli were delivered at various latencies relative to each SB. The history h, ($h = 2$ in this schematic) of SB and response strengths, denoted by n_{sp} and n_{st} respectively, and the last inter-trial interval (gray) together defined the state $\mathbf{s_t} \in \mathbb{R}^{2h+1}$ at stimulus trial t.

B.2 Action space

For the optimization experiments (see Chapter 3), the target was to maximize response strengths evoked at a chosen recording electrode (RE) per trial. To this end, the controller was given two choices at each state: to 'wait' or to 'stimulate'. These choices made up its action set $\mathcal{A}$.

For the adaptive control experiments (see Chapter 5), the controller was provided a larger set of actions. Candidate stimulation and recording electrodes (SEs and REs) were first identified as described in Section 2.3. Ranked response strengths evoked at each SE–RE pair were assessed to identify a suitable target response strength (Fig B.2A). In general, the target was chosen to lie in the first quartile of the ranked response strengths across SEs. This was motivated by our observation that such choices often lay below to the saturating region of each recovery function and offered the controller room to regulate response strengths. An exponential recovery model was fitted for each RE–SE pair. Based on the co-efficient of determination of the fits, an RE, few SEs (between 1 and 5) and the corresponding feasible target was finalized (see also Section 2.5.3). When multiple equally good alternatives existed, the final choices were arbitrarily made from among those. In many experimental sessions, we also included SEs that consistently failed to evoke responses at the candidate RE. These were included as known 'dud-sites' and used for a quick validation of the learned policy.

Once the RE and one or more SEs were selected we used the corresponding recovery model fits to heuristically constrain the action space. To this end, approximate stimulus latencies necessary to achieve target response strength were estimated (see Fig B.2B). An interval spanning the estimated latencies was selected and discretized. Step size varied across experiments. Values used were 0.1, 0.25, 0.3, 0.5, or 1 s. The most used step size was 0.5 s. Discrete latencies at each site were included in the action set $\mathcal{A}$. In few sessions, a choice of multiple stimulus amplitudes (0.5, 0.7 and 0.9 V) for each site and latency was also included in $\mathcal{A}$.

B.3 Reward function

In the optimization study (see Chapter 3), in order to learn the optimal stimulus latency, the controller was appropriately rewarded or punished. As shown in Fig 2.1B, within an episode, at each state the controller could choose between two actions: to 'wait' or to 'stimulate'. An episode terminated either when a pre-set maximum number of states (i.e. maximal latency) was reached, an SB occurred or when a 'stimulate' action was chosen. After each episode, the controller received a terminal reward proportional to the strength of the evoked response. Alternatively, if an SB had occurred or the

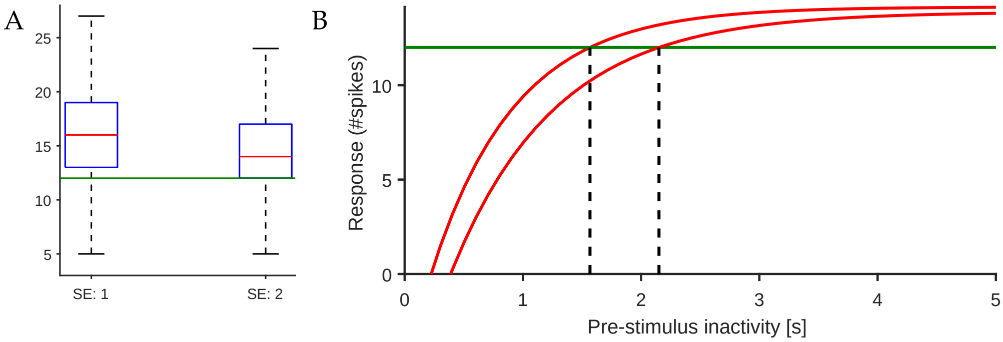

Figure B.2. *(A) The box and whisker plot shows response strengths observed at the RE during open-loop stimulation at multiple SEs as quartiles (box) and the extrema (whiskers). The target response strength (green line) was chosen to lie in the first quartile. (B) Recovery functions fitted to response strengths at the RE for two different SEs in an example network. The target was defined as 12 spikes (green line). Stimulus latencies that achieved target response strengths were estimated (black dashed lines). The range of latencies was chosen to span these values. In this case, latencies between 0.75 s and 3.75 s discretized in steps of 0.5 s were used.*

maximum number of cycles was reached it received a neutral reward (punishment):

$$R(s,a,s') = \begin{cases} i, & \text{if } s' = T_i,\ i \in \{1,\dots,n\} \\ 0, & \text{otherwise} \end{cases} \tag{B.3}$$

For the adaptive control problem taken up in the second half of the thesis, the goal was to clamp response strengths to a pre-defined value (see Chapter 5). The reward function in this case was defined as the negative absolute value of the difference between the observed and desired response strengths. If the controller was interrupted by ongoing activity, it received a constant negative reward, k.

$$R(s,a,s') = \begin{cases} -|n_{ev} - n_{goal}|, & \text{if stimulated} \\ -k, & \text{if interrupted} \end{cases} \tag{B.4}$$

where n_{ev} is the evoked response strength and n_{goal}, the target response strength. k was typically set to n_{goal}. In few sessions, $k = 0$ was used.

B.4 Q-Learning

For the optimization experiments (see Chapter 3), we used online Q-learning as the learning algorithm. It allowed us to learn a Q-function without having a model of the system dynamics, which in general is not available when dealing with biological neuronal networks.

To guarantee full exploration of the state and action space, the controller followed a random policy π_{explore} during training that uniformly chose the state of stimulation. The Q-function was iteratively updated during training sessions as:

$$\begin{aligned} Q_{t+1}(s_t, a_t) =& Q_t(s_t, a_t) + \alpha[R_{t+1} \\ &+ \gamma \max_a Q_t(s_{t+1}, a) - Q_t(s_t, a_t)] \end{aligned} \tag{B.5}$$

where $\alpha = 0.5$ was the learning rate and $\gamma = 0.99$, the discounting factor.

During testing sessions the controller follows a greedy policy without exploration:

$$\pi(s) = \arg\max_a Q(s, a) \tag{B.6}$$

For the optimization problem, since the state space for the control task at hand could be defined as a single discrete variable, a tabular representation of the Q-function was applicable, which is a prerequisite for guaranteed convergence (Watkins and Dayan, 1992). A tabular representation of the Q-function is a suitable choice as long as the biological system can be described by low-dimensional discretized states.

In the second part (see Chapter 5), where the goal was to adaptively clamp response strengths, the state-action space being high-dimensional, a tabular representation of the Q-function was not advisable due to the so called curse of dimensionality (exponentially growing memory demand). We therefore resorted to approximate RL methods to cope with the higher dimensionality. In addition, these methods also offer the ability to generalize the Q-function over the state-action space, i.e., reasonable estimates can be made not only for the state-action pairs already encountered but also for novel ones.

Batch processing methods allow efficient use of collected data while keeping the learning process stable. The technique was employed in experiments to clamp response strengths to pre-defined values. In the batch RL problem, the agent does not interact continually with the system like in online learning, but receives a finite set of state-action transitions and their corresponding rewards from the environment (Lange et al., 2012).

The following methods were employed for function approximation:

Least-Squares Policy Iteration (LSPI) Linear methods have been widely used for Q-function approximation. They are easy to implement and transparent from a debugging and feature-engineering perspective. In the context of control problems, LSPI is a popular approximate policy iteration algorithm proposed by Lagoudakis and Parr. It operates on a fixed set of samples collected by interacting with the environment and stored in the sample set $\mathcal{D}$. Each sample (s, a, r, s') indicates that executing action a at state s resulted in a transition to state s' with an immediate reward of r.

The state-action value function $Q(s, a)$ is approximated using a linear parametric combination of d basis functions (Eq (B.7)).

$$\widehat{Q}(s, a; w) = \sum_{j=1}^{d} \phi_j(s, a) w_j = \phi(s, a)^{\mathrm{T}} w \tag{B.7}$$

where w_j's are real valued parameters, each $\phi_j(s, a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, a basis function and d, the dimensionality of the restricted state-action space. In general, basis functions could be arbitrary functions of the state-action pairs. For our experiments, we directly used our states and actions to approximate Q. The greedy policy π over this approximate Q-function can be obtained as:

$$\pi(s) = \arg\max_{a \in \mathcal{A}} \widehat{Q}(s, a) = \arg\max_{a \in \mathcal{A}} \phi(s, a)^{\mathrm{T}} w \tag{B.8}$$

An initial policy π_0, represented by ϕ and w_0, is fed to a least-squares temporal difference algorithm called LSTDQ (see Lagoudakis and Parr (2003) for details) along with a set of samples for evaluation. It returns the parameters w^{π} of the approximate Q-function, $\widehat{Q}^{\pi}$ from which the new policy π can be determined in a greedy fashion (Wiering and van Otterlo, 2012). The iteration continues in the same manner until w converges (Algorithm 2).

In our experiments, we used a 'growing batch' approach, where the sample set was incrementally extended as the controller gathered more experience. In addition, to cope with non-stationarities in activity dynamics, we resorted to a sliding window approach to the controller's memory. 3000 recent transitions were retained in the sample set while discarding older ones. The discounting factor (γ) was set to 0.99.

Algorithm 2 The LSPI algorithm (adapted from Lagoudakis and Parr (2003)). After initializing the policy π_0 with parameters w_0, the LSTDQ algorithm uses the sample set to return a greedy policy π', parametrized by w'. $\mathcal{D}$ is the sample set, d, the dimensionality of the restricted state-action space, ϕ, the basis functions and γ, the discounting factor. The iteration continues until a stopping criterion, $||w - w'|| < \epsilon$, is met,.

```
$\pi' \leftarrow \pi_0$
**repeat**
    $\pi \leftarrow \pi'$
    $\pi' \leftarrow LSTDQ(\mathcal{D}, d, \phi, \gamma, \pi)$
**until** $\pi \approx \pi'$
```

Neural Fitted Q-iteration (NFQ) When non-linear functions need to be approximated, neural networks – in particular, multi-layer perceptrons – offer a promising approach. However, Q-learning, when implemented directly in neural networks with an online update rule typically suffers from long learning times and poor reliability (Riedmiller, 1999; Riedmiller, 2005).

Weight changes induced by a certain state action pair, causing unpredictable changes in other regions of the network is thought to be behind the poor performance. NFQ attempts to work around this problem by explicitly offering previous knowledge while updating a new sample. Like LSPI, it uses the batch or offline method of reinforcement learning.

In NFQ, the neural value function is updated offline based on a stored set of transition experiences. Experiences were collected in the form of triplets (s, a, s'), by interacting with the biological network and stored in a sample set $\mathcal{D}$. On the collected sample set, we used Rprop, a fast supervised batch learning method, known to be insensitive to the choice of learning parameters.

The learning algorithm consisted of two major steps: (1) The generation of the training set P and the training of these patterns within the multi-layer perceptron repeated for N_{epochs} epochs (Algorithm 3).

We used a 'growing batch' approach limited by 500 transitions to form the sample set. The sliding window approach ensured that the most recent transitions were retained. These were implemented in the functions $Collect\_sample\_trajectory()$ and $Limit\_sample\_set()$ (see Algorithm 3). The number of epochs (N_{epochs}) was set to 5 and the discounting factor (γ) used was 0.99. The number of episodes ($N_{episodes}$) varied

Algorithm 3 NFQ algorithm (adapted from Riedmiller (2005)). After initializing the multi-layer perceptron (*init_MLP*()), a training set P is generated using the collected samples. $input^i$ indicates the i^{th} state-action pair in the sample set, $\mathcal{D}$ and $target^i$ is computed as shown, using the reward function R and the current estimate of the Q-function, Q_k. The training procedure was repeated for N_{epochs} epochs. The inner loop was repeated over a pre-set number of episodes ($N_{episodes}$). The sample set $\mathcal{D}$ grew as the experiment progressed (*Collect_sample_trajectory*()), but was limited to the 500 most recent samples (*Limit_sample_set*()).

```
$episode = 0$
$Q_0 \leftarrow init\_MLP()$
repeat
    $epoch = 0$
    repeat
        $P = \{(input^i, target^i), i = 1, \cdots, |\mathcal{D}|\}$ where:
            $input^i = (s^i, a^i)$ and
            $target^i = R(s^i, a^i, s'^i) + \gamma \max_b Q_k(s'^i, b)$
        $Q_{k+1} \leftarrow Rprop(P)$
        $epoch := epoch + 1$
    until $epoch = N_{epochs}$
    $\mathcal{D} = \mathcal{D} \cup Collect\_sample\_trajectory(\pi)$
    $\mathcal{D} = Limit\_sample\_set(\mathcal{D}, 500)$
    $episode = episode + 1$
until $episode = N_{episodes}$
```

across experiments and were usually set to 40, 50 or 80.

For most of our experiments, we used a variant of NFQ that differed in the following ways:

- Instead of reinitializing the neural network after each epoch, we continuously optimized it across epochs.
- We performed mini-batch training using state-of-the art optimisers like Adam (Kingma and Ba, 2014).
- 'Dropout' was used for regularization (Srivastava et al., 2014).
- Two hidden layers with 100 rectified linear units each, were employed.

For a more detailed description of the RL approach, we refer the reader to Jan Wülfing's PhD thesis, to appear in 2018.

Appendix C

Supplementary figures

C.1 Instabilities arising from switching network modes

This section presents data from 4 networks we observed feedback instabilities arising from a switch in the ongoing network mode (Figs C.1A–D). In these networks, the fluctuations in response strengths were uncorrelated with the action sequences but not with the spontaneous component of activity recorded during the session. In all figures, the response strength sequence was smoothed with an exponential kernel ($\alpha = 0.25$) and binned over 3 or 5 min non-overlapping windows and spike counts ($\mu \pm \sigma$) at the RE in each are plotted. Figs C.1C and D include an initial non-adaptive phase (red-dashed line).

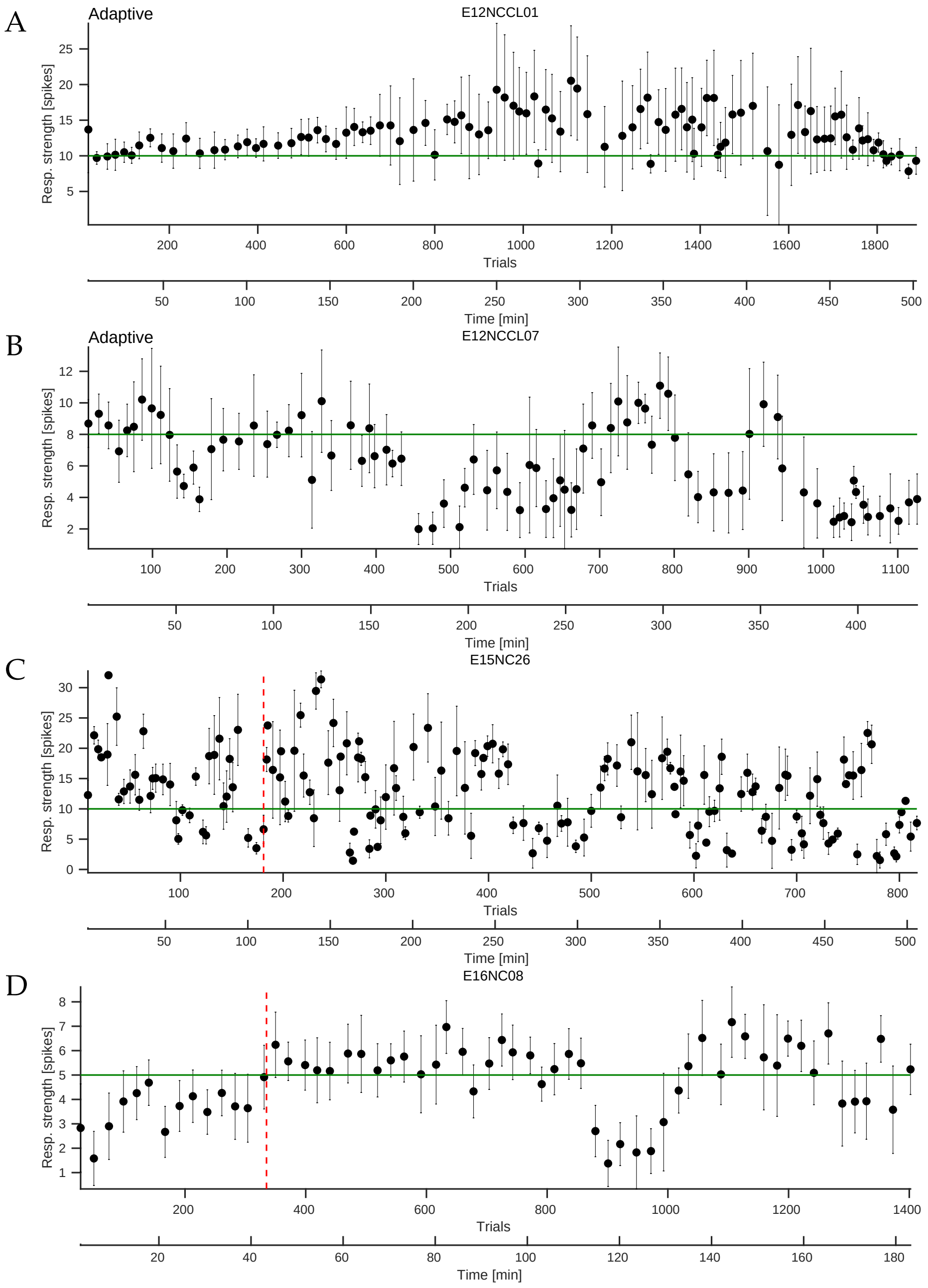

Figure C.1. *Feedback instabilities from switching network modes observed in four networks.*

C.2 Instabilities due to action dimensionality and learning delays

This section presents 6 examples of feedback instabilities arising during the closed-loop session when no mode-switches were observed in ongoing activity (Figs C.2A–F). A summary of networks studied in this configuration is reported in Fig 5.26A. Here, fluctuations in response strengths were correlated with those in the action sequences but not with the spontaneous component of activity. In all figures, binned response strengths were computed as described in Section C.1. The initial non-adaptive phase (red-dashed line) did not show signatures of instabilities.

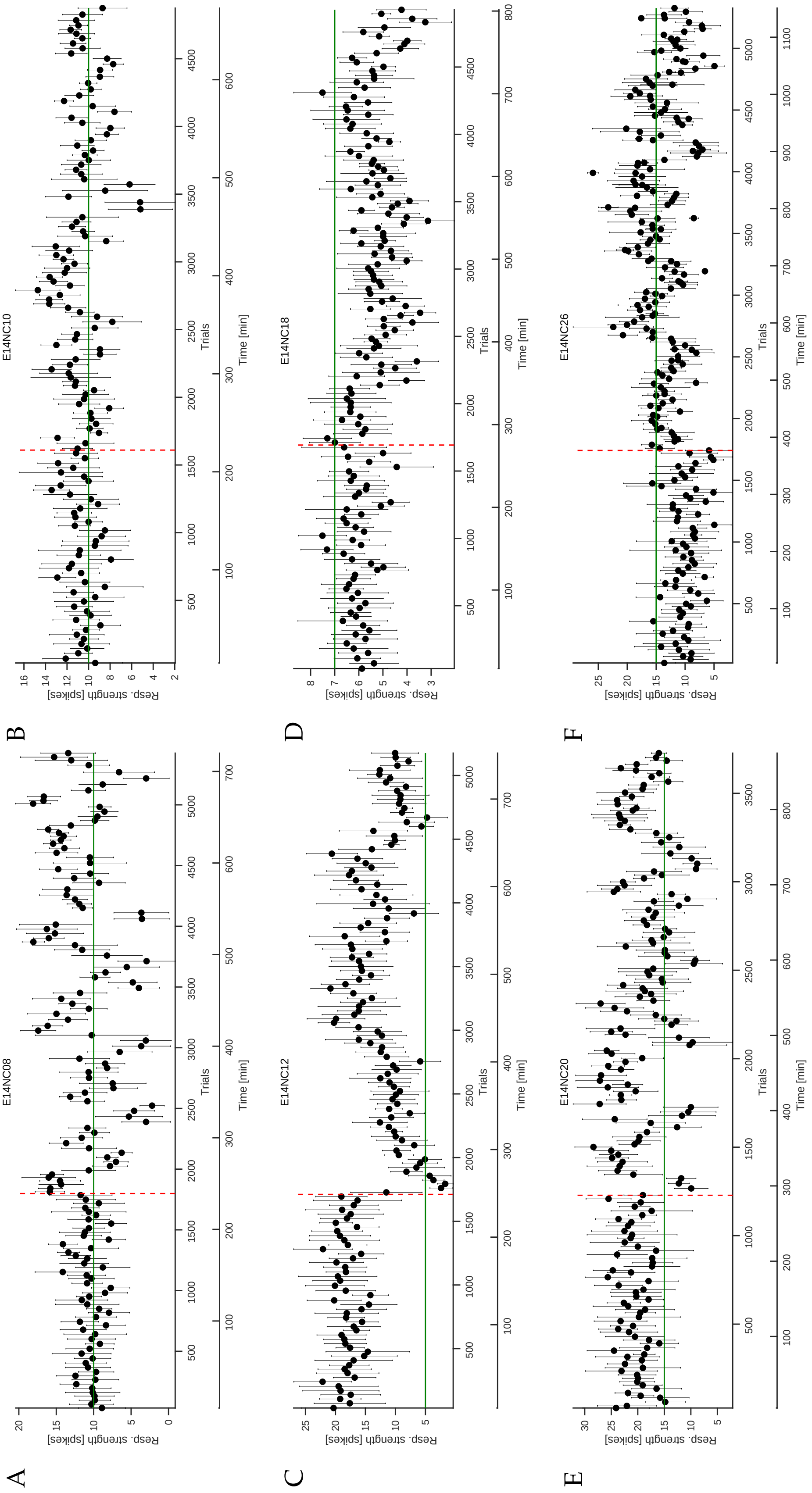

Figure C.2. Feedback instabilities from switching network modes in six example networks.

C.3 Stable high-dimensional adaptive control of response strengths

This section presents 6 examples where stable and adaptive control of response strengths was achieved (Figs C.3A–F). A summary of networks studied in this configuration is reported in Fig 5.26B. Average response strengths, pooled over 27 networks show how after learning, sustained goal-directed interaction was achieved (Fig C.4). Binned response strengths were computed as described in Section C.1.

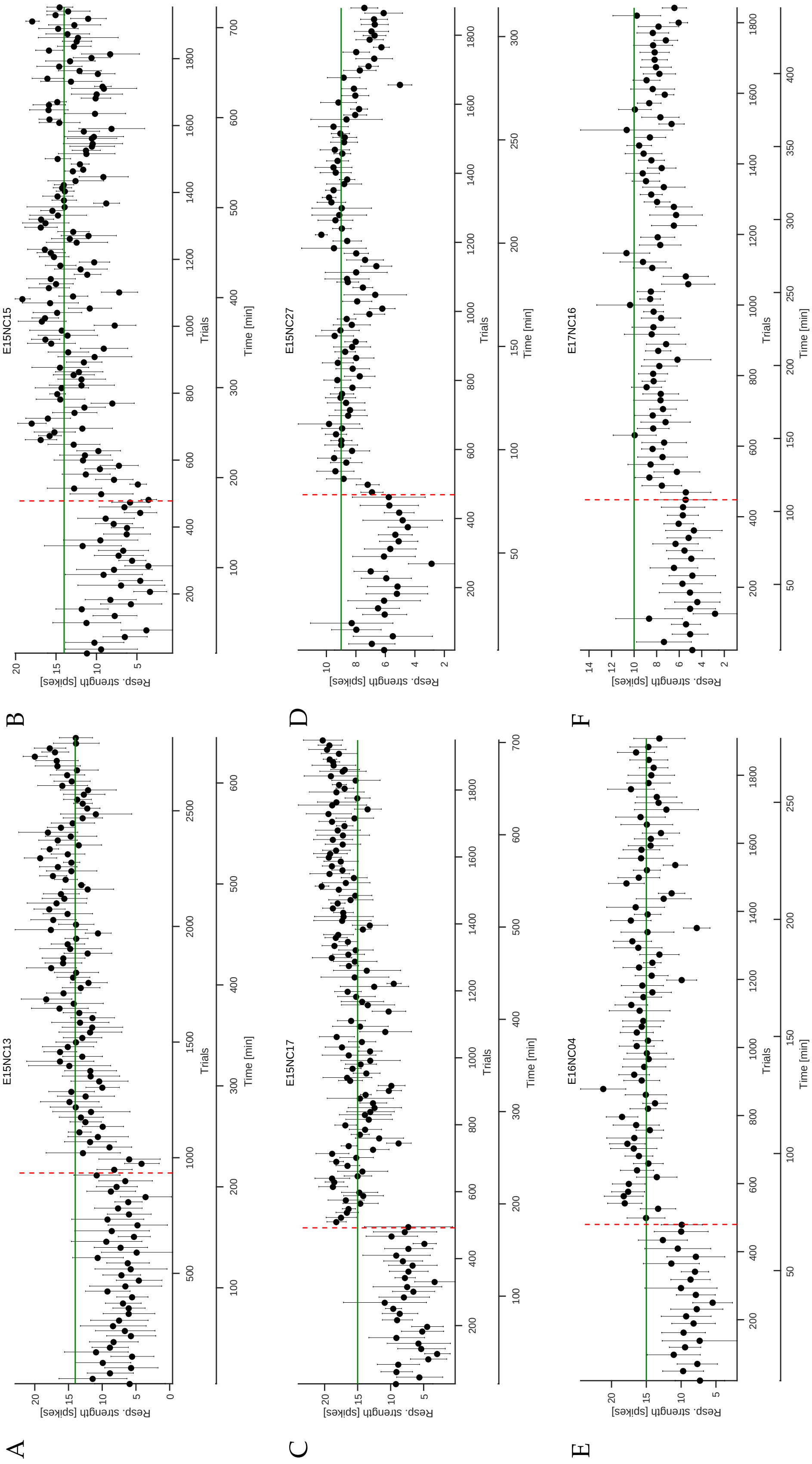

Figure C.3. Stable high-dimensional adaptive control of response strengths in six example networks.

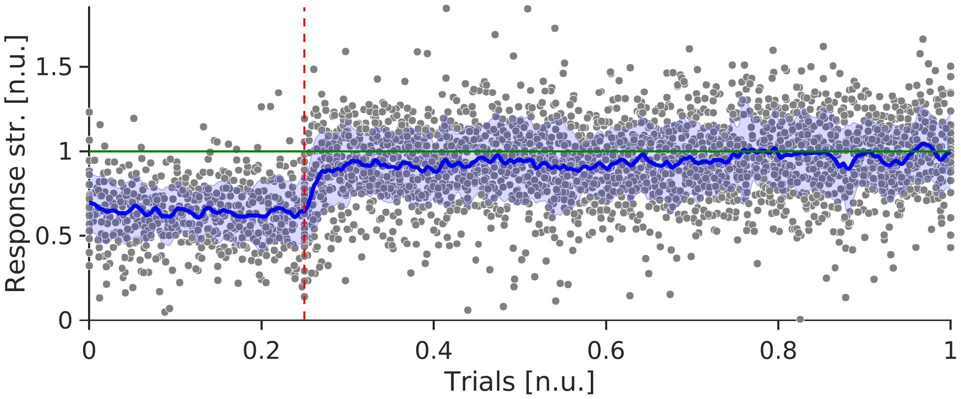

Figure C.4. *Mean response strengths normalized and pooled over 27 networks. The trial sequence in each network was normalized such that the non-adaptive phase spanned from 0 – 0.25 and the adaptive phase from 0.25 – 1. Response strengths were normalized by the target set for the respective networks (green line). Response strength levels across networks improved following the switch to the adaptive phase (red dashed line). Interactions remained goal directed thereafter (blue solid line and shading: $\mu \pm \sigma$).*

Appendix D

List of devices and software

D.1 Devices

Incubator (CB 210)	Binder, Tuttlingen, Germany
Incubator (Heracell 240)	Thermo Fischer Scientific, Germany
Laminar flow bench	H-190, Ehret, Emmendingen, Germany
MEA A/D conversion	MC_Card, MCS, Reutlingen, Germany
MEA filter amplifier	FA60S-BC, MCS, Reutlingen, Germany
MEA pre-amplifier	MEA1060-Inv-BC, MCS, Reutlingen, Germany
MEAs (grid: 6x10 electrode spacing: 500 µm electrode diameter: 30 µm)	MCS, Reutlingen, Germany
Phase contrast microscope	Axiovert 40C, Zeiss, Jena, Germany
Plasma cleaner	Femto A, Diener Electronic, Germany
Stimulus generator	STG2004, MCS, Reutlingen, Germany

D.2 Software

Closed-loop control	Python version 3.3.7, Python Software Foundation
Data analysis	Matlab R2013b – R2017b, The MathWorks, Natick, MA, USA
Data analysis	MEA-Tools (Egert et al., 2002)
Database	Microsoft Access 2013, Microsoft Corp., Redmond, WA, USA
Figure preparation	Inkscape 0.92.2, Inkscape Project
MEA recording	MEABench versions 1.1.4, 1.2.5 (Wagenaar et al., 2005)
Operating system	Ubuntu 16.04, Canonical Ltd., Ubuntu community
Operating system	Ubuntu 10.04 (kernel version `2.6.32-38-generic-pae`), Canonical Ltd., Ubuntu community
Operating system	Windows 7, Microsoft Corporation, Redmond, WA, USA
Reinforcement learning	CLS^2 (Closed-loop Simulation System) version 4.0
Text processor	TeXstudio 2.12.6 with TeX 3.14159265 (TeX Live 2016/Debian)