A STOCHASTIC CONTROL FRAMEWORK FOR THE DESIGN OF OBSERVATIONAL BRAIN-COMPUTER INTERFACES BASED ON HUMAN ERROR POTENTIALS

BY

DAVID A. STEINES

THESIS

Submitted in partial fulfillment of the requirements for the degree of Master of Science in Electrical and Computer Engineering in the Graduate College of the University of Illinois at Urbana-Champaign, 2011

Urbana, Illinois

Adviser:

Assistant Professor Todd Coleman
Abstract

Brain-computer interfaces (BCI) allow human subjects to interact with exogenous systems through neural signals. The goal of this interaction may be to induce behaviors or properties in either the exogenous system or the human subject. In this thesis, we develop a novel framework for designing BCIs based on principles from adaptive control. In particular, we exploit scalp electroencephalography-derived correlates of the human error processing system to recover a subject’s desired policy for the exogenous system. This scheme allows a human subject to control a system through passive observation by critiquing actions taken by the system. We provide a necessary and sufficient condition for convergence and simulations as a proof of concept. Further, we discuss the application of this framework to building co-adaptive BCIs and as a tool for understanding the learning process during BCI interaction.
The fox knows many things, but the hedgehog knows one big thing.

–Archilochus
Acknowledgments

There are always many hands in any project, particularly interdisciplinary efforts such as this. First, I owe thanks to the National Science Foundation Integrative Education and Research Trainee (IGERT) program in Neuroengineering and its visionary principal investigators – Doug Jones, Todd Coleman, Monica Fabiani, and Bob Wickesberg. The IGERT program has introduced me to brain-computer interface research and facilitated my initial foray into cognitive neuroscience and psychophysiology. This thesis would not exist in the absence of this program. I am indebted to my adviser, Todd Coleman, for his deep theoretical knowledge and enthusiasm. I also appreciate the conversations with the many researchers that I work alongside in the BCI Lab. Finally, I owe thanks to my family for their constant support and encouragement.
# Table of Contents

List of Figures .................................................. vi

Chapter 1  Introduction ........................................ 1

Chapter 2  Neural Correlates of Prediction Error and Learning .... 4
  2.1 Survey of Human Error Potentials ........................ 5
  2.2 Equivalence of oFRN and FRN .............................. 8
  2.3 A Model for Error Potential Dynamics .................... 12

Chapter 3  A BCI Stochastic Control Framework .................. 14
  3.1 Canonical Q-Learning .................................... 14
  3.2 Modeling the Human Observer as a Q-Learner .......... 17
  3.3 Problem Statement ....................................... 19
  3.4 Bayesian Control Rule ................................... 20
  3.5 Derivation of Likelihood and Intervention Models .. 22
  3.6 Necessary and Sufficient Conditions for Convergence .. 23

Chapter 4  Simulations and Results ............................ 28
  4.1 Methods .................................................. 28
  4.2 Results .................................................. 29

Chapter 5  Conclusion ........................................... 32

References ......................................................... 33
List of Figures

3.1 Reinforcement Learning. An agent perturbs an environment by eliciting actions, while the environment generates the next state and a reward. ........................................ 15

3.2 A human observer Q-learns the combined dynamics of the agent and environment. The observer critiques the agent through the TD error, $\delta_t$. ................................................................. 17

3.3 A human observer generates error signals, $\delta_t$, according to the parameter $w^*$. The goal is to design a BCI agent that is optimal with respect to $r^*$, the reward function of the observer. ................................................................. 20

4.1 The waterfall game. At each time step, the obstacle in row 1 falls down to row 2, and a new obstacle may be introduced randomly into one of the lanes in row 1. The agent can choose either to remain in his lane or to switch to the other lane. If the agent avoids the obstacle, he stays green. If he hits the obstacle, he turns red. ............................... 28

4.2 Posterior on operation modes after 200 trials. After 200 trials, the BCR places its probability mass on several operation modes. Note that most of the mass is on the true mode, 14. The next plot shows that these operation modes are in an equivalence class, and that the algorithm has converged. ................................................................. 30

4.3 Posterior on reward functions after 200 trials. After 200 trials, the BCR positions all of the probability mass on modes with the true reward function. ........................................ 31

4.4 Probability of the true reward vs. trials. The plot shows the value of the posterior on operation modes containing the true reward against trials. ........................................ 31
Electroencephalography (EEG) has been used by cognitive psychophysicists since the early part of the 20th century to link scalp potentials to human behaviors [19]. These scalp potentials are believed to be produced by superimposing electric fields induced by currents in the cortex of the brain [33]. A sizable library of voltage deflections and rhythms have been discovered that are reliably elicited in a laboratory setting and correlated with human behavioral events [5]. Recently impelled by cheap digital signal processing hardware and clinical motivation, investigators have detected these signals during on-line tasks to control exogenous systems and provide feedback to subjects. Through these applications the nascent field of EEG-based brain-computer interfaces (EEG-BCI) has emerged.

EEG-BCI consists of a wide variety of paradigms that permit communication and control of exogenous systems. In 1988, Farwell and Donchin introduced the P300 speller, which allows subjects to spell sentences by staring at a grid of flashing letters [16]. When the letter they wish to write is illuminated, subjects experience surprise and evoke the P300 ERP. By detecting these voltage deflections on multiple trials, the intended letter can be inferred. Other BCI paradigms exploit different signals. The mu-rhythm motor paradigm allows subjects to control systems through motor imagery [2]. By imagining squeezing the left or right hand, neuronal populations in the cortex can be desynchronized. This event-related desynchronization can be detected in the power spectrum of the 10 Hz mu-rhythm. In general, EEG-BCIs such as the P300 speller and motor imagery are an emerging modality for communication and control of exogenous systems.

There are compelling clinical reasons to study EEG-BCI. At present, it is not possible to restore normal motor function in people with progressive or acute neurological conditions such as Amyotrophic Lateral Sclerosis (ALS), Parkinson’s, or stroke. A significant amount of research has concerned the
application of EEG-BCI to restore motor function in patients by facilitating control of computer cursors and robotic arms through EEG signals [44]. Control of exogenous systems using EEG has been a hallmark of the field since its inception, and it continues to comprise the bulk of new literature entering the field today.

In recent years, attention has shifted to the adaptive nature of the brain during BCI interaction. Several studies have shown that subjects can learn to volitionally modulate or amplify features of their EEG over time through appropriate feedback [2]. Researchers have hypothesized that BCIs might be able to restore normal or close-to-normal motor function by inducing activity-dependent brain plasticity [11]. This hypothesis has motivated a new philosophical perspective that considers EEG-BCI as a problem of co-evolving coupled dynamical systems. Research in this domain suggests that it may be possible to guide learning by training people to modulate or elicit specific scalp potentials. To date, there is mild evidence that this may be possible; but conclusive proof and useful models are lacking [11]. The problem is made difficult because the signals that should be targeted during therapy and the nature of brain-machine co-adaptation are not well understood.

Understanding human-machine co-adaptation is not restricted to clinical applications. EEG signals are inherently non-stationary, and the brain is a complex system that processes multiple impinging stimuli in parallel. If the EEG is truly reflective of cortical activity, many of these processes may contribute to measured scalp potentials. In order to parse this complex signal in a principled manner, it is useful to develop robust control frameworks that both achieve systems engineering objectives and allow for scientific hypotheses about observed signals to be tested.

This thesis attempts to achieve these dual scientific and engineering objectives by developing a principled framework for understanding human-BCI co-adaptation while building a novel BCI based on passive observation. A recent body of literature has identified a class of error potentials that are elicited by a generic, high-level error processing system [21]. Termed the Feedback-Related Negativity (FRN), these error signals are elicited when unexpected outcomes occur during tasks of choice economy [7]. Recent evidence has also indicated that these signals are elicited during observation of an agent that acts on a subject’s behalf [39]. It has been proposed that the FRN is reflective of a reinforcement learning process in the human subject
under both the active and passive observation conditions [21]. Further, it has been shown that the FRN scales with prediction error, suggesting that it encodes the hidden policy that the subject wishes to impose on the BCI agent. By modeling the observer as a reinforcement learning system, principles from stochastic control can be applied to design BCI agents that extract this hidden policy and act optimally with respect to each subject.

This thesis will develop a principled framework to design BCI agents that are optimal with respect to human learners by measuring these prediction error signals. Chapter 2 will review the neural basis of prediction error signals. Further, it will introduce the Feedback-Related Negativity (FRN) and argue that the observational form of the FRN (oFRN) is generated by the same error processing system. Chapter 3 will formalize the reinforcement learning model of the FRN and describe the BCI design problem in detail. The Bayesian control rule will be introduced as a technique for designing co-adaptive BCIs. A necessary condition for convergence will be presented, and sufficient conditions will be summarized. Chapter 4 will present proof-of-concept simulations. Chapter 5 will discuss future work and potential applications. The appendix contains detailed mathematical derivations and code to implement the simulations.
In order for learning to occur, humans must be able to evaluate the outcomes of their actions and make predictions about how an environment will respond to perturbations. For instance, while a subject interacts with a BCI agent, the BCI algorithm elicits feedback. The subject must use this feedback to decide how issuing a certain command has worked and predict how it will tend to work in the future. Recent studies confirm that subjects change their strategy relative to feedback [7]. By incorporating signals that reflect learning into the control scheme, BCI agents can be made optimal with respect to each unique learner, and sensitive to ongoing behavioral adaptation.

While many brain regions and systems participate in the evaluative process underlying human learning, recent psychophysiological studies have identified a generic distributed error processing system that is associated with scalp potentials elicited during tasks of choice economy [30]. Although the system has been termed “error-processing,” this is a bit of a misnomer. Rather, it has been hypothesized that it is a general reinforcement learning system through which a human characterizes the dynamics of their environment and generates predictions of how much utility can be acquired from the environment under certain actions [21]. Over the short term, this system guides action selection and is hypothesized to play a role in longer-term integrative learning through interaction with the ventral striatum of the basal ganglia [21].

Recent studies have identified scalp Event-Related Potentials (ERPs) that reflect this model-building process and are readily detected on single trials using time-frequency signal processing techniques. Specifically, these error potentials have been correlated with theta power over medial and lateral electrode sites during learning [7]. While several studies have explored the application of error potentials to BCI, none has considered theta power on single trials and its application to a co-adaptive BCI architecture, to this
author’s knowledge [26, 9]. In this chapter, we review the neural basis for 
these ERPs and justify the reinforcement learning model that underlies their 
elicitation. We argue that these error potentials – notably the observational 
feedback-related negativity – can be applied to a stochastic control-based 
BCI to train agents through passive observation.

The chapter is organized as follows. Section 1 surveys the major human 
error potentials. Section 2 summarizes several studies pertaining to the ob-
servational feedback-related negativity, and argues that the oFRN and FRN 
are essentially equivalent signals. Section 3 presents a quantifiable model for 
the FRN that can be used to design brain-computer interfaces.

2.1 Survey of Human Error Potentials

Studies investigating error-related EEG signatures have identified a class 
of error ERPs, including the Error-Related Negativity (ERN), Feedback-
Related Negativity (FRN), and Observational Feedback-Related Negativity 
(oFRN). In this section, we review the literature concerning these ERPs and 
justify the reinforcement learning model of the oFRN.

2.1.1 Error-Related Negativity

The Error-Related Negativity (ERN) was jointly discovered in 1990 by sepa-
rate research groups – those of Falkenstein and Gehring [13, 17]. On epoch-
averaged error trials during choice-reaction time experiments, a 10 $\mu$V neg-
avative deflection was observed in frontocentrally located EEG sites approxi-
mately 100 ms after commission of an error. The ERN was shown to scale 
with the magnitude of the commission error, and it was hypothesized that 
the ERN is reflective of a neural system that both detects and compensates 
for errors [17].

Although initially thought to be elicited only on error trials, subsequent 
studies presented evidence that its proposed generator – the anterior cingu-
late cortex (ACC) – was also activated on correct trials [6, 14]. In a follow-up 
study by Falkenstein, a small negativity was noticed after some correct trials 
[14]. In light of this result, later studies proposed that the magnitude of the 
ERN is not reflective of the absolute value of outcomes but rather a devi-
ation in expected outcome [32, 46]. This evidence suggested that the ERN encodes a type of prediction error and led to Holroyd and Coles’s reinforcement learning model [21].

Holroyd and Coles’s reinforcement learning theory of the ERN is based on studies implicating the basal ganglia and midbrain dopamine system in reinforcement learning. According to this theory, the basal ganglia function as a critic that evaluates events and predicts the value of outcomes. When events deviate from expected, the basal ganglia induce phasic changes in the activity of midbrain dopaminergic neurons. These phasic increases and decreases are prediction error signals that indicate the magnitude of deviation in expectation. The prediction error signals are projected from the basal ganglia to the ACC and prefrontal cortex, where they are used to induce behavioral adaptation by compensating executive control systems.

Despite evidence in favor of Holroyd and Coles’s theory of the ERN, an alternative theory known as the Conflict Monitoring Hypothesis has been offered as an explanation [3]. According to this model, the ACC detects conflicts in information processing and recruits executive control systems to resolve them. Conflict is considered the simultaneous activation of incompatible processes. The essence of this model is that it treats the ACC as a passive monitor that attempts to resolve incongruities. This account is distinct from the model advocated by Holroyd and Coles, because the conflict monitoring hypothesis does not treat the ACC as implementing a response-selection function that weights competing motor controllers.

To date, debate remains as to which hypothesis is true. However, several studies of the ERN’s close cousin – the feedback-related negativity – provide compelling evidence that Holroyd and Coles’s reinforcement learning model is a useful framework for characterizing the dynamics of these error potentials.

2.1.2 Feedback-Related Negativity

Several years after discovery of the ERN, a related error signal termed the feedback-related negativity (FRN) was reported by Miltner, Braun, and Coles [31]. Subjects were tasked with detecting an auditory cue, and pressing a button after they thought a second had passed from cue onset. 600 msec after the button press, subjects were presented with an auditory, vi-
sual, or a somatosensory indication of either a correct or incorrect estimate. Approximately 230 to 330 ms after incorrect feedback, epoch-averaged ERP waveforms showed a negative deflection lasting approximately 260 ms that was largest at medial electrode sites. The study presented evidence that this deflection, termed the feedback-related negativity, was related to the ERN but distinct in that it was invoked following feedback as opposed to error commission.

It was postulated that the ERN and FRN are elicited by the same system and that the reinforcement learning model of the ERN would also model the FRN. Several follow-up studies showed that FRN amplitude covaried with expectation deviation in medial electrode sites [20, 24, 23, 22, 25, 1]. These studies investigated a variety of tasks – monetary gambling, choice-reaction, and Erikson Flanker. The reliability of the ERN and FRN in tasks of choice economy suggests that it is elicited by a system that is able to adapt to specific paradigms as necessary. Holroyd and Coles investigated this issue and found that the error processing system is generally context-dependent [23]. Holroyd and Coles’s result is readily understood in the context of reinforcement learning. All reinforcement learning problems are context-dependent, in that agents bring unique reward functions to each problem. The prediction error signals (FRN, ERN, oFRN) are functions of this reward and hence reflect context-dependence.

There has been some debate about the relationship between the FRN, P300, N2, various visual evoked potentials, and other non-error-related components. In an effort to understand how the FRN responds to reward magnitude and valence as opposed to the P300, Yeung had subjects complete a simple gambling task. They found that the P300 is sensitive to reward magnitude but insensitive to reward valence and the feedback negativity exhibits the opposite behavior. This finding suggests that the P300 and FRN originate in separate systems that perform distinct functions [46]. For the purposes of this thesis, the relationship between the FRN and other ERPs is not essential.
2.2 Equivalence of oFRN and FRN

Traditional ERN and FRN studies have investigated ERP elicitation during tasks when the subject is able to act on his or her own behalf. However, since the reinforcement learning model suggests that subjects build models of expectation, it is possible that ERNs and FRNs can be elicited during observation of others since agency is not required to acquire utility. Several recent studies in observational learning and social cognition have considered whether the FRN is elicited during observational learning. A wide body of literature addresses the role of mirror neurons – neural circuits that are active in both the observer and observed during monitoring tasks [38, 4]. Some of these circuits have been attributed to systems that evaluate or predict the actions of others [37]. While the literature on the oFRN is sparse and its relation to the FRN is not entirely clear, mirror neuron studies provide some evidence that the oFRN is reliably elicited during observation. Here, I argue that the oFRN is essentially equivalent to the FRN, but it is amplified by a sense of agency and motivation.

Although not the first paper exploring the oFRN, Donkers designed a slot machine task in which participants did not make choices [12]. This paradigm ensured that the elicitation of the FRN would not be contingent on preceding choices. They observed FRN-like mediofrontal negativity associated with outcomes. Further, they found that the observed FRN was elicited whenever a stimulus was different from the preceding stimulus, irrespective of whether that stimulus averted a loss or a gain. The general morphology of the mediofrontal negativity that they observed was similar to that of the FRN, suggesting that they are elicited by similar processes. Further, they found that the mediofrontal negativity was more right lateralized – but this may be task specific, as monetary gains and losses have often been more right lateralized.

An earlier paper by Miltner et al. studied error potentials for subjects that observed a simulated subject performing a choice reaction time task [30]. Observation was not purely passive, since subjects were instructed to either count the number of errors committed by the simulated subject or to press a button when an error was committed. The response to observed errors was notably similar to the FRN. In particular, they found that the P300 was not enhanced following error observation and that the negativity
was an interruption of a positive wave (P300) approximately 200-350 ms after error commission.

While Donkers and Miltner studied the oFRN under gambling and choice-reaction tasks, Van Schie conducted an Erikson Flanker task, during which subjects both performed the task and observed an experimenter perform the task [39]. Observers were instructed to count the number of errors committed by the experimenter. On observation of incorrect trials, subjects elicited a delayed onset, reduced amplitude, and elongated ERN (230 ms).

Are these observational error potentials localized to ACC as the ERN and FRN? In a study conducted by Shane, subjects participated in a speeded go-nogo task, and then watched an actor perform the same task [40]. The study provided hemodynamic evidence that the ACC is implicated during the observation of another’s errors, suggesting that similar neural circuitry is involved in both self- and observer-commissioned errors.

Yu considered oFRN elicitation during observation of a simple monetary gambling task [48]. As in the case of Donkers, Miltner, and Van Schie, the oFRN was elicited with reduced amplitude. It is notable that under Erikson Flanker tasks, reaction-time tasks, and gambling tasks the oFRN is elicited with a topology and morphology approximate to that of the FRN.

The studies by Miltner, Van Schie and Yu provided compelling evidence that the oFRN is elicited by systems responsible for the FRN. However, during each of these studies, subjects were instructed to count or were provided with a monetary reward. These features make it unclear under which conditions the observational error system is activated. It is essential that the BCI practitioner be able to construct paradigms in which the oFRN is elicited reliably.

To further clarify the conditions under which observational error potentials are elicited, Koban studied how differences in social context could influence the response to observed errors [28]. Participants performed a modified go-nogo task, in which one subject performed the task while another observed. Subjects either cooperated or competed (the observer either desired the active subject to succeed or fail). On error trials during cooperation, subjects tended to elicit an early oFRN component at approximately 130 ms. Further, this oFRN was fronto-centrally located and a deflection in a larger positivity that peaked around 150 to 220ms. They attributed this positivity to an overlap of the visual evoked N1 and P2. This positivity is similar to the
positivity discovered by Falkenstein during active-choice tasks and provides further evidence that the oFRN and FRN are elicited by the same neural system [14].

While the oFRN is elicited during observation, there is evidence that purely passive observation is not sufficient to elicit the signal. Several studies have directly sought neural correlates of agency and empathy during observation tasks [15, 42]. Kang investigated the relationship between self-other overlap and oFRN amplitude [27]. It was demonstrated that oFRN amplitude is significantly reduced when observing nonhuman agents. This result is of particular importance to the BCI practitioner, since it indicates that empathy and self-other overlap may play a significant role in the human learning process and affect signal quality. Although not considered in this thesis, the social context in machine-based learning may be a central issue for the BCI practitioner, particularly if the oFRN is reflective of a learning process that is fundamentally different from human-human interactive learning.

In another study investigating the role of agency, Yeung observed feedback-related negativities in monetary gambling tasks in which subjects both made and did not make choices [45]. The amplitude of the component was reduced in these tasks relative to a task in which the outcomes were contingent upon participants’ choices. This finding suggests that the amplitude of the FRN is sensitive to both subject motivation and sense of agency.

Furthering the argument that agency is essential, Bellebaum argues that the processing of feedback stimuli depends upon the direct relevance for one’s own action planning [1]. Hence, reduced amplitude error potentials are observed in the observational case. Further, FRN amplitude was modulated by reward expectancy during observation, although the difference in amplitude on less probable negative outcomes was more pronounced in the active feedback condition.

The aforementioned studies suggest that the oFRN is reliably elicited in a variety of tasks – including choice reaction time, go-nogo, Erikson Flanker, and monetary gambling. In general, this error potential is of a similar latency and morphology with that of the FRN, although it appears to be modulated and amplified by sense of agency, self-other overlap, and motivation. Topologically, it is elicited in medial electrode sites. In general, subjects interact with BCI systems because they want the system to do or acquire something. Hence, they are generally invested in its performance. This natural coop-
eration suggests that the problem of agency may not be significant for BCI practitioners, but since the oFRN has not been studied as widely as the FRN, it is an important design consideration.

The reduced amplitude of the oFRN poses a special problem for the BCI practitioner. ERPs are generally difficult to detect on single trials. Hence, most paradigms – such as the P300 – use signal averaging to increase SNR. While one approach to this problem is to develop novel signal processing techniques for single-trial ERP detection; another solution is to find correlations between the oFRN and other signal measures. The next section addresses this possibility.

2.2.1 FRN Correlates with Medial Theta Power

Studies investigating the frequency characteristics of the ERN [29] have shown that the ERN reflected enhanced theta activity (4-8 Hz) following incorrect responses. As further evidence, Luu and Tucker observed an FRN in response to error feedback after band-pass filtering their data in the theta range, suggesting that much of the energy in the FRN is concentrated in the theta frequencies [29].

Based on this observation and evidence that the ERN and FRN are elicited by the same error processing system, Cohen hypothesized that increased theta activity over medial electrodes would occur on error trials [10]. They observed that enhanced theta power and ERPs occurred following wins. This study was also notable because they observed larger ERPs under the condition when rewards were relatively infrequent, and smaller ERPs during frequent rewards. This finding suggests that the FRN scales with expectation – hence, as the subject learns about the dynamics of the environment or builds an accurate model of the environment, the tendency to elicit large FRNs is reduced. This result further supports Holroyd and Coles’s reinforcement learning model of the FRN, and provides evidence that theta power is a reliable correlate of the FRN.

The correlation between prediction error signals and theta power is particularly powerful, since signal processing methods to detect theta power are simple to implement. Cohen and Cavanagh have classified FRNs on single trials using complex Morlet wavelets and instantaneous power computed us-
ing Hilbert transforms [7]. These algorithms can be easily implemented in an online BCI.

### 2.3 A Model for Error Potential Dynamics

While the aforementioned error potentials have been strongly correlated with prediction error on epoch-averaged trials, the reinforcement learning model of the FRN predicts that their magnitude should scale with the subject’s deviation in expectation on single trials [21]. This hypothesis has not been widely tested, since the ERP technique relies on epoch-averaged trials that preclude single-trial analysis. While the amplitude of epoch-averaged trials under distinct conditions can be compared, single-trial techniques are not widely applied in psychophysiology.

For BCI practitioners, this has led to the adoption of a simplified model of the FRN. BCI applications that use the FRN often treat it as a binary error signal [36, 26]. While this model can be useful, it arbitrarily discards essential information. If the FRN is a prediction error signal, then it implicitly encodes a hidden variable – the subject’s estimate of the agent’s behavior. Hence, attempting to minimize errors will not guarantee convergence to the subject’s desired policy. When prediction error signals cease to be elicited, this means that the agent is behaving as the subject expects, which is not the same as behaving as the subject desires. Hence, our goal is to design a BCI that extracts the subject’s desired policy. This requires that we model the subject as a learning system and attempt to identify his hidden policy.

By exploiting the relationship between theta power and the FRN, it is possible to implement this model. A recent study has shown that theta power in medial electrode sites scales with outcomes that are worse than expected, while theta power in lateral electrodes tends to scale with outcomes that are both better and worse than expected [7, 8]. In that same study, both theta power and theta phase synchrony between medial and lateral sites were shown to predict behavioral adaptation in subjects that were learning to interact with a stochastic environment. By measuring the combined effect of medial and lateral theta power, the prediction error on single trials can be measured and integrated into a BCI.

While Holroyd and Coles simulate an adaptive critic-actor architecture
as their generative model of the FRN, Cavanagh models the subject as a Q-Learner with a time horizon of individual trials [7, 21]. Q-Learning is a special case of temporal difference learning, and hence these models are closely related. By modeling the human as a Q-Learner, we can exploit its simplicity to gain intuition about how humans learn relative to a BCI. Further, both Cavanagh and another recent study have shown that theta power scales with prediction error on single trials [7, 35]. To date, these are the only two studies that have investigated the reinforcement learning model of the FRN on single trials. Hence, we will model our human observer as a Q-Learner in the sections that follow.

We note that our model hinges on an important assumption that to this author’s knowledge has not been investigated. Notably, we assume that subjects will elicit theta power under the same conditions as in Cavanagh during passive observation. While Cavanagh and others have shown that theta power correlates with prediction error signals, all of these studies have concerned cases where subjects were able to act on their own behalf. As presented earlier in this thesis, there is ample reason to believe that this will happen, since the oFRN generally maintains a morphology and topological distribution similar to that of the FRN. However, we will implicitly test this hypothesis in the scheme that follows.
Chapter 3
A BCI Stochastic Control Framework

In this chapter, we will formalize the model developed in the previous chapter to design a BCI based on human error potentials. The first section will review Q-Learning, since it serves as a model for the human observer. The second section will discuss the specialization of Q-Learning to the BCI design problem. The third section will propose a control framework based on the Bayesian control rule to design the BCI. The fourth section will derive distributions that are necessary for application of the Bayesian control rule. The fifth section will provide necessary and sufficient proofs of convergence. Detailed mathematical derivations and code have been relegated to the appendix.

3.1 Canonical Q-Learning

Reinforcement learning theory considers the problem of an agent that is interacting with an environment in order to obtain a certain amount of utility. For instance, consider a fish looking for food in a shark tank. At each time point, the fish must observe the environment and consider the location of the shark and potential locations of food rewards. Then, it must make a decision – should it swim closer to the food, or away from the shark? After making its decision, it changes its position relative to both the food and shark and must prepare to take another action. The goal of the fish is to maximize its utility by obtaining food rewards in the fewest steps possible while avoiding the shark. However, the fish does not know how the shark will react to each of its movements. The fish does not know how the environment (the shark) will change with each of its decisions, nor does it know how much utility it will acquire with each decision. The fish must learn how to pick actions based on past experience in order to maximize its utility. This is a problem
of reinforcement learning.

Referring to figure 3.1 above, note that the reinforcement learning problem consists of two interacting systems – an agent and environment. In most reinforcement learning problems, the environment is modeled as a Markov Decision Process (MDP). An MDP is a tuple, $M = \{S, A, P, \gamma, R\}$ where:

- $S$ is a finite set of states
- $A$ is a finite set of actions
- $P$ is a state transition probability matrix
- $\gamma$ is a discount factor bounded by $0 \leq \gamma \leq 1$
- $R : S \times A \rightarrow \mathbb{R}$ is a reward function
- $\pi : S \rightarrow A$ is the policy which maps states to actions

We make use of the following notations and constraints:

- $S_{t+1}$ is a random variable, where $S_{t+1} \sim P(S_{t+1}|s_t, a_t)$
- $R_{t+1}$ is a random variable, where $R_{t+1} \sim P(R_{t+1}|s_t, a_t)$
- Outcomes $r$ of $R$ are bounded by $0 \leq r \leq 1$

The goal of the agent is to learn a policy $\pi$ that achieves a certain amount of utility. In general, the goal is to learn a policy that maximizes the expected cumulative discounted utility. Such a policy is said to be optimal and is
denoted by $\pi^*$. When the reward function, $R$, and the transition environment dynamics, $P$, are known, the agent can solve for $\pi^*$ using Bellman’s Equation:

$$V^*(s) = \max_a \left[ r(s, a) + \gamma \sum_{s'} P(s'|s, a)V^*(s') \right]$$  \hspace{1cm} (3.1)

For convenience, we will define the state-action or Q values as follows:

$$V(s) = \max_a Q(s, a)$$  \hspace{1cm} (3.2)

Substituting into equation (3.1) we have,

$$\max_a Q(s, a) = \max_a \left[ r(s, a) + \gamma \sum_{s'} P(s'|s, a) \max_{a'} Q(s', a') \right]$$  \hspace{1cm} (3.3)

from which it follows that,

$$Q(s, a) = r(s, a) + \gamma \sum_{s'} P(s'|s, a) \max_{a'} Q(s', a')$$  \hspace{1cm} (3.4)

This equation is simply a restatement of Bellman’s equation in terms of the state-action values. If we return to the learning problem stated at the beginning of this section, we note that the agent does not know the reward distribution or the state transition dynamics, leaving equation (3.4) of little direct value. However, we can define the observed reward and transition probabilities as follows:

$$r(s, a) = E [R_{t+1}|s_t = s, a_t = a]$$  \hspace{1cm} (3.5)

$$P(s'|s, a) = P(S_{t+1} = s'|s_t = s, a_t = a)$$  \hspace{1cm} (3.6)

which results in the following iterative update expression for the Q-Values:

$$Q_{t+1}(s_t, a_t) = R(s_t, a_t) + \gamma \max_{a'} Q_t(s_{t+1}, a')$$  \hspace{1cm} (3.7)
Figure 3.2: A human observer Q-learns the combined dynamics of the agent and environment. The observer critiques the agent through the TD error, $\delta_t$.

Each time the agent takes an action, he or she recovers a reward. By adding memory through the parameter $\alpha$, we arrive at a form of value iteration over the Q-values. The scheme is called Q-Learning [43]. It is defined according to the following equation:

$$Q_{t+1}(s_t, a_t) = (1 - \alpha)Q_t(s_t, a_t) + \alpha \left[ R(s_t, a_t) + \gamma \max_{a'} Q_t(s_{t+1}, a') \right]$$ (3.8)

Q-Learning is powerful because it does not require the agent to represent the value function explicitly. Instead, an agent navigates its environment, obtains rewards, and updates Q-Values as necessary. Furthermore, Q-Learning allows the agent to readily compute an optimal policy. For a given state, the agent simply chooses the action with the maximal Q value.

### 3.2 Modeling the Human Observer as a Q-Learner

Under canonical Q-Learning, an agent evaluates the environment according to equation (3.8) and computes a policy based on the learned Q values. Now suppose that the agent is not able to implement a policy on his own behalf. Rather, an actor implements a policy on behalf of the Q-Learning observer, as depicted in figure 3.2.

This setup is similar to an actor-critic scheme, in which the observer criti-
cizes the actions taken by the actor. The observer communicates a prediction error signal to the actor that specifies the observer’s deviation in expected reward. The prediction error is derived as follows from equation (3.8):

\[
Q_{t+1}(s_t, a_t) = (1 - \alpha)Q_t(s_t, a_t) + \alpha \left[ R(s_t, a_t) + \gamma \max_{a'} Q_t(s_{t+1}, a') \right] 
\]

\[
= Q_t(s_t, a_t) + \alpha \left[ R(s_t, a_t) + \gamma \max_{a'} Q_t(s_{t+1}, a') - Q_t(s_t, a_t) \right] 
\]

\[
= Q_t(s_t, a_t) + \alpha \delta_t 
\]

where \( \delta_t \) is the prediction error,

\[
\delta_t = R(s_t, a_t) + \gamma \max_{a'} Q_t(s_{t+1}, a') - Q_t(s_t, a_t) \tag{3.10}
\]

We note that this is precisely the model suggested by the reinforcement learning theory of the Feedback-Related Negativity. The observer learns about the environment and elicits a Feedback-Related Negativity, \( \delta_t \), that is proportional to the prediction error. Further, this model does not require that the observer act on his own behalf in order to elicit an FRN. Assuming that the error-processing system builds a model of expected outcomes according to a reinforcement learning scheme such as Q-Learning, the FRN will be elicited during passive observation. While we have modeled the observer as a Q-Learner based on results from the literature, other schemes for the critic are plausible [7].

Based on the work of Cavanagh we assume that the oFRN will scale with medial and lateral theta power on single trials [7]. Although no study has tested this hypothesis to date, a wide body of literature shows the relationship between the FRN and theta power at medial electrode sites [47, 41, 7]. Since the morphology of the oFRN is similar to the FRN (with reduced amplitude) and localized to the ACC, it is likely that theta power will correlate with the oFRN during passive observation. However, it should be noted that our scheme rests on this assumption.

In practice, noise in the EEG measurements and modeling inaccuracies will contribute noise to the prediction error signal. We define \( e_t \) as \( \delta_t \) with additive Gaussian noise:
\[ e_t = \delta_t + n \]  

(3.11)

where \( n \sim N(0, \frac{1}{p}) \), and it is assumed that \( p \) is a constant that is known.

Since our observer is critiquing an agent, we assume that the observer has a deterministic reward function. Therefore, we replace the \( R \) in equation (3.8) with \( r \). We further assume that before the observer begins interacting with a BCI agent he will fix his learning parameter \( \alpha \), discount factor \( \gamma \), and initial Q values, \( Q_0 \). To summarize our model, the human observer can be completely characterized by the following two equations:

\[
Q_{t+1}(s_t, a_t) = (1 - \alpha)Q_t(s_t, a_t) + \alpha \left[ r(s_t, a_t) + \gamma \max_{a'} Q_t(s_{t+1}, a') \right] \\

e_t = r(s_t, a_t) + \gamma \max_{a'} Q_t(s_{t+1}, a') - Q_t(s_t, a_t) + n 
\]

(3.12)

where \( n \sim N(0, \frac{1}{p}) \).

Hence, the human observer is a Q-Learner parameterized by \( w = [Q_0, r, \alpha, \gamma] \).

We now proceed to formulate the BCI control problem.

### 3.3 Problem Statement

*Given a human observer that learns about a BCI agent according to Q-Learning (equation (3.8)) and elicits prediction error signals according to equation (3.12), construct a BCI agent that is optimal with respect to the reward function of the observer.*

According to this problem formulation, the observer belongs to a class of observers – specifically the class of Q-Learners parameterized by \( w = [Q_0, r, \alpha, \gamma] \) for \( w \in W \). It is assumed that both the observer and BCI agent know the dynamics of the environment, \( p(s'|s, a) \). Since the dynamics are known, if the BCI agent can recover the observer’s reward function, \( r \), he can solve Bellman’s equation for \( \pi^* \) and maximize the observer’s expected discounted cumulative reward.

This is an adaptive control problem, because the agent and observer are
3.3 Bayesian Control Rule

Now we formulate a scheme to identify the reward function of the human observer and design a BCI agent that is optimal with respect to a human observer drawn from a class of observers. Throughout this section, we will refer to the human observer as the environment, and the BCI agent as the agent. Figure 3.3 depicts the problem formulation. The environment generates symbols, $o_i \in O$, and the agent generates symbols, $a_i \in A$. Each observation consists of a state and prediction error, $o_i = \{s_{i+1}, \delta_i\}$. The environment is parameterized by $w \in W$ and is fixed before the interaction.
starts. For each $w \in \mathcal{W}$, there is a set $[w]$ consisting of all $w' \in \mathcal{W}$ such that the optimal policy for mode $w'$ is the same as for $w$. In other words, for each $w$ there may exist an equivalence class of $[w]$ which all induce the same optimal agent.

For each environment $w \in \mathcal{W}$, it is assumed that there exists an optimal agent denoted by the measure $Q$. Hence, if $w$ were known, the designer would choose $Q_w$ to be the agent. In general, however, the environment is unknown but is believed to belong to a class of environments, $\mathcal{W}$. The goal is to design an agent $Q$ that is optimal with respect to this class of environments and will converge to the agent for the true environment, $w \in \mathcal{W}$.

This problem can be formulated as an adaptive coding problem over the action-observation sequence, $Z = a_1, o_1, a_2, o_2, \ldots$, generated by the coupled agent and environment [34]. Adaptive coding is the problem of compressing observations from an unknown source. Universal compressors solve this problem by minimizing the average deviation between a predictor and the true source and then constructing code words using the predictor [34]. In our setup, this consists of constructing a predictor, $Q$, and measuring its deviation from the true distribution, $Q_w$, using the KL-Divergence. The difficulty with formulating the problem this way is that the actions are generated by the agent. Hence, doing inference over the action-observation stream includes doing inference over one's own actions, which can lead to paradoxes, since actions themselves do not provide the agent with any information. The Bayesian control rule addresses this problem by treating actions and observations as random variables and using causal conditioning. That is, the Bayesian control rule treats actions as interventions, thereby obeying the rules of causality.

We state the Bayesian control rule as follows:

Given a set of operation modes $P(\cdot | w, \cdot)$ over interaction sequences, $Z$, and a prior distribution $P(w)$ over the parameters $\mathcal{W}$, the probability of the action $a_{t+1}$ is given by:

$$
P(a_{t+1} | \hat{a}^t, o^t) = \sum_w P(a_{t+1} | w, a^t, o^t)P(w | \hat{a}^t, o^t)$$  (3.13)

where the posterior probability over operation modes is given by the recursion...
\[
P(w|\hat{a}^t, o^t) = \frac{P(o_t|w, a^{t-1}, o^{t-1})P(w|\hat{a}^{t-1}, o^{t-1})}{\sum_{w'} P(o_t|w', a^{t-1}, o^{t-1})P(w'|\hat{a}^{t-1}, o^{t-1})} \tag{3.14}
\]

3.5 Derivation of Likelihood and Intervention Models

In order to apply the Bayesian Control Rule, we must compute three things:

1. A likelihood model for observations: \( p(o_t|w, o^{t-1}, a^{t-1}) \)
2. An intervention model for actions: \( p(a_{t+1}|w, a^t, o^t) \)
3. A posterior distribution over operation modes: \( p(w|a^t, o^t) \)

We start by deriving a likelihood model for observations:

\[
p(o_t|m, o^{t-1}, a^{t-1}) = p(e_t, s_{t+1}|w, o^{t-1}, a^{t-1}) \tag{3.15}
\]

\[
= p(e_t|w, o^{t-1}, a^{t-1}, s_{t+1})p(s_{t+1}|w, o^{t-1}, a^{t-1}) \tag{3.16}
\]

\[
= p(e_t|w, a_t, s_{t+1}, s_t)p(s_t|s_t, a_t) \tag{3.17}
\]

where (3.15) follows because \( o_t = (e_t, s_{t+1}) \).

(3.16) is by application of the chain rule for probabilities.

(3.17) follows because the environment dynamics are assumed to be Markov and the same for all \( m \).

Referring to equation (3.17), \( p(s_t|s_{t-1}, a_{t-1}) \) is simply the state transition dynamics of the MDP, which are assumed to be known. To compute the prediction error likelihood, \( p(e_t|w, a_t, s_{t+1}, s_t) \), we note that:

\[
e_t = r(s_t, a_t) + \gamma \max_{a'} Q_t(s_{t+1}, a') - Q_t(s_t, a_t) + n \tag{3.18}
\]

where \( n \sim N(0, \frac{1}{\rho}) \), so that \( e_t \) is normally distributed,
The likelihood model consists of a family of normal distributions parameterized by $\mu$. We note that the operation modes consist of $Q_0$, so that the current Q values can be computed by rolling back to the initial Q values.

Next, we derive the intervention model for actions. Since the state transition dynamics are assumed to be known, given an environment $w$, the optimal action is the action suggested by Bellman’s equation:

$$p(a_{t+1}|w, a^t, o^t) = \begin{cases} 
1 & \text{if } a = \pi(s) \text{ for } \pi \text{ satisfying (3.1)} \\
0 & \text{otherwise}
\end{cases}$$

Finally, we must compute the posterior belief over operation modes. If we discretize the operation modes and place a uniform prior over them, we can update the posterior directly according to equation (3.14).

These three distributions are all that is necessary to apply the Bayesian control rule. As will be shown in the next section, this scheme should converge since the environments are well-parameterized (it is clear which operation modes belong to the same equivalence class) and the Markov chain is ergodic.

### 3.6 Necessary and Sufficient Conditions for Convergence

The Bayesian control rule is guaranteed to converge to the equivalence class of the true operation mode when the operation modes are consistent and exhibit bounded variation. Ortega appeals to divergence process theory to show that these two properties are necessary conditions for convergence [34]. Rather than restating their arguments, we will summarize these two necessary conditions and prove a sufficient condition for convergence based on an analysis of the stability of the nonlinear filter.

Bounded variation can be understood as an ergodicity assumption. For the purposes of this thesis, the Markov chain describing the environment must
be ergodic. For details concerning bounded variation in terms of divergence processes, refer to [34].

Operation modes $w$ and $w^*$ are consistent if and only if $w \in [w^*]$ implies that for all $\epsilon < 0$, there is a $t_0$ such that for all $t \geq t_0$ and all $(a^t, o^{t-1})$, $|P(a_t|w, a^t, o^t) - P(a_t|w^*, a^t, o^t)| < \epsilon$. In words, this means that if we expect $w$ to have the same behavior as $w^*$, then $w$ has to converge to $w^*$’s policy. This condition is really a design condition – if we properly parameterize the operation modes and understand the equivalence classes to which they belong, consistency can be achieved.

While these sufficient conditions aid in controller design, it is convenient to have a necessary and sufficient condition that is easily testable. Recent work by Gorantla and Coleman have shown that necessary and sufficient conditions for achieving reliable message point communication can be stated in terms of the stability of the nonlinear filter [18]. We note that the Bayesian control rule consists of a nonlinear filter that updates its posterior belief according to causal conditioning. We can exploit their result to derive a relation between the prior over operation modes and the cardinality of the equivalence class of the true operation mode. We show that the Bayesian control rule will converge if and only if it is designed with respect to this relation.

We will make use of the following definitions in the lemmas and proofs that follow:

- Define $\nu^w$ to be the prior for which $\nu([w]) = 1$ and $\nu(\{w'\}) = \frac{1}{|w|}$ for any $w' \in [w]$.
- Define $\tilde{\nu}$ to be the prior that is uniformly distributed on $W$.
- Define $\pi^{w}_n$ to be the posterior given by the causal nonlinear filter with initial condition $\pi^{w}_0 = \nu^w$.

\[
\tilde{\pi}^{w}_n = \frac{\sum_{w'} P(o_n|w', a^{n-1}, o^{n-1})\pi^{w'}_{n-1}}{\sum_{w'} P(o_n|w', a^{n-1}, o^{n-1})\pi^{w'}_{n-1}}
\] (3.21)

- Analogously, define $\tilde{\pi}_n$ to the posterior given by (3.21) where the initial condition is $\tilde{\pi}_0 = \tilde{\nu}$.
- Define $\mathbb{P} = \mathbb{P}(w, a_1, o_1, a_2, o_2, \ldots)$
- Define $\tilde{\mathbb{P}} = \mathbb{P}(w, a_1, o_1, a_2, o_2, \ldots)$
We will use the following identity:

\[
\frac{d\mathbb{P}}{d\bar{\mathbb{P}}} = \frac{dv^W}{d\bar{v}}(W) \tag{3.22}
\]

We will define \textit{Contraction} as follows:

\textbf{Definition 3.6.1} \textit{We say that the Bayesian control rule is “contractive” if}

\[
\bar{\pi}_n([W]) \to 1 \text{ in } \bar{\mathbb{P}}. \tag{3.23}
\]

Contraction is the property that the posterior computed by the Bayesian control rule will place all of its mass on the equivalence class of operation modes with probability 1. This is equivalent to stating that the scheme converges on the set of operation modes that induce the same policy. We now provide a necessary and sufficient condition for contraction:

\textbf{Lemma 3.6.2} \textit{Contraction occurs if and only if}

\[
D(\pi^W_n \| \bar{\pi}_n) \to 0 \text{ in } \bar{\mathbb{P}}. \tag{3.24}
\]

\textbf{Proof 3.6.3}

\[
D(\pi^W_n \| \bar{\pi}_n) = \sum_w \pi^W_n \log \frac{\pi^W_n}{\bar{\pi}_n} \tag{3.25}
\]

For simplicity we assume \(\pi^W_n\) has all of its mass at \(w^* \in W\). Therefore, \(\pi^W_n(w^*) = 1\) and \(\pi^W_n(w) = 0\) for all \(w \neq w^*\). So the above expression reduces to

\[
D(\pi^W_n \| \bar{\pi}_n) = \pi^W_n(w^*) \log \frac{\pi^W_n(w^*)}{\bar{\pi}_n(w^*)} \tag{3.26}
\]

Since \(\bar{\pi}_n([W]) \to 1 \text{ in } \bar{\mathbb{P}}\) we have that

\[
D(\pi^W_n \| \bar{\pi}_n) \to 0 \text{ in } \bar{\mathbb{P}}. \tag{3.27}
\]
We exploit lemma 3.6.3 to derive a necessary and sufficient condition that relates the cardinality of the equivalence class of operation modes and an expectation over the modes.

**Lemma 3.6.4** Contraction occurs if and only if

\[
\mathbb{E} \left[ \frac{d\nu^u}{d\nu}(W)|\mathcal{F}_{I,\infty}^Y \right] \biggr|_{u=W} = \frac{1}{||W||}.
\] (3.28)

**Proof 3.6.5**

\[
\mathbb{E} \left[ g(W)|Y^n = y^n \right] = \frac{\mathbb{E} \left[ g(W) \frac{d\nu}{d\nu} Y^n = y^n \right]}{\mathbb{E} \left[ \frac{d\nu}{d\nu} Y^n = y^n \right]} \\
= \frac{\mathbb{E} \left[ g(W) \frac{d\nu}{d\nu} (W)|Y^n = y^n \right]}{\mathbb{E} \left[ \frac{d\nu}{d\nu} (W)|Y^n = y^n \right]} \\
= \sum_w g(w) \frac{d\nu}{d\nu}(w) \mathbb{E} \left[ \frac{d\nu}{d\nu} (W)|Y^n = y^n \right] \pi_n(dw)
\]

Thus, we have

\[
\frac{d\pi_n}{d\pi_n}(W) = \frac{d\nu}{d\nu}(W) \mathbb{E} \left[ \frac{d\nu}{d\nu} (W)|Y^n = y^n \right]
\] (3.29)

26
For a fixed $\nu^n$ and $\bar{\nu}$,

$$D(\pi_n||\bar{\pi}_n) = \mathbb{E} \left[ \log \frac{d\pi_n^u}{d\bar{\pi}_n^u}(W) | Y^n = y^n \right]$$

$$= \mathbb{E} \left[ \log \mathbb{E} \left[ \frac{d\pi_n^u}{d\bar{\nu}^u}(W) | Y^n = y^n \right] | Y^n = y^n \right]$$

$$= \mathbb{E} \left[ \log \frac{d\nu^u}{d\bar{\nu}}(W)|Y^n = y^n \right]$$

$$- \mathbb{E} \left[ \log \mathbb{E} \left[ \frac{d\nu^u}{d\bar{\nu}}(W)|Y^n = y^n \right] | Y^n = y^n \right]$$

$$= \sum_{u'} \log(\frac{1}{||u'||}\mathbb{1}_{\{u'=u\}})\pi_n^u(du')$$

$$- \log \mathbb{E} \left[ \frac{d\nu^u}{d\bar{\nu}}(W)|Y^n = y^n \right]$$

$$= - \log ||u|| - \log \mathbb{E} \left[ \frac{d\nu^u}{d\bar{\nu}}(W)|Y^n = y^n \right]$$

Therefore

$$D(\pi_n^W||\bar{\pi}_n) = D(\pi_n^u||\bar{\pi}_n)_{u=W}$$

$$= - \log ||[W]|| - \log \mathbb{E} \left[ \frac{d\nu^u}{d\bar{\nu}}(W)|Y^n = y^n \right]_{u=W}.$$

This lemma relates the cardinality of the equivalence class of the true operation mode to the prior placed on the operation modes. While this provides a necessary and sufficient condition for convergence, it is not an easily testable condition. Future work will attempt to find an easily testable property of the Markov chain that will guarantee this property, and hence guarantee convergence.
Chapter 4
Simulations and Results

The stochastic control framework was applied to a simple game. Results were obtained that show the posterior contracts to position all of its probability mass over operation modes with the true reward function.

4.1 Methods

The waterfall game is a simple game consisting of a grid with four squares (refer to figure 4.1). Obstacles are denoted by yellow squares, and the agent is denoted by either a red square or a green square. At each time step, obstacles in row 1 fall to row 2, and a new obstacle may be added to row 1. If the obstacle that falls into row 2 collides with the agent, the agent turns red. If the agent avoids the obstacle, it remains green. At each time step, the agent can either stay in its lane, or switch to the other lane. Further, at each time step one of three possible obstacles can be added to row 1. They are shown in the figure. The obstacle position in row 1 is drawn from a uniform distribution. The game consists of 18 states total (a state consists of the

![Figure 4.1: The waterfall game. At each time step, the obstacle in row 1 falls down to row 2, and a new obstacle may be introduced randomly into one of the lanes in row 1. The agent can choose either to remain in his lane or to switch to the other lane. If the agent avoids the obstacle, he stays green. If he hits the obstacle, he turns red.](image-url)
colors of all squares on the board) and two actions. This game was designed because it is simple for observers to follow and it permits a variety of reward functions to be implemented.

Each human observer, \( w \), was parameterized as \( w = [\alpha, Q_0, r] \). The learning rate parameter, \( \alpha \), was discretized so that \( \alpha \in \{0.4, 0.5, 0.6, 0.7, 0.8\} \). Five different sets of \( Q_0 \) were initialized randomly so that each \( Q_0(s, a) \in \{0, 1\} \). Five different reward functions were initialized randomly so that each \( r(s, a) \in \{0, 1\} \). All combinations of the parameters were computed, yielding a total of 125 operation modes.

The BCR was applied according to the likelihood and intervention models derived in the previous chapter. Since the operation mode parameters were discretized, the posterior could be computed directly for each mode. The human agent was modeled as the Q-Learner described in chapter 3, eliciting noisy prediction errors with \( p=1 \). The algorithm is summarized as follows:

Set initial state to \( s = s_0 \).

for \( t=1 \) to 200

Sample \( w \sim p(w|a^t, o^t) \)

Set \( a : V^*(s) = \max_a [r(s, a) + \gamma \sum_{s'} P(s'|s, a)V^*(s')] \)

Obtain \( o = (s', \delta) \)

Update \( p(w|a^t, o^t) \) according to equation (3.14)

end for

Several simulations were run for different observers, \( w^* \). Simulations were run over 200 trials. In practice, each trial would consist of about 750ms stimulus presentation followed by 750ms of processing for an average 1.5s per trial. Convergence in under 200 trials means that the optimal policy is recovered in fewer than 5 minutes.

4.2 Results

The plots in figures 4.2, 4.3 and 4.4 are for a simulation with true operation mode 14 and reward function 3. It is clear from the plots that the Bayesian control rule converged on a subset of the equivalence class of the true operation mode. Further, the scheme converged in about 100 trials. This is an
average time to convergence of approximately 2.5 minutes, which is essential
to maintain subject interest in an online BCI paradigm.

Using this simple waterfall paradigm, subjects can be instructed to im-
plement a wide variety of reward functions. For instance, subjects can be
instructed to collide with all of the obstacles (attempt to keep the agent red),
avoid all of the obstacles (keep the agent green), hit all obstacles on the left,
hit all obstacles on the right, and many others. Further, the task can be used
to test the subject’s sense of agency. For instance, subjects can be instructed
to use motor imagery to guide the agent – even if the paradigm does not
make use of motor imagery in selecting actions. This trick can be used to
test how the subject’s sense of agency influences BCI performance. These
modifications will be considered in future work.
Figure 4.3: Posterior on reward functions after 200 trials. After 200 trials, the BCR positions all of the probability mass on modes with the true reward function.

Figure 4.4: Probability of the true reward vs. trials. The plot shows the value of the posterior on operation modes containing the true reward against trials.
We have developed a stochastic control framework for designing BCIs that exploit the observational Feedback-Related Negativity prediction error signals. Based on simulations, we can use prediction error signals to infer the reward function and desired policy of a human observer. This scheme is particularly useful for design problems in which the subject wishes to train a BCI algorithm, rather than communicate messages on a trial-by-trial basis.

The paradigm can be used to understand a variety of human learners. Although we derived a Q-Learning model for the human observer, other architectures are plausible and may prove more computationally tractable. The Q-Learning model that we have presented here is well-supported in the literature – but most experimental tasks under which Q-Learning has been applied have been simple. BCIs that incorporate prediction error signals over larger state and action spaces will likely encounter forms of learning that are not captured in a simple Q-Learning model. Our scheme is general enough to accommodate enhanced models of human learning.

We note that the framework is sufficiently general to add volitional signals. For instance, it would be possible to add motor rhythm classifications to the observation stream. These signals may accelerate convergence and increase the subject’s sense of agency, thereby increasing the oFRN amplitude. We intend to test the relationship between agency and BCI performance in the future.

There is obvious future work to extend this proof of concept. Human subject testing is of immediate need, to verify that the scheme works online. Further, while we have provided a necessary and sufficient condition for convergence, it is not easily testable. Future work will investigate easily testable conditions for convergence. Finally, a thorough investigation studying the relation between theta power on single trials and the oFRN is needed to better characterize the oFRN dynamics.
References


